Emergent Structure, Semantics and Usage of Social Streams

Dipl.-Ing. Claudia Wagner, Bakk.rer.soc.oec.

DISSERTATION

zur Erlangung des akademischen Grades
eines Doktors der technischen Wissenschaften
der Studienrichtung
Informatik
an der
Technischen Universität Graz

Univ.-Doz. Dipl.-Ing. Dr.techn. Markus Strohmaier
Institut für Wissensmanagement
Technische Universität Graz

Graz 2013

For my parents

TU Graz
Graz University of Technology
Senat

Deutsche Fassung:
Beschluss der Curricula-Kommission für Bachelor-, Master- und Diplomstudien vom 10.11.2008
Genehmigung des Senates am 1.12.2008

EIDESSTATTLICHE ERKLÄRUNG

Ich erkläre an Eides statt, dass ich die vorliegende Arbeit selbstständig verfasst, andere als die angegebenen Quellen/Hilfsmittel nicht benutzt, und die den benutzten Quellen wörtlich und inhaltlich entnommene Stellen als solche kenntlich gemacht habe.

Graz, am …………………………… ……………………………………………..
(Unterschrift)

Englische Fassung:

STATUTORY DECLARATION

I declare that I have authored this thesis independently, that I have not used other than the declared sources / resources, and that I have explicitly marked all material which has been quoted either literally or by content from the used sources.

…………………………… ……………………………………………..
date (signature)

Abstract

Social streams are aggregations of data that are produced by a temporal sequence of users' activities conducted in an online social environment like Twitter or Facebook where others can perceive the manifestation of these activities. Although previous research shows that social streams are a useful source for many types of information, most existing approaches treat social streams as just another textual document and neglect the fact that social streams emerge through user activities. This thesis sets out to explore potential relations between the user activities which generate a stream (and therefore impact the emergent structure of a stream) and the semantics of a stream.

A network-theoretic model of social streams is introduced in this work which allows to formally describe social streams and the structures which emerge from them. Further, several structural stream measures which allow to compare different social streams and two novel measures for assessing the stability of emerging structures of social streams are presented in this work. In several empirical studies this work explores if a relation between semantics and user activities exists and if so to what extent this relation can be exploited for (1) the creation of semantic annotations of social streams and users and (2) the prediction of users' future activities in social streams. This work does not explore causal relations between semantics and user activities, but investigates if relational patterns between semantics and user activities exist and if they can be exploited for developing new methods which allow annotating social streams with semantic and usage metadata.

The empirical results of this work show that activity patterns which can be observed in social streams around different topics, reveal interesting differences. This suggests that a relation between activity patterns and their semantic context exists. Further, this work shows that structural stream properties and activity patterns can indeed be exploited for learning semantic annotations of social streams.

The results of this work highlight not only the relation between semantics and user behavior, but also show that (1) incorporating information about the semantic context may increase the accuracy of predictions about users'

future activities and (2) incorporating information about users' activities and the structure which emerges from them may help to semantically annotate the context in which the activities are observed. This work is relevant for researchers interested in developing semantic technologies for mining social streams or user generated content and researchers working on modeling and predicting users' behavior in online social environments.

Kurzfassung

Der aktuelle Stand der Forschung zeigt klar, dass sogenannte “social streams”, das heißt von Benutzern generierte Textströme die in einer sozialen online Umgebung erzeugt und geteilt werden, eine wertvolle Quelle für verschiedene Arten von Information sind. Obwohl diese Textströme durch Benutzeraktivitäten erzeugt werden, beschränkt sich die semantische Analyse häufig auf den Text alleine, da dieselben Methoden die zur semantischen Analyse von Textdokumenten verwendet werden, auch auf soziale Textströme angewendet werden können. Das Ziel dieser Dissertation ist es den Zusammenhang zwischen Benutzeraktivitäten (und den dadurch entstehenden emergenten Strukturen) und der emergenten Semantik von sozialen Textströmen zu erforschen, um eine Grundlagen für neue Methoden zur semantischen Analyse von sozialen Textströmen zu entwickeln.

Im Rahmen dieser Dissertation wird ein netzwerktheoretisches Model vorgestellt, welches es erlaubt, soziale Textströme und die emergenten Strukturen die daraus entstehen, formal zu beschreiben und zu charakterisieren. Weiters werden in dieser Arbeit verschiedene Maße eingeführt, die es erlauben social streams über ihre strukturellen Eigenschaften zu charakterisieren und zu vergleichen und die es ermöglichen die Stabilität der emergenten Strukturen zu messen. In mehreren empirischen Studien untersucht diese Dissertation mögliche Zusammenhänge zwischen dem Verhalten von Benutzern in sozialen online Umgebungen und der Semantik dieser Textströme. Genauer gesagt untersucht diese Arbeit ob diese Zusammenhänge (1) für die semantische Analyse und die Erstellung von semantischen Annotationen von Textströmen nützlich sind und (2) helfen können zukünftiges Benutzerverhalten im Zusammenhang mit diesen Textströmen vorherzusagen. Der Fokus dieser Arbeit liegt nicht auf der Detektion von kausalen Zusammenhängen, sondern von relationalen Mustern die genutzt werden können, um die Erstellung von semantischen Annotationen und/oder die Benutzermodellierung zu unterstützen.

Die Ergebnisse dieser Arbeit zeigen, dass einerseits die Einbindung von Informationen über das Benutzerverhalten und die daraus entstehenden Strukturen helfen können akkuratere semantische Annotationen zu erzeugen und andererseits semantische Annotation genutzt werden können, um das Benutzerverhalten akkurater vorherzusagen. Diese Arbeit ist relevant für Forscher die an der Entwicklung von semantischen Analysemethoden für soziale Textströme arbeiten und/oder an der Modellierung von Benutzerverhalten in sozialen Medien.

Acknowledgments

First I would like to thank my advisor, Markus Strohmaier, who has been helping me throughout this research and has been a constant source of inspiration. I am truly grateful for all his valuable feedback, his patience and fruitful discussions. Markus is a brilliant advisor and he really knows how to create a stimulating research environment which allows students to grow and enjoy working on exciting problems.

Various friends and co-authors deserve my gratitude. In particular, I would like to thank my colleagues at Joanneum Research and Graz University of Technology, with whom I have discussed my work in countless hours. Special gratitude must go to Philipp Singer, Lisa Posch, Silvia Mitter, Christian Körner, Jan Pöschko, Florian Klien and Simon Walk who supported me and my work at different stages. You all made me looking forward to our productive group meetings on Thursday afternoons, the nightly gaming events and the interesting and funny discussions in our gossip chat.

I also want to sincerely thank Matthew Rowe, Harith Alani, Yulan He and Enrico Motta from The Open University, with whom I had the pleasure of cooperating. They really helped me to progress with my research and understand potential relations between semantics and user behavior on the Web. I also want to express my gratitude to the whole KMi team which made KMi to the most special lab I have ever visited. The warm and welcoming atmosphere in KMi is truly unique and creates an inspiring environment where people enjoy working and make new friends.

My special thank to Peter Pirolli, Les Nelson and the whole Augmented Social Cognition Lab at Xerox PARC and Bernardo Huberman, Sitaram Asur and the whole Social Computing Lab at HP. I feel truly blessed that I was fortunate enough to work with these incredible smart and inspiring researchers. During many interesting discussions they gave me a new perspective on science and I think I have never learned more in such a short period of time than during my internships. Thank you for that and thank you for making me feel very welcome from day one!

My internships, research visits and conference trips did not only allow me to reflect on my work but also to make new friends. I am extremely grateful for all the amazing people I met during this journey. I also want to thank all my friends in Austria who have been with me during the course of this work and long before that. I feel truly blessed for having such strong ties, reciprocated with love.

Last, but not least, I would especially like to thank my parents, Ellen and Hubert Wagner, and my sister Katrin. Without them, I would not be who I am today and all this would not have been possible. Words are not enough to expose how grateful I am for their love and constant support throughout so many years.

And finally, a very special thanks to Peter, with whom I have shared most of my time during this work.

Institutional Acknowledgements

This thesis was supported by a DOC-fForte fellowship of the Austrian Academy of Science.

Contents

1 Introduction

This chapter presents a motivational introduction to the research problems which are addressed by this cumulative thesis, outlines the contributions of this work and describes the structure of this thesis.

1.1 Motivation

Social streams have become a valuable source for many kind of applications including news recommendations [Chen et al., 2010] [Abel et al., 2011], public mood and emotion mining [Bollen et al., 2009] [Choudhury et al., 2012] [Golder and Macy, 2011] [Kouloumpis et al., 2011], personality mining [Golbeck et al., 2011] [Quercia et al., 2011] [Hughes et al., 2012], interests and expertise mining [Guy et al., 2013] [Vosecky et al., 2013] [Weng et al., 2010a] and trend detection [Mathioudakis and Koudas, 2010]. In this work a social stream is defined as an aggregation of data which is produced by a temporal sequence of users' activities conducted in an online social environment like Twitter[1] or Facebook[2] where others can perceive the manifestation of these activities (which are sometimes called *digital traces*). In this work a social online environment is defined as an online environment where information consumption is mainly driven by social relations and/or group memberships.

Although previous research suggests that social streams are a useful source for many types of information, most of this research focuses on analyzing social streams as just another textual document and neglects the fact that social streams emerge through user activities. In this thesis I hypothesize that a relation between the user activities which generate a stream (and

[1] http://twitter.com

[2] http://facebook.com

therefore impact the emergent structure of the stream) and the semantics of the stream exists. Within several empirical studies this work explores if such a relation exists and if so to what extent this relation can be exploited (1) for the creation of semantic annotations of social streams and users and (2) for the prediction of users' future activities in social streams. I want to point out that this work does not explore causal relations between semantics and user activities, but investigates whether relational patterns between semantics and user activities exist and if they can be exploited for developing novel methods which allow annotating social streams with semantic and usage metadata.

This thesis is organized in three parts which set out to explore:

1. *Emergent structure*: the structure of social streams that emerges from user activities.
2. *Emergent semantics*: the extent to which structural stream properties and more generally information about user activities may support the semantic annotation of social streams.
3. *Emergent usage*: to what extent structural, usage and semantic metadata may help anticipating users' future activities in a social stream.

The goal of this first chapter is to establish the concept of social streams, related problems and research questions. Further, the structure of this dissertation is introduced.

1.2 Social Streams

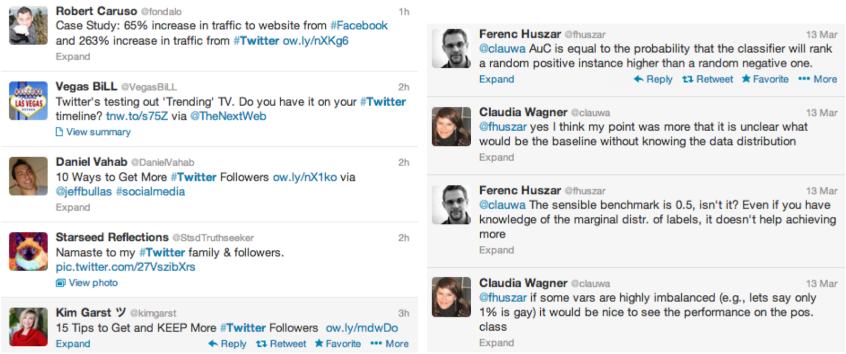
This work focuses on textual social streams – i.e., aggregations of textual data which are generated through users activities in an online social environment where information consumption is mainly driven by social relations and/or group memberships. For example, on Twitter or Facebook information consumption is driven by explicitly defined social networks and group memberships, while on message boards like Boards.ie[3] infor-

[3] http://boards.ie

mation consumption is driven by community memberships where each community corresponds to one forum. Beside social relations and group memberships, novelty is usually another factor which drives information consumption in online social environments. For example, microblogs, social networking sites, forums and message boards display the manifestations of user activities (e.g., the users's status updates or comments) in reverse chronological order.

Depending on the *type of user activity* which generates the data one can differ between different types of social streams such as *status update streams*, *retweet streams*, *user profile update streams*, *comment streams* or *like/voting streams*. Further, depending on the *aggregation type* one may differentiate between *user streams* (which aggregate messages related with one or several users), *resource streams* (which aggregate messages containing one or several resources), and *conversation streams* which are a special type of user streams and only contain messages which are part of a conversation between two or more users. For example, a *user status update stream* aggregates status updates published by all users included in a predefined set of users U, while a *resource status update stream* aggregates status updates which include resources (e.g., links or keywords) defined in the set of resources R. A *user like/voting stream* aggregates items for which at least one user of the predefined set of users U has voted. Figure 1.1 shows an examples of two different types of social streams from Twitter.

In contrast to other stream-based systems where the data structures are formally defined by system developers, social streams are different in the sense that users may collectively generate and impact the data structure of social streams that goes far beyond what the system designers' have envisioned. Emerging syntax conventions, such as RT (retweets), # (hashtags) or @ (replies), are examples of innovations by users or groups of users that superimpose an informal, emerging data structure on social streams. This has made social streams complex and dynamic structures which can be analyzed in a staggering variety of ways, for example, according to the author(s) of messages, the recipients of messages, the links, keywords or hashtags contained in messages, the time stamps of messages or the message types.



Robert Caruso @fondalo 1h
Case Study: 65% increase in traffic to website from #Facebook and 263% increase in traffic from #Twitter ow.ly/nXKg6

Vegas BiLL @VegasBiLL 2h
Twitter's testing out 'Trending' TV. Do you have it on your #Twitter timeline? tnw.to/s75Z via @TheNextWeb

Daniel Vahab @DanielVahab 2h
10 Ways to Get More #Twitter Followers ow.ly/nX1ko via @jeffbullas #socialmedia

Starseed Reflections @StsdTruthseeker 2h
Namaste to my #Twitter family & followers. pic.twitter.com/27VszibXrs

Kim Garst ツ @kimgarst 3h
15 Tips to Get and KEEP More #Twitter Followers ow.ly/mdwDo

(a) A hashtag stream.

Ferenc Huszar @fhuszar 13 Mar
@clauwa AuC is equal to the probability that the classifier will rank a random positive instance higher than a random negative one.

Claudia Wagner @clauwa 13 Mar
@fhuszar yes I think my point was more that it is unclear what would be the baseline without knowing the data distribution

Ferenc Huszar @fhuszar 13 Mar
@clauwa The sensible benchmark is 0.5, isn't it? Even if you have knowledge of the marginal distr. of labels, it doesn't help achieving more

Claudia Wagner @clauwa 13 Mar
@fhuszar if some vars are highly imbalanced (e.g., lets say only 1% is gay) it would be nice to see the performance on the pos. class

(b) A conversation stream.

Figure 1.1: Examples of different types of social streams. A resource stream, or more specific a hashtag stream (aggregation type), consisting of status updates (type of user activity) which contain the resource "#Twitter". A conversation stream (aggregation type) consisting of status updates (type of user activity) which are part of a conversation between two selected users.

To formally describe social stream aggregation this thesis introduces an extensible network-theoretic model which is capable of accommodating the complex and dynamic structure of social streams.

1.2.1 Emergent Usage Patterns, Structure and Semantics

Emergence is what "self-organizing" processes produce [Corning, 2002]. According to Lewes [Lewes, 1879] the emergent is unlike its components in so far and it cannot be reduced to their sum or their difference. Also Kaufman [Kauffman, 1995] describes emergent phenomenons as something where "*the whole is greater than the sum of its parts*".

Usage patterns emerge from social streams if a large group of users behaves similar in a certain context. For example, if a large group of users frequently comments on messages published by a certain type of user (e.g., newbies), one might call this a usage pattern since it depicts a regularity in the behavior of users in a social stream.

Structural patterns and usage patterns are tightly connected since the structure of social streams is directly impacted by the usage behavior of users. That means, if usage behavior becomes stable (in the sense that a large amount of users behave similar) also the structure of social streams might become stable. However, emerging usage patterns do not necessarily cause emerging structure but are a pre-requesite for emerging structure. If users' behavior on social streams would be completely random, no structural patterns and usage patterns could emerge from them. In collaborative tagging systems such as delicious[4], previous research has shown that the tagging vocabulary which is applied to a certain web resource becomes stable over time (see e.g., [Golder and Huberman, 2006], [Cattuto et al., 2007] and [Halpin et al., 2007]). This indicates that the structure which emerges from social tagging systems (a so-called folksonomy [Lambiotte and Ausloos, 2006]) becomes stable over time.

Emerging structure may also be related with emerging semantics since the fact that the structure becomes stable may indicate that a population of agents has *agreed* on something – e.g., on a certain set of tags which describe a resource. *Semantics* lack a precise definition, but are often characterized via a mapping between a syntactic structure and some domain [Aberer et al., 2004]. [Rapaport, 1988] suggests the notion of semantics as correspondence which means that the semantic interpretation function recursively maps symbols to themselves or other symbols. A dictionary is a simple example where the interpretation of symbols (words) is given by the mean of the same symbols. [Aberer et al., 2004] defines *emergent semantics* as a phenomenon where global semantics emerge from a society of agents and represents the common current semantic agreement.

In social streams semantics emerge when a set of users agrees on a common vocabulary to talk about the same entities, concepts and/or resources. If agreement happens, the structure becomes stable and a stable structure is a pre-request for emerging semantics. However, one needs to note that an emerging and stable structure does not imply emerging semantics. For example, assume that users have to use a certain tag to share content with their classmates. In this case a large group of users behaves similar and therefore parts of the tag-resource-user structure would become

[4] `http://delicious.com/`

stable. However, no semantics *emerge* since the stable pattern does not reflect the agreement of a large group of users but was forced by external factors.

1.3 Problem Statement and Research Questions

Social streams are a relatively new but fast growing source of data. However, the nature of those streams is mainly unexplored. Since social streams grow fast, browsing and consuming them efficiently is a challenge for users. Semantic and usage metadata of social streams and individual messages may allow users to browse social streams more efficiently by guiding them towards messages about topics they are interested in and which will receive a lot of attention in the future. *Semantic metadata* may describe the topics a social stream is about, while *usage metadata* may depict how a social stream was used in the past and/or anticipate how it will be used in the future. *Structural metadata* describe structural properties of social streams (e.g., the variety of users participating in the stream, the balance of users' participation efforts, the focus of users' hashtag usage or the focus of users' retweet activities) which emerge from users' activities conducted on social streams.

This thesis is built around the following three general research questions:

- *Emergent structure*: What is the emergent structure of social streams and how can we formally describe it?
- *Emergent semantics*: To what extent may structural metadata and usage metadata contribute to acquiring emerging semantics from streams and annotating streams with semantic metadata?
- *Emergent usage*: To what extent do structural, usage and semantic metadata help predicting future usage activities in social streams?

In the following I elaborate the three general questions around which this work is built, their objectives and the scientific publications that form this thesis:

1.3.1 Emergent Structure

RQ 1: What is the emergent structure of social streams and how can we formally describe it? Social streams are a relatively new and fast growing source of data. Many different types of social streams exist and their structure is not necessarily predefined by system developers but emerges via user activities and is therefore a collectively-generated data structure that may go far beyond what the system designers' have envisioned. Emerging syntax conventions, such as RT (retweets), # (hashtags) or @ (replies), are examples of innovations by users that superimpose an informal, emerging data structure on social streams. This has made social streams complex and dynamic structures which can be analyzed in a staggering variety of ways, for example, according to the author(s) of messages, the recipients of messages, the links, keywords or hashtags contained in messages, the time stamps of messages or the message types.

The objective of this first research question is to create a theoretical foundation of social streams which allows to describe their structure and compare different types of stream aggregations according to their structural properties. In [Wagner and Strohmaier, 2010] a network theoretic model of social streams together with structural stream measures is introduced. Further, this work shows how the model and the measures can be used to explore the nature of social streams.

This thesis also compares different methods for measuring the stability of structures emerging from social streams and shows that not all methods which have been used in previous work on tag streams are equally suitable for measuring stability. Based on this discussion, in [Wagner et al., 2013c] two novel methods for measuring the extent to which the structure of social streams stabilizes are presented. These methods allow to overcome important limitations of previous work.

1.3.2 Emergent Semantics

RQ 2: To what extent may structural metadata and usage metadata contribute to acquiring emerging semantics from streams and annotat-

ing streams with semantics? Since social streams lack explicit semantics, adding semantics to streams is crucial for increasing the searchability and navigability of social streams. The objective of this research question is to explore to what extent structural metadata and usage metadata can be exploited for semantically annotating social streams and to what extent novel social stream mining methods which go beyond existing text mining approaches could benefit from this.

While the structural framework introduced in this thesis is general and can be used to answer a wide range of empirical question about the usage and the semantics of social streams, the scope of the empirical studies presented in this thesis is limited. The following publications address a set of interesting empirical questions which allow gaining insights into the relation between emergent structure, usage and semantics.

In [Wagner et al., 2011], we study to what extent information about users' activities can be exploited to improve the quality of semantic annotations of users. We investigate a novel method for creating semantic annotations of user streams by incorporating information about user activities into the text mining process. This work shows that for predicting the topics of future messages of user streams knowing who communicated with whom is useful since messages authored by a user are more likely to be about similar topics than messages authored by users with whom the user has communicated in the past. Therefore, we conclude that structural metadata which emerge from users' communication activities in social streams may indeed be useful for creating semantic annotations of user streams.

Based on this observation one may hypothesize that the background knowledge of the audience might be useful for creating semantic annotations of social stream messages. In [Wagner et al., 2013b], we test this hypothesis and explore the value of the background knowledge of the audience for the task of semantically annotating social media messages. Our results suggest that the audience of a social stream possesses knowledge which may indeed help to interpret the meaning of a stream's message and can even be more useful than ontological background knowledge. Further, our work suggests that the audience of a social stream is most useful for interpreting the meaning of a stream's message if the stream is created and consumed by a stable and communicative community – i.e., a group of users who is interconnected and has few core users to whom almost everyone is connected.

While the above mentioned work is limited to single types of user activities, in [Wagner et al., 2012a] we analyze the potential of different types of data (generated through different types of user activities) for informing human's expertise judgements as well as computational expertise models. We conducted a comparative user study as well as a prediction experiment on Twitter data. Our results show that different types of data generated through different types of user activities are useful for different computational and cognitive tasks. The task of expertise modeling benefits most from information contained in user lists as opposed to tweet, retweet or bio information. Therefore, we conclude that metadata which may reveal which type of user activities generated the data is useful for creating semantic annotations of Twitter users, since one can select the data which has the highest information value for a given task such as expertise mining.

In [Wagner et al., 2013a] we construct a comprehensive collection of features (including *linguistic style*, *semantics*, *activity patterns* and *social-semantics*) and examine their efficacy in classifying users on Twitter according to two dimensions of interest – personality and profession. Our results show that the classifiers built on an optimal subset of all features obtain an impressive accuracy of around 0.9 across all categories. When examining the features independently, we observed that, not surprisingly, the feature groups differ in their accuracy across the various categories. For example, we found that social-semantic features (including information about users' list memberships and the tweets of their friends) and linguistic style features tend to work well with both personality related attributes and professions. Contrary, we found that activity features were not useful across all classes. This indicates that *how a user says something* (linguistic style) and *what others say about/to a user* (social-semantic) tend to be most useful for identifying users' professional areas and personality related attributes. *What a user says* (semantics) and *how a user behaves online* (activity patterns) tend to reveal less information about his professional areas and personality.

Beside user streams, also hashtag streams gained a lot of interest in the recent past, since on Twitter users started using hashtags to organize their content and annotate it with contextual information. Despite the added value of hashtags which allow to organize messages and join conversations, they lack explicit semantics and it is often hard to guess the meaning of a hashtag [Laniado and Mika, 2010]. Adding hashtags to predefined and ontologically grounded semantic categories is useful since it allows to easily infer the meaning of hashtags and relations between them. In [Posch et al., 2013] we present an empirical study and measure the utility of usage metadata and structural metadata for classifying hashtags into semantic categories. Our work shows that different semantic categories of hashtag streams reveal significantly different usage patterns and consequently reveal significantly different structural properties which can be used to create a semantic classification system of hashtag streams. This indicates that information about emergent usage patterns in social steams as well as properties of emergent structures may indeed support the semantic annotation process.

1.3.3 Emergent Usage

RQ 3: To what extent do structural, usage and semantic metadata help predicting future usage activities in social streams? The objective of this research questions is to explore to what extent structural, usage and semantic metadata may help anticipating users' future activities in social streams and therefore allow annotating them with usage metadata. Specifically, this thesis focuses on usage metadata which capture if anyone will reply to a message or not (i.e., is the message of interest to anyone), how many users will reply to the message (i.e., how interesting is the message from a global perspective) and who will reply to the message (i.e., how interesting is the message from the local perspective of an individual user). Anticipating users' reply behavior is central for predicting the attention focus of a group of users or an individual user and may therefore help to guide users towards content they might want to see and/or react on. To study users' reply behavior we analyze user streams and conversation streams which are a special type of a user streams that aggregate only messages which belong to conversations between a group of users.

While the structural framework introduced in this thesis is general and can be used to answer a wide range of empirical questions related to usage patterns in social streams, the empirical studies related to the prediction of the future usage of social streams are limited in their scope. In the following publications we present a set of interesting empirical questions which help to gain insight into the relation between the usage of social streams and the semantics of those streams.

In [Wagner et al., 2012b] and [Wagner et al., 2012c] we investigate if different user streams (which correspond to the aggregation of data created by certain groups of users) reveal interesting differences in how they are used and which factors help to predict the future evolution of the stream. Concretely, we explore which semantic and usage metadata contribute most to the prediction of how many replies a message will get. We use the number of replies as a proxy measure for attention. Our findings show that message streams of different forums exhibit interesting differences in terms of how attention is generated and therefore suggest that the existence of global, context-free attention patterns is highly questionable. For example, we find that the purpose of a stream as well as the specificity of the stream's topic impact which factors drive the reply behavior of a group of users. Streams around very specific topics require messages to fit to the topical focus of the stream in order to attract attention while streams around more general topics do not have this requirement. In streams that lack specificity everyone can participate, but posts are required to be rather short in order to minimize effort while still containing distinct terms in order to attract attention.

Beside analyzing which features have the power to predict the number of replies a message will receive, this work also investigates the task of predicting who will reply to a message in social streams. In [Schantl et al., 2013] we explore which types of features (structural versus semantic features) contribute most to predicting who will reply to a message within conversation streams on Twitter. Our work suggests that on Twitter conversations are more driven by social factors than by semantic factors –

i.e., that structural metadata which capture the social structure of users are more suitable for predicting who will reply to a message than semantic metadata which describe topical similarities between users.

1.4 Main publications

This cumulative dissertation consists of the following main publications which are associated with the research questions and the topics that are covered in this thesis in Figure 1.2:

- Claudia Wagner and Markus Strohmaier, The Wisdom in Tweetonomies: Acquiring Latent Conceptual Structures from Social Awareness Streams, Semantic Search Workshop at WWW2010, Raleigh, US, 2010.
- Claudia Wagner, Philipp Singer, Markus Strohmaier and Bernardo Huberman, Semantic Stability in People and Content Tagging Streams, under review, 2013.
- Claudia Wagner, Markus Strohmaier and Yulan He, Pragmatic metadata matters: How data about the usage of data effects semantic user models, Social Data on the Web Workshop co-located with 10th International Semantic Web Conference (ISWC2011), Bonn, Germany, 2011.
- Claudia Wagner, Vera Liao, Peter Pirolli, Les Nelson and Markus Strohmaier, It's not in their tweets: Modeling topical expertise of Twitter users, ASE/IEEE International Conference on Social Computing (SocialCom2012), Amsterdam, The Netherlands, 2012.
- Claudia Wagner, Sitaram Asur, Joshua Hailpern, Religious Politicians and Creative Photographers: Automatic User Categorization in Twitter, ASE/IEEE International Conference on Social Computing (SocialCom2013), Washington D.C., USA, 2013.
- Claudia Wagner, Philipp Singer, Lisa Posch and Markus Strohmaier, The Wisdom of the Audience: An Empirical Study of Social Semantics in Twitter Streams, European Semantic Web Conference

(ESWC), Montpellier, France, 2013.

- Lisa Posch, Claudia Wagner, Philipp Singer and Markus Strohmaier, Meaning as Collective Use: Predicting Hashtag Semantics on Twitter, 3rd Workshop on Making Sense of Microposts at WWW2013, Rio de Janeiro, Brazil, 2013.
- Claudia Wagner, Matthew Rowe, Markus Strohmaier and Harith Alani, Ignorance isn't Bliss: An Empirical Analysis of Attention Patterns in Online Communities, ASE/IEEE International Conference on Social Computing (SocialCom2012), Amsterdam, The Netherlands, 2012. (Best Paper)
- Claudia Wagner, Matthew Rowe, Markus Strohmaier and Harith Alani, What catches your attention? An empirical study of attention patterns in community forums, The International AAAI Conference on Weblogs and Social Media (ICWSM2012), Dublin, Ireland, 2012.
- Johannes Schantl, Claudia Wagner, Rene Kaiser and Markus Strohmaier, The Utility of Social and Topical Factors in Anticipating Repliers in Twitter Conversations, ACM Web Science (WebSci2013), Paris, France, 2013.

1.5 Contributions and Implications

The contribution of this thesis is threefold:

- First, this work introduces a network theoretic model of social streams which allows to formally describe the structure emerging from social streams, introduces various measures that allow comparing structural properties of social streams and introduces two novel measures for estimating the stability of the structures emerging from social streams. This model and the measures are general and universally applicable to existing and future manifestations of social streams.
- Second, this thesis empirically shows that structural metadata and usage metadata can be exploited for semantically annotating social

streams.

- Finally, the empirical results of this work show that semantic metadata and information about users' activities on social streams (and the structure emerging from those activities) can be exploited for predicting future activities of users in social streams. However, users' activities may differ depending on the context (e.g., platform context or topical context) and therefore context-specific user model may increase the accuracy of predictions about users' future activities.

The findings presented in this work imply that users' activities in social stream depend on the context (e.g., platform context or topical context) and that context-specific user models may increase the accuracy of predictions about users' future activities; incorporating information about users' activities may help to create more accurate semantic annotations of the context in which the activities have taken place. Therefore, we conclude that the task of semantically annotating social streams may benefit from taking information about users' activities into account, while the task of predicting users' activities may benefit from taking semantic and other contextual information into account.

1.6 Organization of this Thesis

Figure 1.2 visualizes the structure of this dissertation. It shows how the research questions relate to the topics addressed in this work and highlights the relation between individual papers and these topics. This dissertation consists of three parts: emergent structure, emergent semantics and emergent usage. The three parts and their relations are outlined in Section 1.2.1.

The first research question explores structural aspects of social streams. The structure of social streams captures and abstracts users' activities in social streams (i.e., the usage of social streams). The second research question investigates to what extent semantics emerge from social streams by exploring the stabilization of social stream structures and the utility of information about the structure and usage of social streams for semantically annotating them. Finally, the third research question explores the predictability of users' communication behavior and especially focuses on associations between semantics of social streams and users' future behavior in those streams.

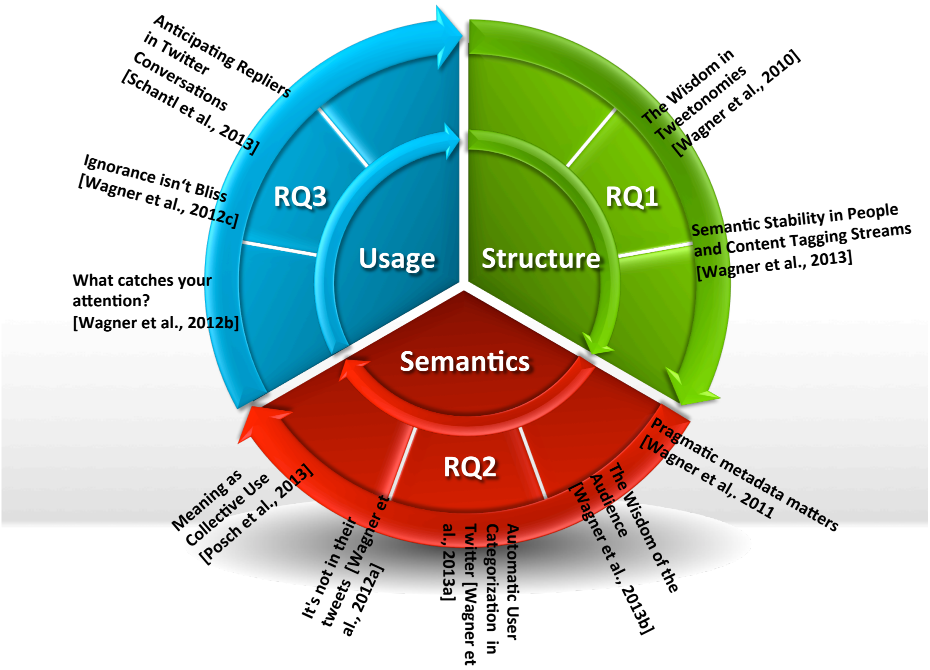

Anticipating Repliers in Twitter Conversations [Schantl et al., 2013]
Ignorance isn't Bliss [Wagner et al., 2012c]
What catches your attention? [Wagner et al., 2012b]
RQ3
Usage
The Wisdom in Tweetonomies [Wagner et al., 2010]
RQ1
Structure
Semantic Stability in People and Content Tagging Streams [Wagner et al., 2013]
Semantics
RQ2
Pragmatic metadata matters [Wagner et al,. 2011
The Wisdom of the Audience [Wagner et al., 2013b]
Automatic User Categorization in Twitter [Wagner et al., 2013a]
It's not in their tweets [Wagner et al., 2012a]
Meaning as Collective Use [Posch et al., 2013]

Figure 1.2: A structural visualization of the relationship between the main research questions of this thesis and the associated publications. Usage patterns emerge from social streams if a large group of users behaves similar in a certain context. Stable usage patterns are a pre-requisite for stable structures. Emerging stable structures may also relate to emerging semantics since the fact that the structure becomes stable may indicate that a population of agents has agreed on something. Finally, semantics may also impact usage patterns since in different semantic contexts user may behave differently.

2 Related Work

For this dissertation, research on usage patterns and emerging structures in social streams, as well as research on exploring emergent semantics and semantic annotations of social streams is relevant. This chapter is intended to give a high-level, incomplete introductory overview of these related research areas. For further details see the corresponding related work sections in the papers included in this dissertation.

2.1 Emergent Structure of Social Streams

Previous research on emerging structures mainly focused on exploring structures emerging from tag streams in collaborative tagging systems such as delicious. One of the first and most important work which shows that stable structural patterns emerge from social tagging systems was presented by [Golder and Huberman, 2006]. Their work shows empirically that the number of tags needed to describe an object consistently converges to a power law distribution as a function of how many tags it receives. A power law distribution is a good sign of stability since, due to the scale invariance property of power law distributions, increasing the number of tagging instances only proportionally increases the scale of the power law, but does not change the parameters of the power law distribution. Further they find that the proportion of frequencies of tags within a given site stabilizes over time. [Halpin et al., 2007] also find stable structural patterns emerging from delicious and propose that stabilization can also be detected by using the Kullback-Leibler (KL) divergence as an information-theoretic metric that describes the difference between the tag distributions of a resource at different points in time. Their work shows that the KL divergence converges relatively quickly to a very small value (with a few outliers).

To simulate the tagging process, [Golder and Huberman, 2006] propose that the simplest model that results in a power law would be the classical Polya urn model. The first model that formalized the notion of new tags was proposed by [Cattuto et al., 2007]. They explore the utility of the Yule-Simon model [Yule, 1925] to simulate tagging data and conclude that it seems to be unrealistic that users are choosing to reinforce tags uniformly from a distribution of all tags that have been used previously. According to Cattuto et al. it seems more realistic to assume that users tend to apply recently added tags more frequently than old ones. Therefore, they present a Yule-Simon model with a long-term memory.

In addition to exploring the imitation behavior which may drive the tagging process, previous research [Dellschaft and Staab, 2008] started exploring the utility of background knowledge which may help to explain the stable emerging structures from social tagging systems. Dellschaft et al. use the word frequency distributions obtained from different text corpora as background knowledge and pick tags according to those frequencies. They show that combining background knowledge with imitation mechanisms improves the simulation results. Although their results appear strong, their evaluation has certain limitations since they only tried to reproduce the shape of the original tag distribution produced by humans using their simulation model without comparing the real tags and their order. Dellschaft et al. as well as Cattuto et al. show that the low-rank part (between rank 1 and rank 7-10) of the ranked tag frequency curves exhibits a flatten slope which is typically not observed in systems strictly obeying Zipf's law [Zipf, 1949]. Therefore, the authors argue that a model which can simulate this tagging-specific shape of a curve is suitable to explain the tagging process. However, recent work by [Bollen et al., 2009] questions that the flatten head of this distributions is a characteristic which can be attributed to the tagging process itself but may only be an artifact of the user interface which suggests up to ten tags.

Though previous research mostly agrees on the fact that stable structures emerge from social tagging systems, it remains unclear what causes the stability. Imitation behavior of users and shared background knowledge are the two main explanatory factors which have been explored in previous research. An alternative explanation is that the emergent structure in social tagging systems is simply a product of the stable structure which we can observe in the natural language usage [Ferrer-i Cancho and Elvevåg, 2010] [i Cancho and Solé, 2001].

Unlike previous work, this thesis goes beyond analyzing structures of resource streams emerging from collaborative tagging systems and introduces an extensible formal model capable of accommodating the complex and dynamic structure of a large variety of social streams found in online social environments. The model which is introduced in this thesis builds on top of previous research (see e.g., [Lambiotte and Ausloos, 2006], [Hotho et al., 2006], [Mika, 2007] and [Heymann et al., 2008]) which introduced a formal model, a so-called *folksonomy*, which describes the structure of social streams emerging from collaborative tagging systems. Further, this thesis highlights that not all methods which have been used in previous work for measuring the stabilization of social streams (or more specific tag streams) are equally suitable and presents two novel methods for assessing if and to what extent the structures emerging from social streams become stable.

2.2 Emergent Semantics and Semantic Annotations of Social Streams

Semantics lack a precise definition, but are often characterized via a mapping between a syntactic structure and some domain [Aberer et al., 2004]. [Rapaport, 1988] suggests the notion of semantics as correspondence which means that the semantic interpretation function recursively maps symbols to themselves or other symbols. A dictionary is a simple example where the interpretation of symbols (words) is given by the mean of the same symbols. *Emergent semantics* is a phenomenon where global semantics emerge from a society of agents and represents the common current semantic agreement [Aberer et al., 2004]. *Semantic annotations* of social streams or users usually aim to reveal the main topics the stream is about or the main topics the user is interested in or knows about. However, semantic annotations of users may also reveal other user attributes (which go beyond interest and knowledge) such as the origin, age, gender or personality of a user. Ideally semantic annotations reflect semantic associations on which a large group of users would agree on when being asked to describe or tag a social stream or user.

There exists an interesting body of work on algorithms which enable the automatic construction of term hierarchies and ontologies which are manually constructed and agreed up on specifications of conceptualizations. For example, [Sanderson and Croft, 1999] describe the extraction of concept hierarchies from a document corpus. They use a simple statistical model for subsumption and apply it to concept terms extracted from documents returned for a directed query. Another line of research (e.g., [Hearst, 1992]) suggests to use lexico-syntactic patterns (e.g., “such as”) to detect hyponymy relations in text. Finally, the use of hierarchical clustering algorithms for automatically deriving term hierarchies from text was, amongst others, proposed in [Cimiano et al., 2005].

Since on the Social Web new data structures such as *folksonomies* (see e.g., [Lambiotte and Ausloos, 2006], [Hotho et al., 2006], [Mika, 2007] and [Heymann et al., 2008]) emerge, the extension and adaption of traditional content and link analysis algorithms and ontology induction algorithms became a key question. Several data mining techniques such as dimensionality reduction and clustering techniques have been applied and adapted to folksonomies. For example, [Schmitz, 2006] describe how they mine from a tag space association rules of the form *If users assign the tags from X to some resource, they often also assign the tags from Y to them.* If resources tagged with t_0 are often also tagged with t_1 but a large number of resources tagged with t_1 are not tagged with t_0, t_1 can be considered to subsume t_0. [Mika, 2007] presents a graph-based approach and shows how lightweight ontologies can emerge from folksonomies in social tagging systems. For mining concept hierarchies he adopts the set-theoretic approach that corresponds to mining association rules as is described by Schmitz et al.. [Heymann and Garcia-Molina, 2006] represents each tag t as a vector (of resources tagged with the tag) and computes cosine similarity between these vectors. That means, they compute how similar the distributions of tags are over all resources. To create a taxonomy of tags, they sort the tags according to their closeness-centrality in the similarity graph. They start with an empty taxonomy and add a tag to the taxonomy as a child of the tag it is most similar to, or as a root node if the similarities are below a threshold. In [Begelman, 2006] [Gemmell et al., 2008] the authors illustrate how tag clusters serving as coherent topics can aid in the personalization of search and navigation. [Gemmell et al., 2008] compare hierarchical agglomerative clustering [Gower and Ross, 1969], maximal complete link clustering [Augustson and Minker, 1970] and k-means clustering [MacQueen, 1967] and show that hierarchical agglomerative clustering performs best in the personalization task, followed by maximal complete link clustering. K-means was the worst of the three methods since e.g. ambiguous tags can pull unrelated tags together and k-means also fails to identify innocuous tags according to [Gemmell et al., 2008]. Hierarchical agglomerative clustering begins using each tag as a single cluster and joins clusters together depending on the level of similarity between the clusters during each stage of the process. Maximal complete link clustering identifies every maximal clique in a graph – i.e., every clique that is not contained in a larger clique. K-means starts with a predetermined number of clusters and populates clusters randomly with tags. Centroids are calculated for each cluster and each tag is reassigned to a cluster based on a similarity measure between itself and the cluster centroid during each iteration. In [Helic et al., 2011] the authors use different algorithms (affinity propagation [Frey and Dueck, 2007], a hierarchical version of the spherical k-means introduced by [Dhillon et al., 2001] and the graph based algorithm introduced by [Heymann and Garcia-Molina, 2006]) which allow learning hierarchical structures from folksonomies. The authors compare and evaluate these algorithms by using the hierarchical structures which they produce as background knowledge for decentralized search. Spherical k-means represents each document and each cluster mean as a high-dimensional unit-length vector and uses cosine similarity as a similarity measure. The authors use the spherical k-means iteratively in a top-down manner to build a tag hierarchy. First the whole input data set is used for clustering the data into 10 clusters and clusters containing more than 10 connected samples are further partitioned. Clusters which contain less than 10 samples are considered as leaf clusters. Algorithms like k-means are quite sensitive to the initial selection of “exemplars”, which are the centers that are selected from actual data points. Affinity Propagation (AP) [Frey and Dueck, 2007] is a new clustering method that overcomes this limitations and accepts a set of similarities between data samples as input. Real-valued messages are exchanged between data points until a high-quality set of exemplars and corresponding clusters gradually emerges. [Strohmaier et al., 2012] present a comprehensive evaluation of various ontology induction algorithms. In addition to adopting semantic evaluation techniques, the authors present and adopt a new technique that can be used to evaluate the usefulness of folksonomies for navigation (see [Helic et al., 2011] and [Strohmaier et al., 2012]). Their results show that centrality based algorithms outperform hierarchical clustering algorithms by a large margin.

Many clustering techniques require to calculate the similarity between tags. [Markines et al., 2009] define, analyze and evaluate different semantic similarity relationships obtained from mining socially annotated data.

When it comes to the automated semantic annotations of textual documents, we can differentiate between three general types of text mining approaches:

- *Bag of Word approaches* (BOW) provide a simple vector space representation of documents where each item in the vector corresponds to a word [Salton and McGill, 1986]. Depending on which weighting schema is used the weight of each word may depend on if the word occurs in the document or not (binary weighting), how often the word occurs in the document (term frequency weighting), how often the word occurs in the document and how often it occurs in all other documents (term frequency inverse document frequency weighting).

  Bag of words approaches allow to annotate individual documents with words which are representative for the documents.

- *Latent Semantic approaches* such as latent semantic analysis (LSA) [Deerwester et al., 1990], probabilistic latent semantic analysis (pLSA)

[Hofmann, 1999] and topic models (such as LDA) [Blei et al., 2003] exploit the implicit higher-order structure in the association of terms with documents. These methods assume that words that are close in meaning will occur in similar pieces of text. Also other dimensionality reduction and clustering methods can be applied to a document-word matrix and allow to reduce the dimensionality of the matrix e.g. by clustering words into semantically coherent groups.

LSA constructs a matrix containing word counts per paragraph (rows correspond to words and paragraphs correspond to columns). Singular Value Decomposition (SVD) is used to reduce the dimensionality of the matrix by deleting elements representing dimensions which do not exhibit meaningful variation. That means, noise in the representation of word vectors is eliminated and the word vectors become shorter, and contain only the elements that account for the most significant correlations among words in the original dataset. Taking the cosine similarity between the two vectors formed by any two rows allows to find similar words which form"latent concepts". Unlike LSA which is based on linear algebra, pLSA and LDA are based on a mixture decomposition derived from a latent class model.

Latent semantic methods allow to annotate individual documents with latent concepts where each concept is represented via a word cluster.

- Explicit Semantic approaches such as Explicit Semantic Analysis (ESA) [Gabrilovich and Markovitch, 2007] indexes documents with respect to an external conceptual space (e.g., the Wikipedia article space), indicating how strongly a given word in the document (and by aggregation also the whole document) is associated to an external concept.

  Explicit semantic methods allow to annotate individual documents with explicit external concepts (e.g., a Wikipedia articles).

Unlike previous work this thesis focuses on analyzing the impact of user activities (and usage patterns and structural patterns emerging from user activities) on creating semantic annotations. Previous research on social streams in collaborative tagging systems found empirical evidence that the emergent semantics of tags in folksonomies are influenced by the pragmatics of tagging – i.e., the tagging practices of individual users [Körner et al., 2010]. This work suggests that the quality of emergent semantics (what a concept means) depends on the usage behavior of users. In this thesis we further explore this hypothesis in the broader context of social streams.

A common approach for evaluating the quality of semantic models is to use them for a certain task. In this thesis three different tasks are considered: semantic annotations of users, semantic annotations of hashtags and semantic annotations of individual messages. Therefore, we review research analyzing semantics of user and hashtag streams, as well as research on semantically annotating and enriching social media messages in the following subsections.

2.2.1 Semantic Annotations of User Streams

A substantial amount of research has been conducted on predicting various user attributes such as gender [Burger et al., 2011] [Rao et al., 2010], age [Rao et al., 2010], ethnicity [Pennacchiotti and Popescu, 2011], mood and sentiment (see e.g., [Bollen et al., 2009], [Choudhury et al., 2012], [Golder and Macy, 2011], [Kouloumpis et al., 2011], personality (see e.g., [Golbeck et al., 2011], [Quercia et al., 2011], and [Hughes et al., 2012]), interests and expertise (see e.g., [Hong and Davison, 2010] and [Vosecky et al., 2013]) from users' digital traces produced on social media. Such predicted user attributes can be seen as semantic annotations of users since they reveal additional information about the user which may help to classify and organize users according to various dimensions of interest.

[Rao et al., 2010] classify Twitter users according to a set of latent user attributes, including gender, age, regional origin, and political orientation. They show that message content is more valuable for inferring the gender, age, regional origin, and political orientation of a user than social network's structure or communication behavior. Rather than performing general user classification, [Pennacchiotti and Popescu, 2011] specifically models a user's political affiliation, ethnicity and affinity for a set of specific business. While their approach combines both user-centric features (profile, linguistic, behavioral, social), and social graph based features, their results suggest that user-centric features alone achieve good classification results, and social graph information has a negligible impact on the overall performance. In addition they found that semantic features (based on LDA) work consistently best across all tasks, followed by profile features. Behavioral features and social network features were less useful.

Similarly, [Hong and Davison, 2010] compare the quality and effectiveness of different standard topic models for analyzing social data. Their results suggest that the best performance can be achieved by training a topic model on aggregated messages per user and for generating semantic annotations of individual messages topic features outperform simple TF-IDF weighted word features. However, for generating semantic annotations of users, TF-IDF based words perform best and topic features do not allow to improve the performance in the user classification task. While these results are quite strong, one important limitation is the data used. The models were created using only 274 users that were pre-selected by Twitter via http://twitter.com/invitations/suggestions (e.g., Health, Food&Drinks, Books). Given this pre-selection by Twitter, it is highly likely that these 274 users mainly discuss those specific topics, and are very popular. Therefore, the generalizability of these results is unknown.

Unlike the above work which focuses on individuals, [Bryden et al., 2013] examine how networks emerging from user communication are closely replicated in the frequency of words used within these communities. In short, users who are strongly connected also talk about similar subjects and therefore use similar words. In addition, [Bryden et al., 2013] also reveal that users who belong to one community tend to show similarities in the length of words they use or in their three letter word ending usage. This suggest that socio-linguistic features may be useful for differentiating users of different communities and therefore allow annotating them with their community membership or the topics associated with a community.

Recently, researchers started exploring different approaches for identifying experts on Twitter and predicting users' expertise areas. [Weng et al., 2010b] present TwitterRank, an adapted version of the topic sensitive PageRank, which allows identifying topical influential Twitter users based on follow relations and content similarities. [Pal and Counts, 2011] compare different network-based features and content/topical features to find authoritative users. [Canini et al., 2010] present an approach to find topically relevant Twitter users by combining standard Twitter text search mechanisms with information about the social relationships in the network. Previous research agrees on the fact that one needs both, content and structural network features, for creating a powerful expert retrieval algorithm. [Kim et al., 2010] use tweets of all users in a given Twitter list to discover characteristics and interests of the users in that list and compare the user interests with those that are perceived by human subjects in the user survey. Their user survey confirms that the words extracted from each set of lists are representative for all the members in the list even if the words are not used by those members.

Another line of research which shows promising results focuses on identifying experts by analyzing the structure and content of behavior-originated data (such as click-through data [Macdonald and White, 2009], search query logs [White et al., 2009] or social annotations [Kang and Lerman, 2011]). It is well-known that experts and novices tend to behave differently in various tasks, but little research exist on how these differences are reflected on social streams. For example [Kang and Lerman, 2011] defined several heuristics which may help to differentiate experts and novices by observing behavior originated data (in their case tags and pictures). Their work shows amongst others that the usage of overly-broad concepts (such as "misc" or "things") as well as the creation of conflicts in their own tagging vocabulary implies novice users.

Although previous research suggests that information about users' activities and how they are performed may help predicting user attributes such as their expertise topics, most of the existing research in the context of social media focuses on exploiting the content produced via user activities. In this thesis we extend existing research on creating semantic annotations of social media users by exploring the utility of activity patterns and linguistic style information for semantically annotating users with various types of semantic metadata such as expertise topics, profession types and personality related attributes.

2.2.2 Semantic Annotation of (Hash)Tag Streams

Hashtags and tags are free form annotations generated by humans and therefore lack explicit semantics. Previous research has shown that it is often hard to guess the meaning of a hashtag [Laniado and Mika, 2010]. Therefore, adding tags and hashtags to predefined and ontologically grounded semantic categories is useful since it makes the implicit semantics of hashtags explicit and therefore allows to easily infer the meaning of hashtags and relations between them.

[Overell et al., 2009] present an approach which allows classifying tags into semantic categories. The authors train a classifier to classify Wikipedia articles into semantic categories, map Flickr tags to Wikipedia articles using anchor texts in Wikipedia and finally classify Flickr tags into semantic categories by using the previously trained classifier. Their results show that their ClassTag system increases the coverage of the vocabulary by 115% compared to a simple WordNet approach which classifies Flickr tags by mapping them to WordNet via string matching techniques.

On Twitter, users have developed a tagging culture by adding a hash symbol (#) in front of a short keyword. The first introduction of the usage of hashtags was provided by Chris Messina in a blog post [Messina, 2007]. In [Huang et al., 2010], the authors say that this kind of new tagging culture has created a completely new phenomenon, called *micro-meme*. The difference between such micro-memes and other social tagging systems is that the participation in micro-memes is an *a-priori* approach, while other social tagging systems follow an *a-posteriori* approach. This is due to the fact that users are influenced by the observation of the usage of micro-meme hashtags adopted by other users. Therefore, the authors claim that users may produce a tweet to use a hashtag which they observed before and that it is likely that they would have never produced the tweet, if they would not have observed the hashtag before.

In social tagging systems the tagging process usually does not start with a tag, but with object which is being tagged. The work of [Huang et al., 2010] suggests that hashtagging in Twitter is more commonly used to join public discussions than to organize content for future retrieval. The role of hashtags has also been investigated in [Yang et al., 2012]. Their study confirms that a hashtag serves as both, a tag of content and a symbol of community membership.

[Laniado and Mika, 2010] explore to what extent hashtags can be used as strong identifiers like URIs are used in the Semantic Web. They measure the quality of hashtags as identifiers for the Semantic Web, defining several metrics to characterize hashtag usage on the dimensions of frequency, specificity, consistency, and stability over time. Their results indicate that the lexical usage of hashtags can indeed be used to identify hashtags which have the desirable properties of strong identifiers.

Although previous research has started to explore usage patterns of hashtags – e.g., the diffusion dynamics of hashtags [Tsur and Rappoport, 2012] and temporal spreading patterns of hashtags [Romero et al., 2011] – it remains unclear to what extent usage patterns of tags and hashtags may help to semantically ground them. In this thesis we partly address this gap by exploring the utility of hashtag's usage patterns for predicting their semantic categories.

2.2.3 Semantic Annotation of Individual Messages in Social Streams

Understanding and modeling the semantics of individual messages is important in order to support users in consuming social streams efficiently – e.g., via filtering social streams by users' interests or recommending tweets to users. However, one drawback of many state-of-the-art text mining approaches (such as *Bag of Words*) is that they suffer from the sparsity of microblog messages (i.e., the limited length of messages). Hence, researchers got interested in exploring those limitations and develop methods for overcoming them. Two commonly used strategies for improving short text classification are: (a) improving the classifier or feature representation and (b) using background knowledge for enriching sparse textual data.

Improving the classifier or feature representation: [Sriram et al., 2010] present a comparison of different text mining methods applied on individual Twitter messages. Similar to our work, they use a message classification task to evaluate the quality of the outcome of each text mining approach. Limitations of their work are that they only use 5 broad categories (news, opinions, deals, events and private message) in which they classify tweets. Further, they perform their experiments on a very small set of tweets (only 5407 tweets) which were manually assigned to the aforementioned categories. Their results show that authorship plays a crucial role since authors generally adhere to a specific tweeting pattern – i.e., a majority of tweets from the same author tend to be within a limited set of categories. However, the authorship feature requires that tweets of the same authors occur in the training and test dataset.

Latent semantic models such as topic models provide a method to overcome data sparsity by introducing a latent semantic layer on top of individual documents. [Hong and Davison, 2010] compare the quality and effectiveness of different standard topic models in the context of social streams and examine different training strategies. To assess the quality and effectiveness of different topic models and training strategies the authors use them in two classification tasks: a user and message classification task. Their results of the message classification task show that the overall accuracy for classifying messages into 16 general Twitter suggest categories (e.g., Health, Food&Drinks, Books) is almost twice as accurate when using topics as features rather than raw TF-IDF features. For the user classification task the contrary is the case – i.e., TF-IDF features are significantly better than topic features in the user classification task. This suggests that latent semantic models indeed provide a way to address sparsity and if sparsity is a problem, latent semantic models may achieve better results than bag of word models.

Enriching sparse textual data with background knowledge: In [Phan et al., 2008] the authors present a general framework to build classifiers for short and sparse text data by using hidden topics discovered from huge text collections. Their empirical results show that exploiting those hidden topics improves the accuracy significantly within two tasks: “Web search domain disambiguation” and “disease categorization for medical text”. [Hotho et al., 2003] present an extensive study on the usage of background knowledge from WordNet for enriching documents and show that most enrichment strategies can indeed improve the document clustering accuracy. However, it is unclear if their results generalize to the social media domain since the vocabulary mismatch between WordNet and Twitter might be bigger than between WordNet and news articles.

In this thesis a novel approach to overcome the sparsity of individual messages and to create semantic annotations of message by exploiting the background knowledge of the intended audience of a message is presented.

2.3 Emergent Usage Patterns in Social Streams

Usage patterns on social streams are regularities in human behavior which can be observed in online social environments and emerge if a large group of users shows similar behavior within a given context. For example, if users within a certain user group are far more likely to reply to messages which contain certain keywords than this is a usage pattern which is valid within a certain context. Another example of a usage pattern in social streams would be that messages authored by certain types of users (e.g. experts) are far more likely to stimulate discussions and get replies within a certain context.

In this thesis we focus on one specific type of human behavior in online social environments, namely the reply behavior. In the following sections we review research on predicting the number of replies a message will receive in the future which can be seen as proxy for popularity or attention and research on anticipating who will reply – i.e., conversation dynamics.

2.3.1 Popularity and Attention in Social Streams

The number of replies have been used as proxy measure for attention as well as for popularity in previous research. Within this context [Cheng et al., 2011] consider the problem of reciprocity prediction and study this problem in a communication network extracted from Twitter. They essentially aim to predict whether a user A will reply to a message of user B by exploring various features which characterize user pairs and show that features which approximate the relative status of two nodes are good indicators of reciprocity.

The work described in [Rangwala and Jamali, 2010] explores the task of predicting the rank of stories on Digg. They find that the number of early comments and their quality and characteristics are useful indicators. [Szabo and Huberman, 2010] study content popularity on Digg and YouTube. They demonstrate that early access patterns of users can be used to forecast the popularity of content and show that different platforms reveal different attention patterns. For example, while Digg stories saturate fairly quickly (about a day) to their respective reference popularities, YouTube videos keep attracting views throughout their lifetimes.

[Hong et al., 2011] investigate the problem of predicting the popularity of messages on Twitter measured by the number of future retweets. One of their findings is that the likelihood that a portion of a user's followers will retweet a new message depends on how many followers the user has and that messages which only attract a small audience might be very different from the messages which receive huge numbers of retweets. The work presented in [Naveed et al., 2011] explores the relation between the content properties of tweets and the likelihood of the tweets being retweeted. By analyzing a logistic regression model's coefficients, [Naveed et al., 2011] find that the inclusion of a hyperlink and using terms of a negative valence increase the likelihood of the tweet being retweeted. The work of [Macskassy and Michelson, 2011] explores the retweet behavior of Twitter users by modeling individual micro-cosm behavior rather than general macro-level processes. They present four retweeting models (general model, content model, homophily model, and recency model) and find that content based propagation models are better at explaining the majority of retweet behaviors in their data.

[Asur et al., 2011] explore the popularity of topics (concretely trending topics on Twitter) rather than individual messages and demonstrate empirically how factors such as user activity and number of followers, do not contribute strongly to trend creation and its propagation. In fact, they find that the resonance of the content with the users of the social network plays a major role in causing trends – i.e., the retweet activity of users' followers.

Although it is to assume that different communities of users which are created around different topics of interest can be identified on most social media applications, previous research did not explore differences in how attention is generated in these communities. This thesis closes this gap by exploring the idiosyncrasies of the reply-behavior of different topical communities which are created around different forums on the largest Irish message board, Boards.ie [5]. We also provide an extended set of semantic, structural and usage features to assess the effects that community and focus features have on reply behavior, something which has not been explored previously.

2.3.2 Conversation Dynamics in Social Streams

Previous research has focused on exploring how users use Twitter in general, and to what extent this platform is used for conversational purposes. For example, in one of the first papers about Twitter usage intention, [Java et al., 2007] find that Twitter is often used for discussing events of daily life, sharing information or URLs, reporting news and for conversations. Java et al. show that 21% of Twitter users participate in conversations, and 1/8 of all Twitter messages are part of conversations. They use the *@mention* sign as indicator for a conversation. [Macskassy, 2012] show that 92% of dialogues are between two people and that the average number of messages in dialogues is less than 5 tweets. [Honeycutt and Herring, 2009] evaluate conversations in Twitter and give insight about the nature of the *@mention* usage. They find that *@mention* is used in 90.96% of

[5] http://www.boards.ie/

cases for addressivity reasons, and that the median/mean number of users participating in conversations is 2/2.5. [Naaman et al., 2010] develop a content based categorization system for Twitter messages and find that most users focus on themselves (so-called "meformers") while less users are "informers".

Understanding the nature and dynamics of conversations on social media applications like Twitter was also subject of previous studies. For example, in [Chelmis and Prasanna, 2012] the authors explore the problem of predicting directed communication intention between users who did not communicate with each other before. The authors use various network and content features and conduct a link prediction experiment to assess the predictive power of those features. Their work is limited to predicting only new communication links between users.

[Sousa et al., 2010] explore if the reply behavior of Portuguese Twitter users and find that it is mainly driven by social factors rather than topical factors. For users with larger and denser ego-centric networks, they observe a slight tendency for separating their connections depending on the topics discussed. Their work focuses on three broad topics (sport, religion and politics) and therefore they only analyze the replies of messages which belong to one of these topics.

[Wang and Huberman, 2012] study the predictability of online interactions at the group and the individual level. They measure the predictability of online user behavior by using information-theoretic methods applied to real time data of online user activities from Epinions[6], a who-trust-whom consumer review site and Whrrl[7], a location based online social network game. Their work shows that the users' interaction sequences have strong deterministic components. In addition, they show that individual interactions are more predictable when users act on their own rather than when attending group activities. The work presented in [Chen et al., 2011] describes an approach for recommending interesting conversations to Twitter users. They use topic and tie strength between users and preferred thread length as factors to recommend conversations. Their ap-

[6] http://www.epinions.com/

[7] http://en.wikipedia.org/wiki/Whrrl

proach gives interesting insights about which conversations different types of users prefer.

Work described in [Rowe et al., 2011b] considers the task of predicting discussions on Twitter. The authors find that certain features are associated with increased discussion activity - i.e., the greater the broadcast spectrum of the user, characterized by in-degree and list-degree levels, the greater the discussion activity. Further, in [Rowe et al., 2011a] the authors explore factors which may impact discussions on message boards and show, amongst others, that content features are better indicators of seed posts than user features.

A special form of users' reply behavior is users' question answering behavior. [Paul et al., 2011] present a study of question asking and answering behavior on Twitter. They examine which characteristics of the asker might improve his/her chances of receiving a response. They find that the askers' number of followers and their Twitter account age are good predictors of whether their questions will get answered. However, the number of tweets the asker had posted or his/her frequency of use of Twitter do not predict whether his/her question will get answered. Finally, they examine the relationship between asker and replier and find that 36% of relationships are reciprocal and 55% are one-way. Surprisingly, 9% of answerers are not following the askers.

This thesis explores, similar to the work presented in [Sousa et al., 2010], to what extent social and topical factors may drive conversations on Twitter. However, unlike the work of [Sousa et al., 2010] which focuses on the Portuguese Twitter this thesis focuses on the English Twitter.

3 Papers

This cumulative dissertation consists of ten scientific publications. In the following I give a detailed overview about the contributions of other authors to this publications. Further, I describe the relation between different publications and the research questions of this thesis.

3.1 Contributions to the Scientific Publications

The following section details the contributions of other researchers to the scientific publications presented in this thesis.

The network theoretic model of social streams as well as the structural stream measures which were introduced in the following paper originate from lengthy discussions between the two authors and other members of our research lab. The dataset was crawled and analyzed by the author of this thesis.

- Claudia Wagner and Markus Strohmaier, The Wisdom in Tweetonomies: Acquiring Latent Conceptual Structures from Social Awareness Streams, Semantic Search Workshop at WWW2010, Raleigh, US, 2010.

The novel measures for assessing if and to what extent structures emerging from social streams become stable, have been developed by the author of this thesis. The empirical study about the stability of emerging structures in tag streams was conducted by the author of this thesis and Philipp Singer. The datasets were partly crawled by the author of this thesis and were partly available online.

- Claudia Wagner, Philipp Singer and Markus Strohmaier, Semantic Stabilization in Social Tagging Streams, under review, 2013.

The design of the experiments and evaluation of the results stem from extensive discussions between all participating authors. The datasets was available online[8] and the experiment was conducted by the author of this thesis.

- Claudia Wagner, Markus Strohmaier and Yulan He, Pragmatic metadata matters: How data about the usage of data effects semantic user models, Social Data on the Web Workshop co-located with 10th International Semantic Web Conference (ISWC2011), Bonn, Germany, 2011.

All authors contributed to the design of the experiment and the user study. The dataset for the experiment was crawled by the author of this thesis and the experiment was conducted by the author of this thesis. The user study was conducted by Vera Liao and the author of this thesis. The data collected from the user study was analyzed by Vera Liao.

- Claudia Wagner, Vera Liao, Peter Pirolli, Les Nelson and Markus Strohmaier, It's not in their tweets: Modeling topical expertise of Twitter users, ASE/IEEE International Conference on Social Computing (SocialCom2012), Amsterdam, The Netherlands, 2012.

The design of the experiments and evaluation of the results stem from extensive discussions between all participating authors. The dataset was crawled by Joshua Hailpern and Sitaram Asur. The feature engineering, classification experiments and evaluation were conducted by the author of this thesis.

- Claudia Wagner, Sitaram Asur, Joshua Hailpern, Religious Politicians and Creative Photographers: Automatic User Categorization in Twitter, ASE/IEEE International Conference on Social Computing (SocialCom2013), Washington D.C., USA, 2013.

The idea for this work and the experimental setup were elaborated by all participating authors. The dataset was crawled and analyzed by Lisa

[8] http://www.icwsm.org/2012/submitting/datasets/

Posch as part of her master thesis. The classification experiments and statistical tests were conducted by Philipp Singer and Lisa Posch. The audience-integrated topic models were trained by the author of this thesis and the correlation analysis between structural stream measures and classification performance was conducted by the author of this thesis.

- Claudia Wagner, Philipp Singer, Lisa Posch and Markus Strohmaier, The Wisdom of the Audience: An Empirical Study of Social Semantics in Twitter Streams, European Semantic Web Conference (ESWC), Montpellier, France, 2013.
- Lisa Posch, Claudia Wagner, Philipp Singer and Markus Strohmaier, Meaning as Collective Use: Predicting Hashtag Semantics on Twitter, 3rd Workshop on Making Sense of Microposts at WWW2013, Rio de Janeiro, Brazil, 2013.

The design of the experiments and interpretation of the results presented in the following two papers stem from extensive discussions between all participating authors. The datasets was available online[9]. The features were developed by the author of this thesis, while the classification and regression models were trained and tested by Matthew Rowe.

- Claudia Wagner, Matthew Rowe, Markus Strohmaier and Harith Alani, Ignorance isn't Bliss: An Empirical Analysis of Attention Patterns in Online Communities, ASE/IEEE International Conference on Social Computing (SocialCom2012), Amsterdam, The Netherlands, 2012. (Best Paper)
- Claudia Wagner, Matthew Rowe, Markus Strohmaier and Harith Alani, What catches your attention? An empirical study of attention patterns in community forums, The International AAAI Conference on Weblogs and Social Media (ICWSM2012), Dublin, Ireland, 2012.

The design of the experiment as well as the features were elaborated by the author of this thesis. The dataset was crawled by Johannes Schantl and the statistical tests and classification experiment were conducted by Johannes Schantl as part of his master thesis.

[9] http://www.icwsm.org/2012/submitting/datasets/

- Johannes Schantl, Claudia Wagner, Rene Kaiser and Markus Strohmaier, The Utility of Social and Topical Factors in Anticipating Repliers in Twitter Conversations, ACM Web Science (WebSci2013), Paris, France, 2013.

3.2 Emergent Structure

Social streams can be seen as complex adaptive systems – i.e., "*systems that have a large numbers of components, often called agents, that interact and adapt or learn*" [Holland, 2006]. Many different types of social streams exist and their structure is not necessarily predefined by system developers but emerges via user activities and is therefore a collectively-generated data structure that may go far beyond what the system designers' have envisioned. *Emergence* is what "self-organizing" processes produce [Corning, 2002]. According to Lewes [Lewes, 1879] the emergent is unlike its components in so far and it cannot be reduced to their sum or their difference. Also Kaufman [Kauffman, 1995] describes emergent phenomenons as something where "*the whole is greater than the sum of its parts*".

The objective of the following research question is to explore the structure of social streams and structural phenomenons which may emerge over time.

3.2.1 RQ 1: What is the emergent structure of social streams and how can we formally describe it?

In the first publication we introduce a network theoretic model of social streams which allows to formally describe the structure of social streams. Further, we introduce several structural stream measures and show how the model and the measures can be used to explore the nature of social streams.

In the second publication we discuss state-of-the-art methods for measuring if and to what extent social streams (or more specific tag streams) stabilize over time and highlight that not all methods are equally suitable for measuring the stability of social streams. Based on this discussion, we present two novel methods for measuring stabilization of social streams. Our results empirically show in two substantially different tagging systems that the tag distributions of objects also become stable if users are not exposed to the tags which have been previously assigned to the object. This result is striking since it suggests that imitation cannot be the only factor which causes the stable patterns which arise when a large group of users tag an object.

The Wisdom in Tweetonomies: Acquiring Latent Conceptual Structures from Social Awareness Streams

Claudia Wagner
JOANNEUM RESEARCH
Steyrergasse 17
8010 Graz, Austria
claudia.wagner@joanneum.at

Markus Strohmaier
Graz University of Technology and Know-Center
Inffeldgasse 21a
8010 Graz, Austria
markus.strohmaier@tugraz.at

ABSTRACT

Although one might argue that little wisdom can be conveyed in messages of 140 characters or less, this paper sets out to explore whether the *aggregation of messages* in social awareness streams, such as Twitter, conveys meaningful information about a given domain. As a research community, we know little about the structural and semantic properties of such streams, and how they can be analyzed, characterized and used. This paper introduces a network-theoretic model of social awareness stream, a so-called "tweetonomy", together with a set of stream-based measures that allow researchers to systematically define and compare different stream aggregations. We apply the model and measures to a dataset acquired from Twitter to study emerging semantics in selected streams. The network-theoretic model and the corresponding measures introduced in this paper are relevant for researchers interested in information retrieval and ontology learning from social awareness streams. Our empirical findings demonstrate that different social awareness stream aggregations exhibit interesting differences, making them amenable for different applications.

1. INTRODUCTION

In the last decade, the emergence of social media applications such as Wikipedia, Del.icio.us and Flickr has inspired a community of researchers to tap into user-generated data as an interesting alternative to knowledge acquisition. Instead of formally specifying meaning *ex-ante* through for example agreed-upon ontologies or taxonomies, the idea was to capture meaning from user-generated data *ex-post*.

With the emergence of social awareness streams, popularized by applications such as Twitter or Facebook and formats such as activitystrea.ms, a new form of communication and knowledge sharing has enriched the social media landscape. Personal awareness streams usually allow users to post short, natural-language messages as a personal stream of data that is being made available to other users. We refer to the aggregation of such personal awareness streams as *social awareness streams*, which usually contain a set of short messages from different users. Although one could argue that little wisdom can be conveyed in messages of 140 characters or less, this paper sets out to explore whether the *aggregation of messages* in different social awareness streams conveys meaningful information about a given domain.

Extracting structured knowledge from unstructured data is a well-known problem which has extensively been studied in the context of semantic search, because semantic search attempts to consider the meaning of users' queries and of available web resources. To extract the meaning of available web resources, different methods have been proposed which mainly rely on the content of web pages, their link structure and/or collaboratively generated annotations of web pages, so-called folksonomies. Social awareness streams provide a rich source of information, which can for example be used to improve semantic search by revealing possible meanings of a user's search query and by providing social annotations of web resources.

Since social awareness streams differ significantly from other information sources, such as web pages, blogs and wikis (e.g., through their lack of context and data sparseness), chats and newsgroups (e.g., through the way how information is consumed on social awareness streams, namely via social networks) and social tagging systems (e.g., through their structure and purpose), their applicability for knowledge acquisition and semantic search is still unclear. To address these differences and capture information structures emerging from social awareness streams we introduce the concept of a "tweetonomy", a three-mode network of social awareness streams.

This paper sets out to explore characteristics of different *social awareness stream aggregations* and analyzes if and what kind of knowledge can be extracted from social awareness streams through simple network transformations. The overall objectives of this paper are 1) to define a network-theoretic model of social awareness streams that is general enough to capture and integrate emerging usage syntax, 2) to define measures that characterize different properties of social awareness streams and 3) to apply the model together with the measures to study semantics in Twitter streams. Our experimental results show that different types of social awareness streams exhibit interesting differences in terms of the semantics that can be extracted from them. Our findings have implications for researchers interested in ontology learning and information retrieval from social awareness streams or general studies of social awareness streams.

The paper is organized as follows: First we introduce a network-theoretic model of social awareness streams as a tripartite network of users, messages and resources. Then, we propose several measures to quantify and compare differ-

WWW2010, April 26-30, 2010, Raleigh, North Carolina.
.

ent properties of social awareness streams. Subsequently, we characterize four different types of social awareness streams which have been aggregated from Twitter for a given search query *semantic web*, by computing several structural stream measures, such as the social and topical diversity of a stream. We investigate if and what kind of knowledge can be acquired from different aggregations of social awareness streams by transforming them into lightweight, associative resource ontologies. Finally, we relate our work to other research in this area and draw conclusions for future work.

2. SOCIAL AWARENESS STREAMS

Social awareness streams are an important feature of applications such as Twitter or Facebook. When users log into such systems, they usually see a stream of messages posted by those they follow in reverse chronological order. That means information consumption on social awareness streams is driven by explicitly defined social networks. Although messages in social awareness streams can be targeted to specific users, they are broadcasted to everyone who follows a stream and can be public or semi-public (i.e., only visible to users belonging to a user's social network).

Messages usually consist of words, URLs, and other user-generated syntax such as hashtags, slashtags or @replies. Hashtags are keywords prefixed by a hash (#) symbol which enrich short messages with additional (often contextual) information. Hashtags are, amongst others, used to create communication channels around a topic or event and to annotate term(s) with additional semantic metadata (e.g., #need[list of needs][1]). Slashtags[2] are keywords prefixed by a slash symbol (/) to qualify the nature of references in a message. So called @replies are usernames prefixed by an at (@) symbol and are used to mention users or target messages to them.

In contrast to other stream-based systems where data structures are formally defined by system developers (such as the tripartite data structure of folksonomies), social awareness streams are different in the sense that they have yielded an emerging, collectively-generated data structure that goes far beyond what the system designers' have envisioned. Emerging syntax conventions, such as RT (retweets), # (hashtags) or @ (replies), are examples of innovations by users or groups of users that superimpose an informal, emerging data structure on social awareness streams. This has made social awareness streams complex and dynamic structures which can be analyzed in a staggering variety of ways, for example, according to the author(s) of messages, the recipients of messages, the links, keywords or hashtags contained in messages, the time stamps of messages or the message types.

2.1 Tweetonomy: A Tripartite Model of Social Awareness Streams

Based on the existing tripartite structure of *folksonomies* [14] [7] [16] [5], we introduce a tripartite model of social awareness streams, a so-called "tweetonomy", which consists of messages, users and content of messages.

While a taxonomy is a hierarchical structure of concepts developed for classification, a folksonomy refers to the emerging conceptual structure that can be observed when a large group of users collaboratively organizes resources. In a tweetonomy nobody classifies or organizes resources, but users engage in casual chatter and dialogue. Our motivation for introducing *tweetonomies* as a novel and distinct concept is rooted in our interest in knowledge acquisition from this new and different form of discourse, i.e. to explore whether we can acquire latent hierarchical concept structures from social awareness streams such as Twitter of Facebook.

To formally define emerging structures from social awareness streams we present the model of a tweetonomy and introduce qualifiers on the tripartite structure that allow to accommodate user generated syntax. We formally define a tweetonomy as follows:

Definition 1. *A tweetonomy is a tupel* $T := (U_{q1}, M_{q2}, R_{q3}, Y, ft)$*, where*

- *U, M and R are finite sets whose elements are called users, messages and resources.*
- *Qualifier q1 represents the different ways in which users can be related to a message. For example, a user can be the author of a message* (U_a)*, or a user can be mentioned in a message in a variety of ways, such as being mentioned via an @reply* ($U_@$)*, or being mentioned via slashtags*[3] *such as /via, /cc and /by, which can represented as* U_{via},U_{cc} *and* U_{by}*. Future syntax can be accommodated in this model by adding further types of relations between users and messages.*
- *Qualifier q2 represents the different types of messages M supported by a social awareness stream. Messages in social awareness streams can have different qualities depending on the system. For example, the Twitter API distinguishes between public broadcast messages* (M_{BC})*, conversational direct messages* (M_D)*, and retweeted messages* (M_{RT})*. Future syntax can be accommodated in this model by adding further message types.*
- *Qualifier q3 represents the different types of resources that can be included in a social awareness stream. Resources can be keywords* (R_k)*, hashtags* (R_h)*, URLs* (R_l) *or other informational content occurring in messages of a social awareness stream.*
- *Y is a ternary relation* $Y \subseteq U \times M \times R$ *between U, M, and R.*
- *ft is a function which assigns to each Y a temporal marker,* $ft : Y \to \mathbb{N}$.

If we mention U, M or R without any qualifier, we refer to the union of all qualified sets of them. According to the definition, we use U_a to refer to the set of users who authored messages of stream, U_m to refer to the set of users who are mentioned in messages of a stream and R_h, R_k and R_l to refer to the set of resources in messages which are hashtags, keywords and URLs.

To define and characterize social awareness streams as well as individual messages, we can use the tripartite model to represent them as a tuples of users, messages and resources. For example, the following Twitter message: "*RT@tim new blog post: http://mydomain.com #ldc09*" created by a user *alex* can formally be represented by the tweetonomy shown in Figure 1.

[1] http://epic.cs.colorado.edu/helping_haiti_tweak_the_twe.html

[2] http://factoryjoe.com/blog/2009/11/08/

[3] http://factoryjoe.com/blog/2009/11/08/

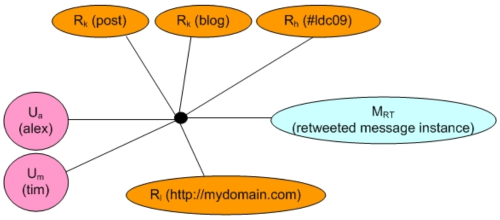

R_k (post)
R_k (blog)
R_h (#ldc09)
U_a (alex)
M_{RT} (retweeted message instance)
U_m (tim)
R_l (http://mydomain.com)

Figure 1: Example of a simple tweetonomy

2.2 Aggregations of Social Awareness Streams

The tripartite structure provides a general model to distinguish different aggregations of social awareness streams. Depending on the task and scope of investigations, researchers usually have to make choices about which aspects of social awareness streams to study. By making these choices, they usually produce different aggregations of the stream of data, that capture different parts and dynamics of streams. The introduced tripartite model allows to make these choices explicit.

In the following, we use the tweetonomy model to 1) define a subset of different aggregations of social awareness streams and 2) to demonstrate the nature and characteristics of different aggregations. Based on the model, we can distinguish between three basic aggregations of social awareness: resource streams $S(R')$, messages streams $S(M')$ and user streams $S(U')$. They are defined in the following way:

A resource stream $S(R')$ is a tupel $S(R') = (U, M, R, Y', ft)$, where $Y' = \{(u, m, r) \,|\, r \in R' \vee \exists r' \in R', \tilde{m} \in M, u \in U : (u, \tilde{m}, r') \in Y\}$ and $R' \subseteq R$ and $Y' \subseteq Y$. In words, a resource stream consists of all messages containing one or several specific resources $r' \in R'$ (e.g. a specific hahstag, URL or keyword) and all resources and users related with these messages.

A user stream $S(U')$ is a tupel $S(U') = (U, M, R, Y', ft)$, where $Y' = \{(u, m, r) \,|\, u \in U' \vee u' \in U', \tilde{m} \in M, r \in R : (u', \tilde{m}, r) \in Y\}$ and $U' \subseteq U$ and $Y' \subseteq Y$. In words, a user stream contains all messages which are related with a certain set of users $u \in U'$ and all resources and further users which are related with these messages. On Twitter, examples of user stream aggregations include user lists and user directory streams. User list and user directory stream aggregation contain all messages which have been authored by a defined set of users and all resources and users related with these messages. While user list streams are maintained by the user who has created the list, user directory streams, such as the one provided by wefollow[4], allow users to add themselves to existing or new lists.

A message stream $S(M')$ is a tupel $S(M') = (U, M, R, Y', ft)$, where $Y' = \{(u, m, r) \,|\, m \in M'\}$ and $M' \subseteq M$ and $Y' \subseteq Y$. In words, a message stream contains all messages of a certain type (e.g. conversational direct messages or retweeted messages) and their related resources and users.

In addition all streams can be restricted to a specific time window in which the stream is recorded. For example, $S(M')[t_s, t_e]$ denotes a message stream recorded within the time window t_s and t_e. Formally $S[t_s, t_e]$ can be defined as follows: $S[t_s, t_e] = (U \times M \times R, Y, ft)$, where $ft : Y \rightarrow \mathbb{N}, t_s \leq ft \geq t_e$.

2.3 Properties of Social Awareness Streams

Since for a given keyword (e.g., *semantic web*) different types of social awareness stream aggregations (e.g., the *semanticweb* hashtag or keyword stream or various user directory or user list streams denoted by the label *semanticweb*) can be analyzed, we introduce several stream measures in order to be able to compare different stream aggregations and quantify their differences. It appears intuitive that different aggregations of social awareness streams would yield different stream properties and characteristics. However, as a community we know little about *how* our aggregation decisions influence what we can observe. For example: What kind of streams are most suitable to identify links to web resources or hashtags for a given user query? What kind of streams and what kind of network transformations are most suitable for identifying synonyms or hyponyms (e.g. for hashtags)? What kind of streams are effective for identifying experts for a given topic? What kind of streams are topically diverse vs. topically focussed and narrow?

In the following, we introduce a number of measures that can be applied to different aggregations of social awareness streams in order to answer such questions, and to enable a quantitative comparison of different stream aggregations.

2.3.1 Social Diversity

The social diversity of a stream measures the variety and balance of users authoring a stream, i.e. the social variety and social balance of a stream. The Stirling measure [20] captures three qualities of diversity: variety (i.e., how many individual users participate in a stream), balance (i.e., how evenly the participation is distributed among these users), and similarity (i.e., how related/similar those users are). That means, although we do not use the concepts of similarity yet, the proposed diversity measures could be extended by including the concept of similarity.

To measure the social variety we can count the number of unique users $|U_a|$ who authored messages in a stream. For normalization purposes we can include the stream size $|M|$. The social variety per message $SVpm$ represents the mean number of different authors per message and is defined as follows:

$$SVpm = \frac{|U_a|}{|M|} \tag{1}$$

The maximum social variety $SVpm$ of a social awareness stream is 1. A social variety $SVpm$ of 1 indicates that every message has been published by another user. The social variety can also be interpreted as a function which illustrates how the number of authors in a stream grows over time and with increasing number of messages. For example, the interpretation of the social variety over time is defined as follows:

$$SVpt(t) = \frac{|U_a[t]|}{|M[t]|} \tag{2}$$

The variable $|M[t]|$ represents the number of messages within the time interval t and $|U_a[t]|$ denotes the number of authors of these messages.

To quantify the social balance of a stream, we can define an entropy-based measures, which indicates how democratic a stream is. Specifically, we call the distribution of authors U_a for messages M of a given stream, $P(M|U_a)$. Given this

[4] http://wefollow.com

number, we define the social balance of a stream as follows:

$$SB = -\sum_{u \in U_a} P(m|u) * log(P(m|u)) \tag{3}$$

A low social balance indicates that a stream is dominated by few authors, i.e. the distribution of messages per author is not even. A high social balance indicates that the stream was created in a balanced way, i.e. the distribution of messages per author is even.

For example, if on a stream author A has published 3 messages, author B has published 1 message and author C has as well published 1 message in a stream, we would say the social balance of this stream is equal to:

$$SB = -\frac{3}{5} * log(\frac{3}{5}) - \frac{1}{5} * log(\frac{1}{5}) - \frac{1}{5} * log(\frac{1}{5}) \approx 1.37 \tag{4}$$

2.3.2 Conversational Diversity

The conversational diversity of a stream measures how many users communicate via a stream and can be approximated via the conversational variety and conversational balance of a stream. To measure the number of users being mentioned in a stream (e.g., via @replies or slashtags), we can introduce $|U_m|$ for $u_m \in U_m$. The conversational variety per message $CVpm$ represents the mean number of different users mentioned in one message of a stream and is defined as follows:

$$CVpm = \frac{|U_m|}{|M|} \tag{5}$$

The conversational variety can in the same way as the social variety be interpreted as a function over time and message. The conversational balance of a stream can be defined in the same way as the social balance, as an entropy-based measure (CB) which quantifies how predictable conversation participants are on a certain stream.

2.3.3 Conversational Coverage

From the number of conversational messages $|M_c|$ and the total number of messages of a stream $|M|$, we can compute the conversational coverage of a stream, which is defined as follows:

$$CC = \frac{|M_c|}{|M|} \tag{6}$$

The conversational coverage measures the mean number of messages of a stream that have a conversational purpose.

2.3.4 Lexical Diversity

The lexical diversity of a stream can be approximated via the lexical variety and lexical balance of a stream. To measure the vocabulary size of a stream, we can introduce $|R_k|$, which captures the number of unique keywords $r_k \in R_k$ in a stream. For normalization purposes, we can include the stream size ($|M|$). The lexical variety per message $LVpm$ represents the mean vocabulary size per message and is defined as follows:

$$LVpm = \frac{|R_k|}{|M|} \tag{7}$$

In the same way as the social variety we can interpret the lexical variety as a function which illustrates the growth of vocabulary over time and with increasing number of messages. The lexical balance LB of a stream can, in the same way as the social balance, be defined via an entropy-based measures which quantifies how predictable a keyword is on a certain stream.

2.3.5 Topical Diversity

The topical diversity of a stream can be approximated via the topical variety and topical balance of a stream. To compute the topical variety of a stream, we can use arbitrary surrogate measures for topics, such as the result of automatic topic detection or manual labeling methods. In the case of Twitter we could use the number of unique hashtags $r_h \in R_h$ as surrogate measure for topics. The topical variety per message $TVpm$ represents the mean number of topics per message and is defined as follows:

$$TVpm = \frac{|R_h|}{|M|} \tag{8}$$

The topical variety can also be interpreted as a function which illustrates the growth of the hashtag vocabulary over time and with increasing number of messages. The topical balance TB can, in the same way as the social balance, be defined as an entropy-based measures which quantifies how predictable a hashtag is on a certain stream.

2.3.6 Informational Diversity

The informational diversity of a stream can be approximated via the informational variety and informational balance of a stream. To measure the informational variety of a stream, we can compute the number of unique links in messages of a stream $|R_l|$ for $r_l \in R_l$. The informational variety per message $IVpm$ is defined as follows:

$$IVpm = \frac{|R_l|}{|M|} \tag{9}$$

In the same way as the social variety measure, the informational variety measure can be interpreted as a function which illustrates how the number of different links shared via a stream grows over time and with increasing number of messages. The informational balance IB can, in the same way as the social balance, be defined as an entropy-based measures which quantifies how predictable a link is on a certain stream.

2.3.7 Informational Coverage

From the number of informational messages $|M_i|$ and the total number of messages of a stream $|M|$ we can compute the informational coverage of a stream which is defined as follows:

$$IC = \frac{|M_i|}{|M|} \tag{10}$$

The informational coverage indicates how many messages of a stream have a informational character.

2.3.8 Spatial Diversity

The spatial diversity of a stream measures the variety and balance of geographical message annotations in a stream, i.e. the spatial variety and spatial balance of a stream. The more spatial diverse a stream is the more messages it contains which were published on different locations and the more even the message distribution is across these locations. The spatial variety per message $SPVpm$ of a stream is defined via the number of unique locations of messages in a stream $|L|$

and the number of messages $|M|$ and is defined as follows:

$$SPVpm = \frac{|L|}{|M|} \quad (11)$$

In the same way as the social variety measure, the spatial variety measure can be interpreted as a function which illustrates how the number of different geo-locations grows over time and with increasing number of messages. The spatial balance *SPB* can, in the same way as the social balance, be defined as an entropy-based measures which quantifies how balanced messages are distributed across these geo-locations.

2.3.9 Temporal Diversity

The temporal diversity of a stream can be approximated via the temporal variety and temporal balance of a stream. The more temporal diverse a stream is the more messages it contains which were published at different moment in time and the more even the message distribution is across these timestamps. The temporal variety per message *TPVpm* of a stream is defined via the number of unique timestamps of messages $|TP|$ and the number of messages $|M|$ in a stream and is defined as follows:

$$TPVpm = \frac{|TP|}{|M|} \quad (12)$$

In the same way as the social variety, the temporal variety measure can be interpreted as a function which illustrates how the number of different timestamps grows over time and with increasing number of messages. The temporal balance *TPB* can, in the same way as the social balance, be defined as an entropy-based measures which quantifies how balanced messages are distributed across these message-publication-timestamps.

3. METHODOLOGY AND EXPERIMENTAL SETUP

To explore the nature of different social awareness stream aggregations which can be created for a given keyword and semantic models emerging from them, we conducted the following experiments. We studied different social awareness streams for the topic *semantic web* which were all recorded within the same time window. We investigated stream properties and semantics by adopting the introduced measures and by applying various network transformations.

Since measures for similarity and relatedness are not well developed for three-mode networks yet, the tripartite structure is often reduced to 3 two-mode networks with regular edges. These 3 networks model the relations between resources and users (N_{RU} network), resources and messages (N_{RM} network) and messages and users (N_{MU} network). To avoid subsubscriptions from now on we use RM, RU and MU instead of N_{RM}, N_{RU} and N_{MU}.

For example, the resource-user network RU can be defined as follows: $RU = (R \times U, E_{ru}), E_{ru} = \{(r,u) \,|\, \exists i \in I : (r,u,i) \in E\}, w : E \rightarrow \mathbb{N}, \forall e = (r,u) \in E_{ru}, w(e) := |i : (r,u,i) \in E|$. In words, the two-mode network RU links users to the resources that they have used or with which they have been mentioned in at least one message. Each link is weighted by the number of times a user has used or has been mentioned with that resource. The RU network can be represented as a matrix of the form $RU = v_{ij}$ where $v_{ij} = 1$ if user u_i is related with resource r_j. Since the resource-user network RU is an unqualified network, several qualified or semi-qualified networks (e.g. the resource-author network RU_a or the hashtag-author network R_hU_a), which are specializations of the resource-user network, can be deduced.

The resource-message RM and message-user MU networks are defined in the same way as the resource-user network: In words, the two-mode network RM links resources to messages in which they have been used at least once. Each link is weighted by the number of times a resource was used in a message. The two-mode network MU links messages to users which have authored them or are mentioned in them. Each link is weighted by the number of times a message was related with a user.

In order reveal associations between resources, we extracted non-qualified resource-message networks RM and semi-qualified resource-author networks RU_a from different social awareness stream aggregations. By multiplying the corresponding two-mode network matrices with their transpose (e.g., $O_R(RM) = RM * RM^T$), we transformed them into non-qualified one-mode networks of resources ($O_R(RM)$ and $O_R(RU_a)$), which can be considered as lightweight, associative resource ontologies [16]. From these non-qualified resource ontologies, we extracted semi-qualified resource networks, namely resource-hashtag networks RR_h and resource-link network RR_l, which we again transformed into associative resource ontologies ($O_R(RR_h(RM))$, $O_R(RR_l(RM))$, $O_R(RR_h(RU_a))$ and $O_R(RR_l(RU_a))$). Different ontologies relate resources which occur in the same contexts of messages/users/hashtags/links and therefore tend to have similar meanings according to Harris' distributional hypothesis [2].

The qualities of different resource ontologies depend on the different ways they are created: For example the $O_R(RM)$ ontology relates resources which co-occur in different messages and weight their relations by the number of times they co-occur. That means, a strong association exists between two resources if they share a large percentage of messages, regardless whether these associations were created by the same users or not. The $O_R(RU_a)$ ontology relates resources which are used by the same users. Relations between resources are weighted by the number of individual users who have used both resources, regardless whether these resources were used in one or different messages of them. The $O_R(RR_h(RM))$ and $O_R(RR_h(RU_a))$ network weight relations between resources by the number of times they co-occur with common hashtags. That means, a strong association exists between two resources if they share a large percentage of hashtags. The $O_R(RR_l(RM))$ and $O_R(RR_l(RU_a))$ network weights relations between resources by the number of times they co-occur with common URLs. That means, between two resources exists a strong association if they share a large percentage of links. In the $O_R(RR_l(RM))$ and $O_R(RR_h(RM))$ network resources co-occur if they are related with the same message (regardless whether these resources were associated via one or several users), while in the $O_R(RR_l(RU_a))$ and $O_R(RR_h(RU_a))$ network resources co-occur if they have been authored by the same user (regardless whether these resources were used in one or several messages of one user).

Since the different qualities of resource ontologies heavily depend on the different two-mode networks from which they originate, we also compared different two-mode networks in terms of their most important resource rankings. As a reminder, in the resource-message network RM a resource is important if it occurs in many different messages, while in the resource-author network RU_a a resource is important if it is used by many different users. In the resource-hashtag networks, $RRh(RM)$ and $RRh(RU_a)$, a resource is important if it co-occurs with many different hashtags. In the resource-link networks, $RRl(RM)$ and $RRl(RU_a)$, the resource ranking depends on the number of different links with which a resource co-occurs. If for example a resource `#semanticweb` appears in 50 percent of all messages of a stream which have all been generated by one user, this resource would have a high rank in the resource-message RM network, but a very low rank in the resource-author RU_a network. If the resource `#semanticweb` occurs together with certain URL in a message, which was retweeted many times by different users, the resource `#semanticweb` would have a high rank in the resource-message RM and resource-author RU_a network, but a very low rank in the resource-link RR_l and resource-hashtag RR_h network.

To assess the quality of different two-mode networks we assumed that hashtags tend to be semantic richer than other resources, because hashtags are often used to add additional contextual information to messages. Under this assumption we were able to quantitatively assess the semantic richness of different two-mode networks by computing the number of hashtags which appear under the top n resources (for n=15, 50, 100).

3.1 Dataset

We analyzed and compared the following social awareness stream aggregations from Twitter which were all related to one topic, *semantic web*. The stream aggregations were recorded in 2 time intervals: from 16th of Dec 2009 to 20th of Dec 2009 and from 29th of Dec 2009 to 1st of Jan 2010. While the first time interval represents 4 "normal" days without specific events or disturbances, we included the second time window due to the occurrence of a particular event (New Years Day) to surface differences in different stream aggregations.

- The *semanticweb* hashtag stream[5] $S(R_h)$ is a resource stream which includes all public messages containing the resource `#semanticweb` and all resources and users related with these messages. $S(R_h)$ is defined as follows: $S(R_h) = (U, M, R, Y', ft)$, where $Y' = \{(u, m, r) \mid r \in \{\#semanticweb\} \vee \exists r' \in \{\#semanticweb\}, \tilde{m} \in M, u \in U : (u, \tilde{m}, r') \in Y\}$ where $R_h \subseteq R$ and $Y' \subseteq Y$.

- The *semanticweb* keyword stream[6] $S(R_k)$ consists of all public messages containing the keyword `semanticweb` and `semweb`, a common abbreviation, and all resources and users related with these messages. $S(R_k)$ is defined as follows: $S(R_k) = (U, M, R, Y', ft)$, where $Y' = \{(u, m, r) \mid r \in \{semanticweb, semweb\} \vee \exists r' \in \{semanticweb, semweb\}, \tilde{m} \in M, u \in U : (u, \tilde{m}, r') \in Y\}$ where $R_k \subseteq R$ and $Y' \subseteq Y$.

- The *semweb* user list stream[7] $S(U_{UL})$ is a user stream which contains all public messages published by users of the authoritatively defined *semweb* user list and all resources and users related with these messages. We have chosen this list, because of its high authority for the topic *semantic web*. The list was created by Stefano Bertolo (*user sclopit*[8]), who is a Project Officer at the European Commission in the field of Knowledge Representation and Content Management. At the time we crawled the list (23th of November 2009), 141 users $u \in U_{UL}$ were included. $S(U_{UL})$ is defined as follows: $S(U_{UL}) = (U, M, R, Y', ft)$, where $Y' = \{(u, m, r) \mid u \in U_{UL} \vee u' \in U_{UL}, \tilde{m} \in M, r \in R : (u', \tilde{m}, r) \in Y\}$ where $U_{UL} \subseteq U$ and $Y' \subseteq Y$.

- The *semanticweb* wefollow user directory stream[9] $S(U_{UD})$ is a user stream which contains all public messages of users of the collaboratively created *semanticweb* directory and all resources and users related with these messages. We have chosen this directory, because it contains a large number of users. At the time we crawled the directory (23th of November 2009) it consisted of 191 users $u \in U_{UD}$. $S(U_{UD})$ is defined as follows: $S(U_{UD}) = (U, M, R, Y', ft)$, where $Y' = \{(u, m, r) \mid u \in U_{UD} \vee u' \in U_{UD}, \tilde{m} \in M, r \in R : (u', \tilde{m}, r) \in Y\}$ where $U_{UD} \subseteq U$ and $Y' \subseteq Y$.

Stream	$\lvert M\rvert$	$\lvert U_a\rvert$	$\lvert U_m\rvert$	$\lvert R_k\rvert$	$\lvert R_h\rvert$	$\lvert R_l\rvert$
$S(R_h)$	156	60	41	182	103	111
$S(R_k)$	210	105	66	618	108	133
$S(U_{UL})$	2183	86	770	4683	544	898
$S(U_{UD})$	4544	139	1559	6059	805	1300

Table 1: Number of messages ($|M|$), authors ($|U_a|$), users ($|U_m|$), keywords($|R_k|$), hashtags ($|R_h|$), and links ($|R_l|$) mentioned in messages of hashtag $S(R_h)$, keyword $S(R_k)$, user list $S(U_{UL})$, and user directory $S(U_{UD})$ stream aggregations.

3.2 Properties of Different Twitter Streams

To analyze and compare different stream aggregations we computed serval basic stream properties (see Table 1) and previously defined stream measures (see Figure 2).

From Figure 2 we can see that both analyzed resource streams (i.e., the hashtag and keyword stream) have a slightly higher informational variety $IVpm$ and informational coverage IC than the analyzed user streams (i.e., user list and user directory streams). This result suggests that researchers who want to sample messages from social awareness streams that contain links would benefit from focusing on hashtag or keyword streams (as opposed to other types of streams).

Figure 2 also shows that both analyzed resource streams have a higher social diversity (which is reflected via the social variety ($SVpm$) and social balance (SB) measure) than the analyzed user streams. Specially, if we compare the social balance (SB) of different stream aggregations, we can see that the analyzed hashtag stream has a significant higher social balance. This indicates that hashtag streams may be more democratic than other types of streams, since the participation of different authors (i.e., the number of messages they produce) seems to be more balanced.

[5] `http://twitter.com/search?q=\%23semanticweb`

[6] `http://twitter.com/\#search?q=semanticweb`

[7] `http://twitter.com/sclopit/semweb`

[8] `http://twitter.com/sclopit`

[9] `http://wefollow.com/twitter/semanticweb`

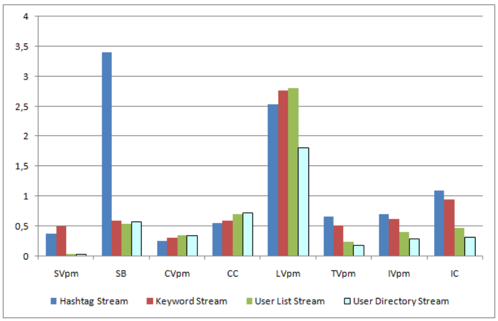


Figure 2: Social- (SVpm), Conversational- (CVpm), Lexical- (LVpm), Topical- (TVpm) and Informational (IVpm) Variety per message, Social Balance (SB), Informational- (IC) and Conversational Coverage (CC) of different Social Awareness Streams.

Since user list streams are closed and authoritatively defined sets of users, it seems plausible that these streams would contain less participants compared to open resource streams. However, the fact that the number of messages contained in the user directory stream is more than double the messages contained in the user list stream (although the number of authors is less than 40 percent higher) indicates that users registered in a user directories produce more messages. Since the social balance of the user directory stream is rather low, few authors seem to produce a major part of messages.

It is also interesting to note that the topical variety $TVpm$ is higher for the analyzed resource streams as for the analyzed user streams. Figure 2 shows that in a hashtag stream, more than every second message contains another hashtag (in addition to the one hashtag which is needed to assign the message to the hashtag stream), whereas the hashtag quota of other streams is lower.

3.3 Results

The aim of our empirical work was to explore if and what kind of knowledge can be acquired from different aggregations of social awareness streams by transforming them into lightweight, associative resource ontologies. The lightweight ontologies expose how related two resources are but do not contain any information about the semantics of relations.

In the following, we present our first results of analyzing emerging semantics conveyed by one user stream (the *semweb* user list stream $S(U_{UL})$) and one resource stream (the *#semanticweb* hashtag stream $S(R_h)$), which are both related with the topic *semantic web*.

Table 2 gives qualitative insights into the emerging semantics of different two-mode networks which were later transformed into resource ontologies. From Table 2 we can see that hashtag streams are in general rather robust against external events (such as New Years Eve), while user list stream aggregations are more perceptible to such “disturbances” (see Figure 3).

If we compare the 15 most important resources in different networks extracted from the same authoritative list of users (the *semweb* user list $S(U_{UL})$), we can observe that in all of them (except in one) the most important resources are mainly words which are not relevant for the topic *semantic web*. Only the resource-hashtag network $RR_h(RM)S(U_{UL})$ seems to be a positive exception and ranks resources (such as `#linkeddata, data, #goodrelations, #semanticweb, source, #distributed, link, #http, #rdf, page, great, web`) high, which are obviously relevant for the topic *semantic web*. This indicates that in a user stream of experts for a certain topic, resources which co-occur with many different hashtags tend to be very relevant for the expertise topic (or topic of common interest) of the group. A more detailed look into the most frequent hashtags of the analyzed user list stream (e.g., `#linkeddata, #semanticweb, #googrelations, #rdf, #rdfa`) confirms this assumption. One possible explanation for this phenomenon is that experts use a very fine-granular vocabulary to talk about their expertise topic and create a detailed hashtag vocabulary to add additional information to their messages and to assign them to appropriate communication channels.

The good quality of the $O_R(RR_h(RM))S(U_{UL})$ ontology, compared to other ontologies extracted from the user list stream aggregation, can amongst others be explained through hashtags' quality of revealing contextual information. Hashtags seem to be more appropriate for estimating the context of resources and identifying semantic similar resources via their common contexts.

For us it was surprising that URLs do not show similar characteristics as hashtags. At the beginning of our work we assumed that URLs might be as well a very appropriate context indicator. However, resource ontologies generated from resource-link networks do not reveal relevant concepts for the topic *semantic web*. These ontologies contain many general resources such as `type`, `source`, `blog` and `read`. These resources heavily occur with many common links, but do not reveal interesting knowledge about the stream aggregation topic.

4. DISCUSSION OF RESULTS

Our empirical findings confirmed our assumption that hashtag streams are in general rather robust against external events (such as New Years Eve), while user list stream aggregations are more perceptible to such “disturbances”. Nevertheless, it would be reasonable to assume that a stream of messages produced by experts in a given domain would result in meaningful semantic models describing resources within the domain and relations between them. Our findings however suggest that this is not necessarily the case. Not only are user list streams prone to external disturbances, the different types of network transformations also influence the resulting semantics.

Research on emerging semantics from folksonomies [16] showed that ontologies extracted from concept-instance networks (which are equivalent to resource-message networks in our model) are more appropriate for concept-mining than concept-user networks (which are equivalent to resource-user networks in our model), but ignore the relevance of individual concepts from the user perspective. Therefore, concept-instance networks might give an inaccurate picture of the community. This line of research would suggest to compute resource ontologies from resource-user networks rather than resource-message networks of social awareness stream aggregations in order to get an accurate picture of the community participating in the stream. Our results however indicate that resource-author networks (and ontologies gen-

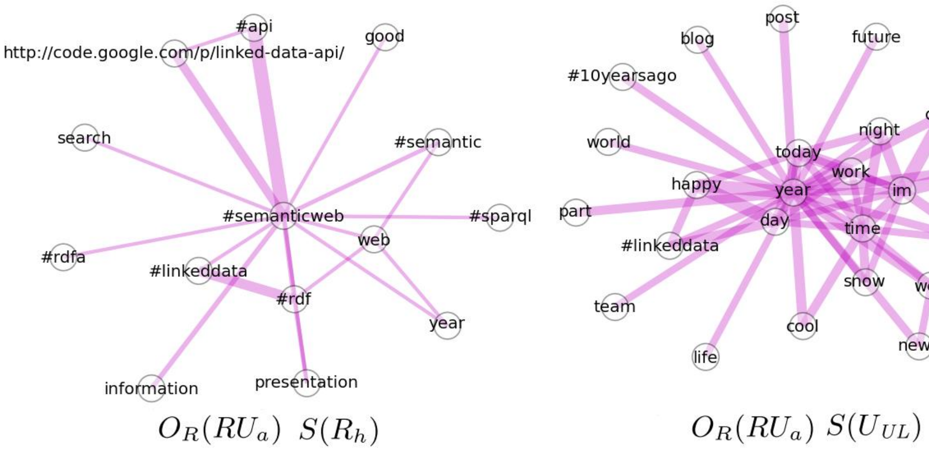


Figure 3: The resource ontology $O_R(RU_a)S(R_h)$ computed from the resource-user network of the `#semanticweb` hashtag stream shows an emerging semantic model which is able to describe the meaning of the stream‘s label *#semanticweb*, while the $O_R(RU_a)S(U_{UL})$ ontology computed from the resource-user network of the *semweb* user list stream shows that resource-user network transformations are perceptible for disturbances.

erated from them) are very prone to "disturbances", such as New Years Eve or the Avatar movie start, because these networks relate resources if they have common authors (regardless if they were used in one or several messages). Therefore, if for example all users post one message which contains happy new year greetings, resources such as `happy` or `year` become very important, although the majority of messages in this stream might be about *semantic web*. Our results indicate that hashtag-resource transformations have the power to reduce the non-informational noise of social awareness streams and reveal meaningful semantic models describing the domain denoted by the stream aggregation label (e.g., *semantic web*).

5. RELATED WORK

Semantic analysis of social media applications is an active research area, in part because on the one hand social media provide access to the "collective wisdom" of millions of users while on the other hand it lacks explicit semantics. Exploiting the "collective wisdom" of social media applications and formalizing it via ontologies, is therefore a promising and challenging aim of current research efforts.

Our work was inspired by Mika's work [16] who explored different lightweight, associative ontologies which emerge from folksonomies through simple network transformations. In general, automatic construction of term hierarchies and ontologies has been studied in both, the information retrieval and the semantic web communities: Sanderson and Croft describe in [18] the extraction of concept hierarchies from a document corpus. They use a simple statistical model for subsumption and apply it to concept terms extracted from documents returned for a directed query. Another line of research (e.g., [3]) suggests to use lexico-syntactic patterns (e.g., "such as") to detect hyponymy relations in text. Finally, the use of hierarchical clustering algorithms for automatically deriving term hierarchies from text was, amongst others, proposed in [1].

Since on the Social Web new data structures such as folksonomies (consisting of users, tags and resources) emerge, the extension and adaption of traditional content and link analysis algorithms and ontology learning algorithm became a key question. Markines et al. [15] define, analyze and evaluate different semantic similarity relationships obtained from mining socially annotated data. Schmitz et al. [19] describe how they mine from a tag space association rules of the form *If users assign the tags from X to some resource, they often also assign the tags from Y to them.* If resources tagged with t_0 are often also tagged with t_1 but a large number of resources tagged with t_1 are not tagged with t_0, t_1 can be considered to subsume t_0. Mika [16] presents a graph-based approach and shows how lightweight ontologies can emerge from folksonomies in social tagging systems. For mining concept hierarchies he adopts the set-theoretic approach that corresponds to mining association rules as is described by Schmitz et al.. Heymann at al. [4] represents each tag t as a vector (of resources tagged with the tag) and computes cosine similarity between these vectors. That means, they compute how similar the distributions of tags are over all resources. To create a taxonomy of tags, they sort the tags according to their closeness-centrality in the similarity graph. They start with an empty taxonomy and add a tag to the taxonomy as a child of the tag it is most similar to, or as a root node if the similarities are below a threshold.

In our past work we studied quantitative measures for tagging motivation [12] and found empirical evidence that the emergent semantics of tags in folksonomies are influenced by the pragmatics of tagging, i. e. the tagging practices of individual users [11]. This work as well was inspired by the hypothesis that the quality of emergent semantics (what concepts mean) depends on the pragmatics of users participating in a stream (how concepts are used).

	Top 15 resources
$RM\ S(R_h)$	#semanticweb, semantic, source, web, #linkeddata, twitter, #rdf, data, link, 2010, technology, present, #singularity, tool, #ontology
$RU_aS(R_h)$	#semanticweb, semantic, web, #rdf, #linkeddata, data, link, present, #sparql, good, 2010, technology, search, #semantic, http://code.google.com/p/linked-data-api/
$RR_h(RM)S(R_h)$	#semanticweb, source, #linkeddata, semantic, data, link, web, #rdf, 2010, twitter, present, http://ouseful.wordpress.com/2009/12/15, pipelink, http://code.google.com/p/linked-data-api/ , #sparql
$RR_h(RU_a)S(R_h)$	#semanticweb, semantic, web, #linkeddata, #rdf, data, link, http://code.google.com/p/linked-data-api/ , #semantic, #api, present, people, nobot, real, explain
$RR_l(RM)S(R_h)$	#semanticweb, source, semantic, twitter, web, #linkeddata, #rdf, tool, present, link, data, technology, #singularity, 800, entry
$RR_l(RU_a)S(R_h)$	#semanticweb, web, semantic, #rdf, tool, #linkeddata, 800, entry, exce, technology, list, source, #wiki, link, #owl
$RM\ S(U_{UL})$	type, year, data, good, #linkeddata, time, imo, 2010, source, great, make, web, work, day, watch
$RU_aS(U_{UL})$	year, make, great, 2010, work, web, day dont, happy, good, time, imo, interest, data, nice
$RR_h(RM)S(U_{UL})$	#linkeddata, data, #goodrelations, #semanticweb, source, #distributed, link, #http, #rdf, page, great, web, good, #bold, work
$RR_h(RU_a)S(U_{UL})$	year, make, happy, data, day, 2010, web, dont, great, interest, time, today, page, idea, future
$RR_l(RM)S(U_{UL})$	type, source, #semanticweb, #linkeddata, 2010, data, web, semantic, blog, state, post, make, new, twitter, read
$RR_l(RU_a)S(U_{UL})$	make, work, people, cool, time, read, thing, blog, new, book, help, language, change, talk, post

Table 2: Most important resources (ranked via their frequency) extracted from the resource-message (RM), the resource-author (RU_a), the resource-hashtag ($RR_h(RM)$ and $RR_h(RU_a)$), and the resource-link ($RR_l(RM)$ and $RR_l(RU_a)$) networks of a selected hashtag stream $S(R_h)$ and user list stream $S(U_{UL})$

Our work differs from existing work (1) through our focus on social awareness streams which have a more complex and dynamic structure than folksonomies and (2) through our focus on stream aggregations and data preprocessing. The aim of this work was to explore the initial step of building ontologies from social awareness streams, i.e. to explore how different stream aggregation and simple network transformations can influence what we can observe.

In general, little research on social awareness streams exists to date. Some recent research investigates user‘s motivation for microblogging and microblogging usage by analyzing user profiles, social interactions and activities on Twitter: A study by [8] shows that the rate of user activities on Twitter is driven by the social network of his actual friends. Users with many friends tend to post more updates than users with few friends. The work distinguishes between two different social networks of a user, the “declared” social network made up of followers and followees and the sparser and simpler network of actual friends. In [13], the authors performed a descriptive analysis of the Twitter network. Their results indicate that frequent updates might be correlated with high overlap between friends and followers. The work of [10] provides many descriptive statistics about Twitter use, and hypothesizes that the differences between users network connection structures can be explained by three types of distinct user activities: information seeking, information sharing, and social activity. In [21] an algorithm for identifying influential Twitter users for a certain topic is presented.

Other research focuses on analyzing content of social awareness stream messages, e.g. to categorize or cluster them or to explore conversations. For example, in [6] the authors examined the functions and usage of the @ (“reply/mention”) symbol on Twitter and the coherence of conversations on Twitter. Using content analysis, this line of work developed a categorization of the functional use of @ symbols, and analyzed the content of the reply messages. Recent research explores sentiments, opinions and comments about brands exposed on Twitter [9] and produces characterization of the content of messages of social awareness streams [17]. Naaman et al. examine how message content varies by user characteristics, personal networks, and usage patterns.

In the light of existing research and to the best of our knowledge, the network-theoretic model introduced in our paper represents the first attempt towards formalizing different aggregations of social awareness streams.

6. CONCLUSION AND FUTURE WORK

As our knowledge about the nature and properties of social awareness streams is still immature, this paper aimed to make following contributions: 1) We have introduce a network-theoretic model of social awareness streams, a so-called tweetonomy, which provides a formal, extensible framework capable of accommodating the complex and dynamic structure of message streams found in applications such as Twitter or Facebook. 2) We have defined and applied a number of measures to capture interesting characteristics and properties of different aggregations of social awareness streams and 3) Our empirical work shows that different aggregations of social awareness streams exhibit interesting different semantics.

While the network-theoretic model of social awareness streams is general, the empirical results of this paper are limited to a single concept (*semantic web*). It would be interesting to expand our analysis to a broader variety of social awareness streams and to conduct experiments over greater periods of time. For example, it seems plausible to assume that streams for hashtags such as #www2010 or #fun would differ significantly from a stream for the hashtag #semanticweb. We leave the task of applying our model to the analysis of a broader set of social awareness streams to future research. When it comes to the semantic analysis of social awareness streams, the extent to which different streams approximate the semantic understanding of users that are participating in these streams is interesting to investigate. While we have tackled this issue by selecting a narrow domain (*semantic web*), more detailed evaluations that include user feedback are conceivable. In addition, the semantic analysis conduced is based on simple network transformations. In future work, it would be interesting to study whether more sophisticated knowledge acquisition methods which, for example, exploit external background knowledge (such as WordNet[10] and DBpedia[11]) would produce different results. Another interesting issue raised by our investigations is the extent to which the semantics of social awareness streams are influenced by tweeting pragmatics of individual users or user groups and vice versa.

The network-theoretic model of this paper is relevant for researchers interested in information retrieval and ontology learning from social awareness streams. The introduced stream measures are capable of identifying interesting differences and properties of social awarness streams. Our empirical results provide evidence that *there is some semantic "wisdom" in aggregated streams of tweets*, but different stream aggregations exhibit different semantics and different extraction methods influence resulting semantic models: While some semantic models and aggregations of streams are rather robust against external events (such as New Years Day), other models and aggregations of streams are more perceptible to such "disturbances", and lend themselves to different purposes.

6.1 Acknowledgments

The research presented in this work is in part funded by the Know-Center and the FWF Austrian Science Fund Grant P20269 TransAgere. The authors want to thank Jan Pöschko and Florian Klien for their valuable comments on this work and Ronald Fellmann for supporting us with data aggregation issues.

7. REFERENCES

[1] P. Cimiano, A. Hotho, and S. Staab. Learning concept hierarchies from text corpora using formal concept analysis. *Journal of Artificial Intelligence Research (JAIR)*, 24:305–339, 2005.

[2] Z. Harris. Distributional structure. *The Structure of Language: Readings in the philosophy of language*, 10:146–162, 1954.

[3] M. A. Hearst. Automatic acquisition of hyponyms from large text corpora. In *Proceedings of the 14th conference on Computational linguistics*, pages 539–545, Morristown, NJ, USA, 1992. Association for Computational Linguistics.

[4] P. Heymann and H. Garcia-Molina. Collaborative creation of communal hierarchical taxonomies in social tagging systems. Technical Report 2006-10, Computer Science Department, April 2006.

[5] P. Heymann, G. Koutrika, and H. Garcia-Molina. Can social bookmarking improve web search? In *WSDM '08: Proceedings of the international conference on Web search and web data mining*, pages 195–206, New York, NY, USA, 2008.

[6] C. Honey and S. C. Herring. Beyond microblogging: Conversation and collaboration via twitter. In *System Sciences, 2009. HICSS '09. 42nd Hawaii International Conference on System Sciences*, 2009.

[7] A. Hotho, R. Jäschke, C. Schmitz, and G. Stumme. Bibsonomy: A social bookmark and publication sharing system. In *Proceedings of the Conceptual Structures Tool Interoperability Workshop at the 14th International Conference on Conceptual Structures*, pages 87–102, 2006.

[8] B. A. Huberman, D. M. Romero, and F. Wu. Social networks that matter: Twitter under the microscope. *ArXiv e-prints*, December 2008.

[9] B. J. Jansen, M. Zhang, K. Sobel, and A. Chowdury. Twitter power: Tweets as electronic word of mouth. *J. Am. Soc. Inf. Sci. Technol.*, 60(11):2169–2188, 2009.

[10] A. Java, X. Song, T. Finin, and B. Tseng. Why we twitter: understanding microblogging usage and communities. In *WebKDD/SNA-KDD '07: Proceedings of the 9th WebKDD and 1st SNA-KDD 2007 workshop on Web mining and social network analysis*, pages 56–65, New York, NY, USA, 2007.

[11] C. Körner, D. Benz, A. Hotho, M. Strohmaier, and G. Stumme. Stop thinking, start tagging: Tag semantics emerge from collaborative verbosity. In *19th International World Wide Web Conference (WWW2010).* ACM, April 2010.

[12] C. Körner, R. Kern, H. Grahsl, and M. Strohmaier. Of categorizers and describers: An evaluation of quantitative measures for tagging motivation. In *21st ACM SIGWEB Conference on Hypertext and Hypermedia (HT2010).* ACM, June 2010.

[13] B. Krishnamurthy, P. Gill, and M. Arlitt. A few chirps about twitter. In *WOSP '08: Proceedings of the first workshop on Online social networks*, pages 19–24, New York, NY, USA, 2008.

[14] R. Lambiotte and M. Ausloos. Collaborative tagging as a tripartite network, Dec 2005.

[15] B. Markines, C. Cattuto, F. Menczer, D. Benz, A. Hotho, and G. Stumme. Evaluating similarity measures for emergent semantics of social tagging. In *WWW '09: Proceedings of the 18th international conference on World wide web*, pages 641–650, New York, NY, USA, 2009.

[16] P. Mika. Ontologies are us: A unified model of social networks and semantics. *Web Semant.*, 5(1):5–15, 2007.

[17] M. Naaman, J. Boase, and C.-H. Lai. Is it all about me? user content in social awareness streams. In *Proceedings of the ACM 2010 conference on Computer supported cooperative work*, 2010.

[18] M. Sanderson and B. Croft. Deriving concept hierarchies from text. In *SIGIR '99: Proceedings of the 22nd annual international ACM SIGIR conference on Research and development in information retrieval*, pages 206–213, New York, NY, USA, 1999. ACM.

[19] P. Schmitz. Inducing ontology from flickr tags. In *Proceedings of the Workshop on Collaborative Tagging at WWW2006*, Edinburgh, Scotland, May 2006.

[20] A. Stirling. A general framework for analysing diversity in science, technology and society. *Journal of The Royal Society Interface*, 4(15):707–719, August 2007.

[21] J. Weng, E. peng Lim, J. Jiang, and Q. He. Twitterrank: Finding topic-sensitive influential twitterers. In *Third ACM International Conference on Web Search and Data Mining (WSDM 2010)*, 2010.

[10] http://wordnet.princeton.edu/

[11] http://dbpedia.org/

Semantic Stability in People and Content Tagging Streams

Claudia Wagner
JOANNEUM RESEARCH
8010 Graz, Austria
claudia.wagner@joanneum.at

Philipp Singer
Graz University of Technology
8010 Graz, Austria
philipp.singer@tugraz.at

Markus Strohmaier
U. of Koblenz & GESIS
50667 Cologne, Germany
strohmaier@uni-koblenz.de

Bernardo Huberman
HP labs
Palo Alto, CA 94304, USA
bernardo.huberman@hp.com

ABSTRACT

Social tagging systems have gained in popularity over the past few years. One potential disadvantage of social tagging systems is that users may never manage to reach a consensus on the description of the objects in the system, due to the lack of a centralized vocabulary. However, previous research has empirically shown that the tag distributions of objects become stable over time as more and more users tag them.

This work presents a critical review of the methods which have been used in previous research to assess semantic stability and proposes two novel methods which overcome the identified limitations. We empirically show in two substantially different tagging systems (a people-tagging and a content-tagging system) that our methods are more suitable for measuring semantic stability than existing methods and that the tag distributions of objects also become stable if users are not exposed to the tags which have been previously assigned to the object. This result is striking since it suggests that imitation cannot be the only factor which causes the stable patterns which arise when a large group of users tag an object.

1. INTRODUCTION

Social tagging systems, such as delicous, flickr, wefollow and others, are systems that allow users to collaboratively tag entities such as URLs, photos or people. Previous research has shown that social tagging systems exhibit social dynamics that yield interesting emergent properties, such as emergent structure or emergent semantics. A potential disadvantage of tagging systems is that due to the lack of a centralized vocabulary users may never manage to reach a consensus or may never produce a stable agreement on the description of objects. Stable description of objects are essential for attaining meaningful object interoperability across distributed systems and for maintaining the efficacy of object searching on local systems [19].

However, empirical studies on streams of tags show that the tag distributions of heavily tagged objects can become stable as more and more users tag them (see e.g., [11], [4] and [13]).

This work is structured as follows: We start by discussing state-of-the-art methods for measuring semantic stabilization in tag streams, including previous work by Golder and Huberman [11], Halpin et al. [13], Dellschaft and Staab [8] and others. We highlight that not all methods which have been used in previous work are equally suitable for measuring the semantic stability of tagging systems, and that some important limitations hinder progress towards a deeper understanding about the social-semantic dynamics involved. Based on this discussion, we present two novel methods for measuring semantic stabilization which overcome important limitations of previous work. Next, we apply our methods to a series of social tagging datasets and show in what ways our method can assert semantic stabilization.

This work makes the following contributions:

- We identify a number of limitations in existing methods of asserting semantic stabilization in social tagging systems. Our discussion of these limitations then gives rise to the development of more valid methods for measuring semantic stabilization.
- We present empirical investigations in different types of social tagging systems (content and people[1] tagging systems), and show that also if imitation is unlikely to happen during the tagging process, tag distributions of objects become stable.

The results of this work are relevant for scientists interested in leveraging social tagging methods for semantic purposes (such as mapping or annotation) and for system designers interested in engineering semantic consensus in open social tagging systems.

2. RELATED WORK

There already exists an interesting body of work on how people assign tags to themselves and other people and what guides their decisions during tagging. In the organizational context the work of Muller [22] and Farrell et al. [9] explore

[1] we refer to user accounts here which may belong to individuals or groups

how people apply tags to other people. In [22], the author classifies the tag usage of users in a cooperate system and finds that users have a preference to apply tags related to expertise to themselves, and to apply tags related to roles to others.

Social tagging systems such as delicious, where users tag web sites, also have been heavily studied in the recent past. One of the first and most important works was presented by Golder and Huberman [11]. Their work empirically shows that the number of tags needed to describe an object consistently converges to a power law distribution as a function of how many tags it receives. Further, they find that the proportion of frequencies of tags within a given site stabilizes over time. The authors attribute the stabilization of the tag proportion of a given resource to the way how users choose tags for a resource. The intuition is that if users would choose tags randomly no stabilization could be observed (or at least not at an early stage).

While Golder and Huberman do not propose a measure to quantify the stabilization, Halpin et al. [13] suggest to use the Kullback-Leibler (KL) divergence between the tag distributions of a resource at different points in time to measure the stabilization. Their work shows that the KL divergence converges relatively quickly to a very small value (with a few outliers). However, it remains unclear what relatively quickly means and in addition it is also unclear which value is small enough to claim stability.

In this work we show that tag distributions produced by a *random* tagging process may also converge towards zero (and thereby seem to stabilize) as the number of tag assignments grows based on inference from existing methods. From that example, one can see that existing methods which focus on visual analytics of the convergence characteristics of tag distributions have certain limitations. We propose to compare the convergence characteristics of the tag distribution produced by a random tagging process with the convergence characteristics of the real tag distributions. This comparison allows to test the significance of the stability and to quantify to what extent the stability which can be observed goes beyond what one would expect.

To simulate the tagging process, Golder and Huberman [11] propose that the simplest model that results in a power law would be the classic Polya urn model. The first model that formalized the notion of new tags was proposed by Cattuto et al. [4]. They explore the utility of the Yule-Simon model [26] to simulate tagging data and conclude that it seems to be unrealistic that users choose to reinforce tags uniformly from a distribution of all tags that have been used previously. According to the authors it seems more realistic to assume that users tend to apply recently added tags more frequently than old ones. Therefore, they present a Yule-Simon model with a long-term memory.

In addition to exploring imitation behavior, previous research also started exploring the utility of background knowledge as an additional explanatory factor which may help to simulate the tagging process. Bollen and Halpin [3] conduct a user study to explore the impact of tag suggestions on the stabilization process. Their results show that power law forms regardless of whether tag suggestions are provided to the user or not. This indicates that background knowledge has a much stronger impact than imitation. While our methodology and dataset are different, our results from the empirical people tagging study in Twitter are in line with the results from their user study on content tagging. Dellschaft et al. [8] use word frequency distributions obtained from different text corpora as background knowledge and pick tags according to those frequencies. They show that combining background knowledge with imitation mechanisms improves the simulation results. Although their results are very strong, their evaluation has certain limitations since they only aim to reproduce the shape of the original tag distribution produced by humans with their simulation model, but do e.g., not evaluate the accuracy at rank N of their simulated tag distributions. Dellschaft et al. [8] as well Cattuto et al. [4] show that the low-rank part (between rank 1 and rank 7-10) of the ranked tag frequency curves exhibits a flatten slope which is typically not observed in systems strictly obeying Zipf's law [27]. Therefore, they argue that a model which can simulate this tagging-specific shape of a curve is suitable to explain the tagging process. However, recent work by Bollen and Halpin [3] questions that the flatten head of these distributions is a characteristic which can be attributed to the tagging process itself. Instead, it may only be an artifact of the user interface which suggests up to ten tags.

3. SEMANTIC STABILITY IN TAGGING SYSTEMS

Emergent semantics is a phenomenon where global semantics emerge from a society of agents and represent the common current semantic agreement [1]. In the following, we empirically explore to what extent global semantics emerge from tagging systems by studying the formation of stable patterns in people-tagging in Twitter and content-tagging in delicious. Tags are free-form words which user associate with an object. First, we use existing methods for measuring the stability of tag distributions and discuss their limitations. Second, we present two novel methods which overcome these limitations and show how these methods can be used to identify stability and semantic agreement in tagging systems which go beyond what one would expect from a random tagging process. Finally, we discuss the commonalities and idiosyncrasies of different tagging systems.

3.1 Datasets

3.1.1 People-Tags in Twitter

People tagging has emerged in organizations [22, 9] as well as on social media applications [14, 16] as a mean to organize people, share selectively content and collaboratively build and maintain profiles of individuals or groups of users. User lists (where list names can be interpreted as tags) have been introduced by Twitter as a mean for users to organize and group their contacts and browse and consume the content of selected groups of users. Furthermore, user lists can also be used for coping with a possible information overload on Twitter [7]. Unlike in content tagging systems like delicious, where the tagging process starts with the content the user wants to tag, the tagging process in Twitter starts with the tag (aka the user list name). This means that the user first creates a new tag and then looks for users she wants to assign this tag to. In Twitter, users are not provided with any tag or user suggestions when creating user lists.

Our people-tag dataset consists of a sample of highly tagged Twitter users (who are usually celebrities) and a sample of users who are less frequently tagged. We selected these users as follows: In our previous work [24] we crawled the 100 largest *Wefollow* directories for Twitter handles and the top 500 users for each of those directories. From this dataset we selected users randomly by using the listed_count attribute in their profile as a sampling criterion.

The first sample of users contains 100 randomly selected users which are mentioned in more than 10k lists (heavily tagged users). The second sample contains 100 randomly selected users which are mentioned in less than 10k lists and more than 1k lists (less frequently tagged users).

For each of these sample users we crawled the full history of lists to which the user was assigned. We do not know when a user was assigned to a list but we know the order in which a user was assigned to different lists. Therefore, we can study the tagging process over time by using consecutive list assignments as a proxy for time.

3.1.2 Content-Tags in Delicious

Content tagging functionalities which are provided by systems such as delicious, flickr or youtube, allow users to assign tags to web content. The tagging process usually starts with the content to tag and users associate free form words (aka tags) with the content. Incentives of users for tagging content include future retrieval, contribution and sharing, attracting attention, play and competition, self-presentation, and opinion expression [20]. In delicious, users are provided with tag suggestions based on the tags which have been previously assigned to the object being tagged.

We use the delicious dataset which was crawled by [12] between January 2004 and December 2005. The dataset consists of around 140 million tag assignments, around 17 million resources, around 532k users and around 2,4 million tags. The dataset shows a continuous growth in number of users, tags and resources. From this dataset we selected the 100 resources which were tagged by most users (between 14k and 4k users). Furthermore, we selected a random sample of 100 resources which were tagged by fewer users (between 1k and 4k).

3.2 Method 1: Stable Tag Proportions [11]

Golder and Huberman [11] analyze the relative proportion of tags assigned to a given object as a function of the number of tag assignments and find a stable pattern in which the proportions of each tag are nearly fixed. Usually this fixed proportion is reached after the first 100 or so tag assignments.

In our work we find that not only the tags of web sites, but also the tags of users, give rise to a stable pattern in which the relative proportions of each tag are nearly fixed (see Figure 1 and Figure 2). This indicates that although users keep creating new tags and assign them to objects, the proportion of the tags which are assigned to an object becomes stable.

Limitations.

Golder and Huberman's work suggests that the stability of tag proportions indicates that users have agreed on a certain vocabulary. However, we argue that also tag distributions produced by a random tagging process become stable as more tag assignments take place since the total sum of the tag frequency vector from which the relative tag proportions are computed increases. Therefore, the impact of a constant number of tag assignments decreases over time. That means in an initial stage the relative proportion of tags per object can easily be changed by assigning N random tags to an object, since the sum of the frequency vector is relatively low compared to N. After users keep adding more and more tags the sum of the frequency vector increases while N remains stable. Therefore, it is not surprising that over time the relative tag proportions stop changing (see *law of large number*).

However, the stable tagging patterns which are shown in Figure 1 and Figure 2 go beyond what can be explained by a random tagging model, since a random tagging model produces similar proportions for all tags (see Figure 3). Hence, small changes in the tag frequency vector are enough to change the order of the ranked tag (i.e., the relative importance of tags for the object which is tagged). For real tag distributions this is not the case since these tag distributions are distributions with heavy tails, which means that few tags are used far more often than most others. Therefore, the relative proportion of the tags assigned to an object will not be similar for all tags. We exploit these observations for defining our novel measure in Section 3.5.

3.3 Method 2: Stable Tag Distributions [13]

Halpin et al. [13] present a method for measuring the semantic stabilization by using the Kullback Leibler divergence between the tag distributions of one object at different points in time. The Kullback-Leibler divergence is also known as relative entropy or information divergence between two probability distributions.

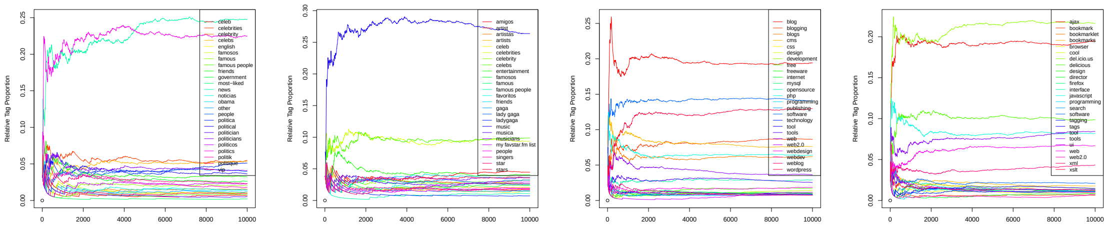
In their case tag distributions are defined as rank-ordered tag frequencies of the top 25 highest ranked unique tags per object. They use one month as a time window rather than using a fixed number of tag assignments as we do or Golder and Huberman [11] did. This is important since their measures, per definition, converge towards zero if the number of tag assignments is constant as we will show later.

Similar to Halpin et al. [13] we define the tag distribution as the rank-ordered tag frequencies of the top 25 highest ranked tags of each user or each website. In our case, the rank of a tag depends on how many users have assigned the tag to that users. Alternatively, we could also use the number of times a tag was assigned to a user. However, this ranking method would be prone to spammers. We use a constant number M of consecutive tag assignments and compare the KL divergence of tag distributions after N and $N + M$ consecutive tag assignments. Using a fixed number of consecutive tag assignments allows us to explore the properties of a random tag distribution which is generated by drawing M random samples from a uniform multinomial distribution.

In Figure 4, each point on the x-axis consists of $M = 10$ consecutive tag assignments. The black dotted line indicates the KL divergence of a random tag distribution. One can see from this figure that not only the tag distributions of users seem to converge towards zero over time (with few outliers), but also random tag distributions do.

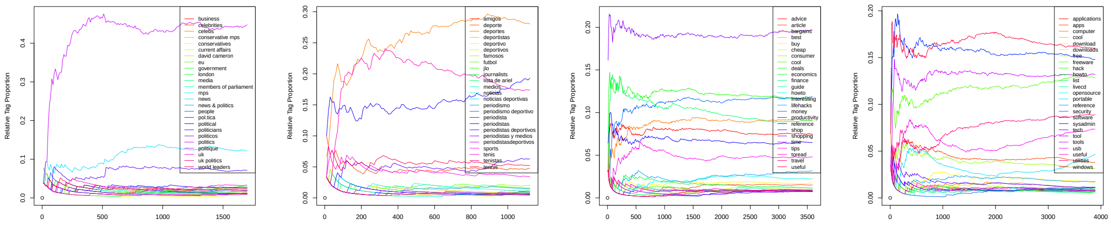
Limitations.

One needs to note that one single tag assignment in month j has more impact on the shape of the tag distribution of a resource than one single tag added in month $j + 1$, if we assume the number of tags which are added per month is



(a) Barack Obama (b) Lady Gaga (c) website id 75 (d) website id 86

Figure 1: Relative tag proportions of two heavily tagged objects (i.e., users or websites).



(a) David Cameron (b) Guille Salatino (c) website id 1929 (d) website id 1933

Figure 2: Relative tag proportion of two less frequently tagged objects (i.e., users or websites).

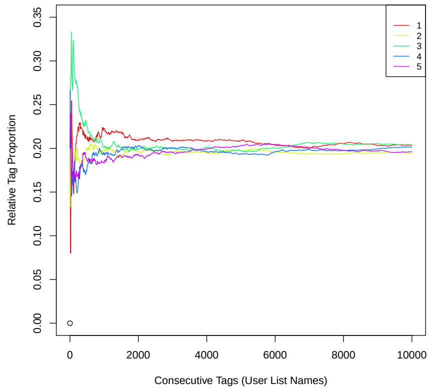


Figure 3: Relative tag proportion of a random tagging process where each tag assignment on the x-axis corresponds to picking one of the tags uniformly at random. In this example we have five tags which are represented by a number.

relatively stable over time. Only if the number of tag assignments per resource varies a lot across different months, convergence can be interpreted as semantic stabilization. This suggests that without knowing the distribution of tag assignments per month, the measure proposed by Halpin et al. is not useful since one never knows if stabilization can be observed due to the fact that users agreed on a certain vocabulary or due to the fact that the tagging frequency in later months was lower than in earlier months.

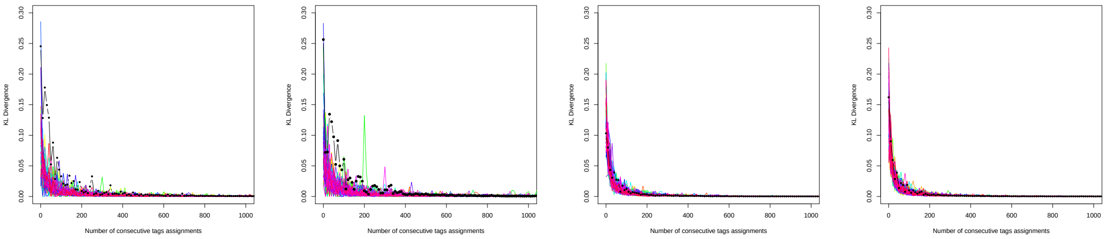
In our work, we propose to compare the KL divergence of a randomly generated tag distribution over time with the KL divergence of real tag distributions. This analysis reveals how much faster users reach consensus compared to what one would expect. Though this method already improves the original approach suggested by Halpin et al. [13], it is still limited since it does not reflect the intuition that it is more important that users agree on highly ranked tags than on lower ranked once. We will address this limitation with the new measure which we propose in section 3.5.

3.4 Method 3: Power Law Fits [21]

A power law distribution produced by tagging is a good sign of stability since, due to the scale invariance property of power law distributions, increasing the number of tagging instances only proportionally increases the scale of the power law, but does not change the parameters of the power law distribution. A power law distribution is defined by the function:

$$y = cx^{-\alpha} + \epsilon \qquad (1)$$

Both c and α are the constants characterizing the power law distribution and ϵ represents the uncertainty in the observed values. The most important parameter is the scaling



(a) Heavily tagged users (b) Less frequently tagged users (c) Heavily tagged websites (d) Less frequently tagged websites

Figure 4: KL divergence between the tag distributions at consecutive time points. Each colored line corresponds to one user, while the black dotted line depicts a randomly simulated tag distributions.

parameter α as it represents the slope of the distribution [3, 5]. It is also important to note that real world data nearly never follows a power law for the whole range of values. Hence, it is necessary to find some minimum value, which we call $xmin$. Based on this value we can then say that the tail of the distribution[2] follows a power law [5]. If we transform the power law distribution to the log-log scale, it takes the form of a linear function with slope α.

First we visualize the rank frequency tag distributions (see Figure 5) and the complementary cumulated distribution function (CCDF) of the probability tag distribution (see Figure 6) on a log-log scale for our objects at interest (i.e., users and websites). All figures show that for heavily tagged and less frequently tagged objects, few tags are applied very frequently while the vast majority of tags are used very rarely. Figure 6(a) and Figure 6(b) show that the tag distributions of heavily and less frequently tagged users are dominated by a large number of tags which are only used once. Heavily and less frequently tagged websites show a similar but less drastic pattern (cf. Figure 6(c) and Figure 6(d)). Further inspections of our data reveal that for both less frequently tagged users and heavily tagged users, around 86% of all tags are only used once, while for less frequently tagged websites and heavily tagged websites around 65% tags are used only once. Furthermore, we found less variance in the tag distributions of websites compared to the tag distributions of users. This indicates that the tagging process for websites seems to be very similar for different types of websites, while the tagging process for users seems to be more prone to the idiosyncrasies of users.

Figure 6 reveals the tail of the tag distributions (starting from a tag frequency 2) for both, users and websites, is close to a straight line. The straight line, which is a main characteristic for power law distributions plotted on a log-log scale, is more visible for heavily tagged objects than for less frequently tagged objects. Furthermore, the straight line seems to be more dominant for website tagging than for user tagging. We can now hypothesize that a power law distribution could be a good fit for our data if we look at the tail of the distribution with a potential $xmin \geq 2$.

In order to find the appropriate $xmin$ value we use the method of Clauset et al. [5] implemented by Alstott et al. [2] which finds this optimal value by selecting a value for $xmin$ for which the *Kolmogorov-Smirnov distance D* is minimized. For finding the scaling parameter α we use a *maximum likelihood estimation.* As suggested in previous work [3, 5], we also look at the Kolmogorov-Smirnov distance D of the corresponding fits.

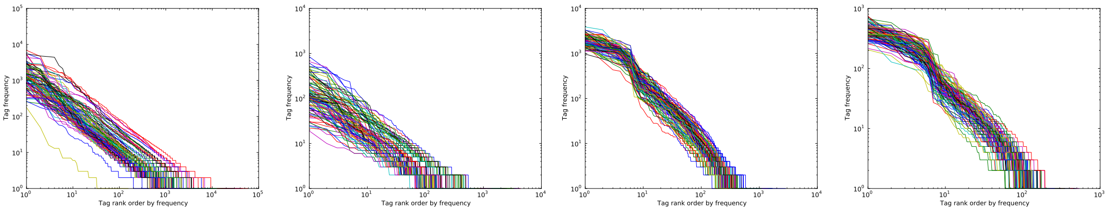
Table 1 shows the parameters of the best power law fits, averaged over all heavily tagged or less frequently tagged objects. One can see from this table that the α values are very similar for all datasets and also fall in the typical range of power law distributions. The low standard deviations indicate high similarity of the fits of distinct objects. Not surprisingly, the standard deviation is slightly higher for less frequently tagged objects. The parameters in table 1 indicate that the power law fits are slightly better for heavily tagged objects than for less frequently tagged objects as also suggested by Figure 5 and Figure 6.

Although our results suggest that its is likely that our distributions have been produced by a power law function, it is highly recommended to further investigate whether other heavy-tailed candidate distributions are better fits than the power law [5, 2]. Hence, we compare our power law fit to the fit of the *exponential function*, the *lognormal function* and the *stretched exponential (Weibull) function.* We use the *log-likelihood ratios* to indicate which fit is better.

The exponential function represents the absolute minimal candidate function to describe a heavy-tailed distribution. This means that if the power law function is not a better fit than the exponential function, we can hardly judge that the distribution is heavy-tailed at all. The lognormal and stretched exponential function represent more sensible heavy-tailed functions. Clauset et al. [5] and Alstott et al. [2] point out that there are only a few domains where the power law function is a better fit than the lognormal or the stretched exponential.

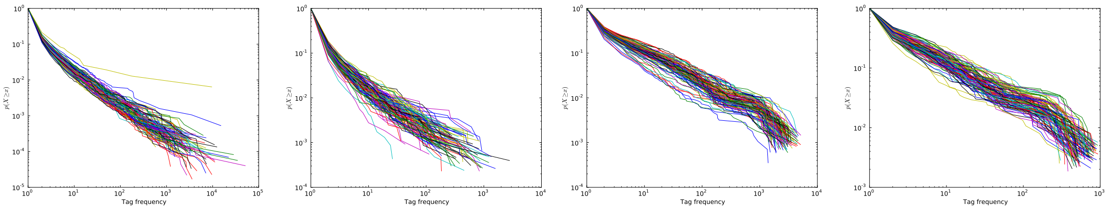
Our results confirm this since we do not find significant differences between the power law fit and the lognormal fit (for both heavily tagged and less frequently tagged users and websites). However, most of the time the power law function is significantly better than the stretched exponential function. Finally, we find that the power law function is a significantly better fit than the exponential function for all heavily tagged users, for most less frequently tagged users and for all websites (both heavily and less frequently

[2] note that we use the term *tail* to characterize the end of a distribution in the sense of probability theory



(a) Heavily tagged users (b) Less frequently tagged users (c) Heavily tagged websites (d) Less frequently tagged websites

Figure 5: Tag frequency plots for heavily tagged and less heavily tagged objects (i.e., users or websites) on log-log scale



(a) Heavily tagged users (b) Less frequently tagged users (c) Heavily tagged websites (d) Less frequently tagged websites

Figure 6: CCDF plots for heavily tagged and less frequently tagged objects (i.e., users or websites) on log-log scale

tagged). This indicates that the tag distributions of heavily tagged objects and most less frequently tagged objects are clearly heavy tail distributions and the power law function is a reasonable well explanation. However, it remains unclear from which heavy tail distribution the data have been drawn since several of them produce good fits.

Limitations.

As we have seen before, one limitation of this method is that it is often very difficult to determine which distribution has generated the data since several distributions with similar characteristics may produce an equally good fit. Furthermore, one needs to note that the automatic calculation of the best $xmin$ value for the power law fit has certain consequences since $xmin$ might become very large and therefore the tail to which the power law function is fitted may become very short. Finally, there is still a discussion going on among researchers about the informativeness of scaling laws [15], since some researchers suggest that many ways exist to produce scaling laws and some of those ways are idiosyncratic and artifactual [23, 18].

3.5 Novel Methods for Measuring Semantic Stability

In the following sections we propose two novel methods for measuring the semantic stability of tag distributions based on the benefits and drawbacks of existing methods which we discussed in the previous section. Our proposed methods incorporate two intuitions:

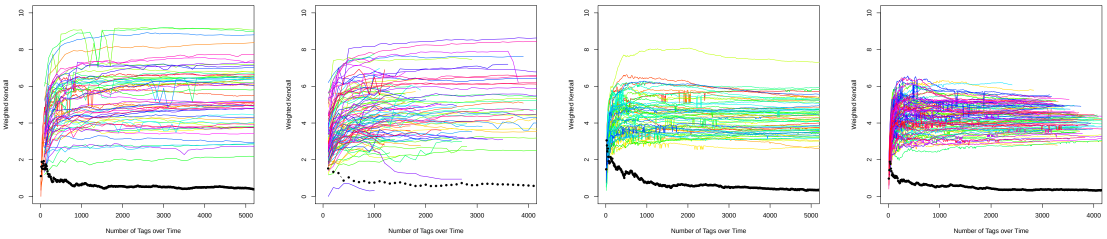
1. It is more important that the ranking of frequent tags remains stable than the ranking of less frequent tags since frequent tags are those which might be more relevant for an object. Frequent tags have been applied by many users and therefore stable patterns of these tags can be interpreted as “agreement”.
2. Semantic stability of a random tagging process needs to be included as a lower bound of stability since we are interested in exploring stable patterns which go beyond what can be explained by a random process.

3.5.1 Weighted Stable Tag Ranking: $K_{ws}(\sigma 1, \sigma 2)$

We propose to use a weighted rank agreement of tags per resource as a function of the number of consecutive tag assignments. Our intuition is that for higher ranked tags it

Table 1: Parameters of the best power law fits

	α	std	$xmin$	std	D	std
Heavily tagged users	1.9793	0.0841	4.5500	1.9818	0.0299	0.0118
Less frequently tagged users	2.0558	0.1529	3.1200	0.0570	0.0570	0.0218
Heavily tagged websites	1.6513	0.0839	4.5700	5.5718	0.0372	0.0104
Less frequently tagged websites	1.7131	0.1002	4.1400	2.8566	0.0558	0.0144



(a) Heavily tagged users (b) Less frequently tagged users (c) Heavily tagged web sites (d) Less frequently tagged web sites

Figure 7: Weighted version of Kendall τ. The black dotted line shows the weighted rank agreement of a random tagging process over time, while each colored line corresponds to one user or one website.

is more important that the ranking remains stable than for lower ranked tags, since higher ranked tags are those which are more relevant for the object being tagged. may have agreed on, in case semantic stabilization happens. We rank the tags for each resource according to their relative tag proportion after N tag assignments. Identical values (i.e., rank ties) are assigned a rank equal to the average of their positions in the ascending order of the values.

We use a weighted version of Kendall τ to measure the agreement between the ranking of tags after N and $N + M$ tag assignments. We measure difference between the total number of concordant pairs and discordant pairs and weight pairs proportionally to the product of the weights of the two elements being inverted or being stable [17]:

$$K_w(\sigma 1, \sigma 2) = \sum_{i,j:i<j} w_i w_j I_{ij} \qquad (2)$$

The variable I_{ij} is 1 if i and j is a concordant pair and -1 if it is a discordant pair. The intuition is that changing the rank of highly-relevant tags should result in a higher penalty than changing the rank of irrelevant tags. Therefore, we define the weight of each tag i as a function of it's rank in $\sigma 1$ as follows $w_i = \frac{1}{log(rank_i+1)}$. When using the log at base 2, the weight of tag at rank 1 is 1, while the weight of the tag at rank 3 is $\frac{1}{2}$ – i.e., the tag at rank 3 is only half as important as the tag at rank 1.

Further, while position weights address the question of swaps occurring near the beginning or the end of a ranked list of elements, many times the importance of the swap crucially depends on the similarity of the elements being swapped. In this work we use the distance between the relative proportion of two tags to define their similarity – i.e., two tags at different ranks are more similar if the distance between their relative tag proportions is small. The weighted and scaled version K_{ws} of Kendall's tau penalizes each inversion by the distance between the pair of elements inverted and the weight of the elements [17]:

$$K_{ws}(\sigma 1, \sigma 2) = \sum_{i,j:i<j} w_i w_j D_{ij} I_{ij} \qquad (3)$$

D_{ij} is the distance matrix $T \times T$ where T is the number of tags. We define the distance between tag t_1 and t_2 as the distance between the relative tag proportions: $|prop(t_1) - prop(t_2)|$.

For example, assume we have two tag distributions *(0.5, 0.4, 0.1)* and *(0.2, 0.6, 0.2)* over the following list of tags *(football, soccer, team)*. These distributions lead to the following ranked tag lists: *(football, soccer, team)* and *(soccer, team, football)*.

Since $football$ and $soccer$ (at rank 1 and 2) and $football$ and $team$ (at rank 1 and 3) are discordant pairs, but $soccer$ and $team$ (at rank 2 and 3) is a concordant pair, K_w is defined as $\frac{1}{log(3)} * \frac{1}{log(4)} - (\frac{1}{log(2)} * \frac{1}{log(3)} + \frac{1}{log(2)} * \frac{1}{log(4)})$. The scaled an weighted Kendall τ is defined as follows $K_{ws} = \frac{1}{log(3)} * \frac{1}{log(4)} * |0.4 - 0.1| - (\frac{1}{log(2)} * \frac{1}{log(3)} * |0.5 - 0.4| + \frac{1}{log(2)} * \frac{1}{log(4)} * |0.5 - 0.1|)$.

Figure 7 shows that the weighted rank agreement of users' and websites' tag distributions tend to become rather stable over time, while the rank agreement of randomly produced tag distributions slightly decreases and is in general lower than for real tag distributions. This indicates that our measure is suitable for identifying stable patterns in tag distributions and allows exploring to what extent they go beyond what one would expect.random tagging process and stability which may arise from real tagging data. We define the tag distribution of an object as the rank frequency distribution of the top 25 tags per object, since the proposed measure is only defined for conjoint lists of elements.

The comparison of users' (see Figure 7(a) and 7(b)) and websites' tag distributions (see Figure 7(c) and 7(d)) shows that there is less variance across different websites than across different users. Again, this indicates that the tagging process of different websites is more similar than the tagging process of different users. We hypothesize that users' idiosyncrasies or special happenings in their life may impact the tagging process and cause the variance which we observe.

Figure 7 also shows that the stable patterns are more easily visible for heavily tagged object than for less frequently tagged objects and that for most objects stability arrises after few hundred tag assignments.

3.5.2 Rank Biased Overlap: $RBO(\sigma 1, \sigma 2, p)$

The weighted and scaled Kendall τ measure which we presented in the previous section only allows to compare the agreement between conjoint lists of elements. A common practice is to limit the two lists to their top k elements and only compare the rank agreement between these conjoint sub-lists.

In the following we present an alternative method which allows to compute the weighted rank agreement between non-conjoint lists. The so called rank biased overlap (RBO) was developed by Webber et al. [25] and is defined as follow:

$$RBO(\sigma 1, \sigma 2, p) = (1-p) \sum_{d=1}^{\infty} \frac{\sigma 1_{1:d} \cap \sigma 2_{1:d}}{d} p^{(d-1)} \quad (4)$$

Let $\sigma 1$ and $\sigma 1$ two not necessarily conjoint lists of ranking. Let $\sigma 1_{1:d}$ and $\sigma 2_{1:d}$ be the ranked lists at depth d. The rank biased overlap falls in the range $[0, 1]$, where 0 means disjoint, and 1 means identical. The parameter p determines how steep the decline in weights is. The smaller p, the more top-weighted the metric is. In our work we empirically chose $p = 0.99$. We got similar results when choosing lower values of p, but more variation. The more top-weighted the metric, the more extreme the RBO values. However, the RBO of real tag distributions was always significantly higher than the RBO of random tag distributions.

Figure 8 and 9 show the rank biased overlap of the tag distributions of users and websites over time. The RBO value between the tag distribution after N tag assignments and after $N + M$ tag assignments is high, if the M new tag assignments do not change the ranking of the (top-weighted) tags. One can see from Figure 8 and Figure 9 that the RBO of a randomly generated tag distribution is pretty low and remains low as more and more tags are added over time. On the contrary, the RBOs of users' and websites' tag distributions increases as more and more tags are added. This indicates that the RBO measure allows identifying a consensus in the tag distributions which may emerge over time and which goes beyond what one would expect from a random tagging process. Again, we can see that for heavily tagged objects more agreement can be observed than for less frequently tagged objects and that the agreement for websites is in general slightly higher than for users.

4. DISCUSSION OF RESULTS

Our work highlights important limitations of existing methods for measuring semantic stability of tagging systems and introduces two novel methods which overcome the identified limitations. Our results empirically show that the proposed methods are suitable for measuring semantic stability and assessing if the stability goes beyond what one would expect if the tagging process would be a random process.

Overall the comparison of people and content tagging reveals that although the tagging process itself is different (e.g., in Twitter the tagging process starts with the tag, while in delicious it starts with the resource; further in Twitter users do not see which tags have been previously assigned to the user they are tagging during the tagging process, while in delicious users get tag suggestions based on what other users previously assigned to the object), the outcome (i.e., the tag distributions of objects) is surprisingly similar. We find that regardless of the type of object which is tagged and regardless of the visibility of tags which other users have been previously assigned to the object, the tag distributions of objects become stable. This result is striking since it suggests that imitation cannot be the only factor which causes the stable patterns that arise when a large group of users tag an object.

Although stable patterns arise from both tagging systems, we also found interesting differences when comparing the patterns from people and website tagging. First of all, we can see that the variance of distinct websites is lower than the variance of distinct users. That means, the stable patterns which we observe for websites are very similar for distinct websites, while the stable patterns of users may vary. One possible explanation for this higher variance across users might be the idiosyncrasies of different users (or types of users) or special happenings or events in their life.

Second, our results reveal that tags assigned to users include a high percentage of tags that are used only once. In contrast, there are fewer tags that are used only once for websites, which may indicate that tags assigned to people tend to be more idiosyncratic or personal than those tags assigned to websites.

This may not only be attributed to the different nature of the object being tagged, but also to the fact that on Twitter users do not get any tag suggestions, while on delicious they do get tag suggestions. As also shown by [3] the presence of tag suggestions may lead to tag distributions with a shorter long tail – i.e., tag suggestions may provoke a higher agreement between users who tag an object.

Finally, our results show that more heavily tagged objects (for both people and website tagging) tend to have more stable tag distributions.

5. CONCLUSIONS AND OUTLOOK

In this work, we present a critical review of existing methods for measuring semantic stability in tagging systems and propose two novel methods which overcome the identified limitations. Using data from two substantially different types of tagging systems (user lists in Twitter and bookmarks in delicious) allows us to conclude that our proposed measures are suitable for measuring semantic stability which goes beyond what one would expect from a random tagging process.

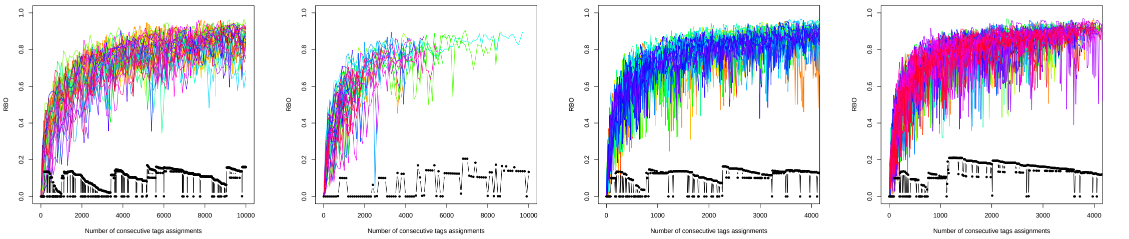
In future work we aim to extend our empirical investigations to further types of tagging systems and aim to investigate the factors which may cause the stabilization. Previous research mainly focused on exploring the imitation behavior of users as a potential cause for the stabilization. However, our own empirical results from the people-tagging study on Twitter, as well the experimental study of Bollen and Halpin [3] suggest that tagging systems become stable over time regardless of whether tag suggestions are provided to the user or not. Therefore, we aim to investigate background knowledge as well as the regularities and stability of natural language (see e.g., [27], [6] and [10]) as two alternative factors which may explain the stabilization process which can be observed in social tagging systems.

Acknowledgments

Claudia Wagner is a recipient of a DOC-fForte fellowship of the Austrian Academy of Science.

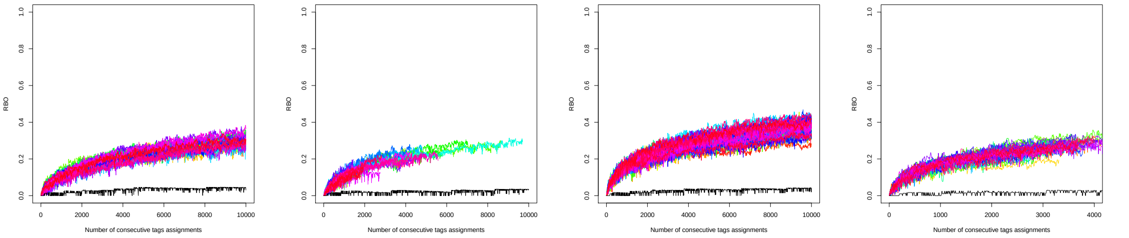
6. REFERENCES

[1] K. Aberer, P. Cudré-Mauroux, A. M. Ouksel, T. Catarci, M.-S. Hacid, A. Illarramendi, V. Kashyap, M. Mecella, E. Mena, E. J. Neuhold, O. D. Troyer, T. Risse, M. Scannapieco, F. Saltor, L. D. Santis, S. Spaccapietra, S. Staab, and R. Studer. Emergent semantics principles and issues. In Y.-J. Lee, J. Li, K.-Y. Whang, and D. Lee, editors, *Proceedings of the 9th International Conference on Database Systems for*



(a) Heavily tagged users (b) Less frequently tagged users (c) Heavily tagged websites (d) Less frequently tagged websites

Figure 8: Rank Biased Overlap (RBO) measures with $p = 0.9$. The black dotted line shows the weighted rank agreement of a random tagging process over time, while each colored line corresponds to one user or one website.



(a) Heavily tagged users (b) Less frequently tagged users (c) Heavily tagged websites (d) Less frequently tagged websites

Figure 9: Rank Biased Overlap (RBO) measures with $p = 0.99$. The black line shows the weighted rank agreement of a random tagging process over time, while each colored line corresponds to one user or one website.

Advanced Applications (DASFAA'04), volume 2973 of *Lecture Notes in Computer Science*, pages 25–38. Springer, 2004.

[2] J. Alstott, E. Bullmore, and D. Plenz. powerlaw: a python package for analysis of heavy-tailed distributions. 2013.

[3] D. Bollen and H. Halpin. The role of tag suggestions in folksonomies. In *Proceedings of the 20th ACM conference on Hypertext and hypermedia*, HT '09, pages 359–360, New York, NY, USA, 2009. ACM.

[4] C. Cattuto, V. Loreto, and L. Pietronero. Semiotic dynamics and collaborative tagging. *Proceedings of the National Academy of Sciences*, 104(5):1461–1464, 2007.

[5] A. Clauset, C. R. Shalizi, and M. E. J. Newman. Power-law distributions in empirical data. *SIAM Rev.*, 51(4):661–703, Nov. 2009.

[6] A. Cohen, R. N. Mantegna, and S. Havlin. Numerical analysis of word frequencies in artificial and natural language texts. *Fractals*, 1997.

[7] S. de la Rouviere and K. Ehlers. Lists as coping strategy for information overload on twitter. In *Proceedings of the 22nd international conference on World Wide Web companion*, WWW '13 Companion, pages 199–200, Republic and Canton of Geneva, Switzerland, 2013. International World Wide Web Conferences Steering Committee.

[8] K. Dellschaft and S. Staab. An epistemic dynamic model for tagging systems. In *HT '08: Proceedings of the nineteenth ACM conference on Hypertext and hypermedia*, pages 71–80, New York, NY, USA, 2008. ACM.

[9] S. Farrell, T. Lau, S. Nusser, E. Wilcox, and M. Muller. Socially augmenting employee profiles with people-tagging. In *Proceedings of the 20th annual ACM symposium on User interface software and technology*, UIST '07, pages 91–100, New York, NY, USA, 2007. ACM.

[10] R. Ferrer-i Cancho and B. Elvevåg. Random Texts Do Not Exhibit the Real Zipf's Law-Like Rank Distribution. *PLoS ONE*, 5(3):e9411+, Mar. 2010.

[11] S. Golder and B. A. Huberman. Usage patterns of collaborative tagging systems. *Journal of Information Science*, 32(2):198–208, April 2006.

[12] O. Görlitz, S. Sizov, and S. Staab. Pints: peer-to-peer infrastructure for tagging systems. In *Proceedings of the 7th international conference on Peer-to-peer systems*, IPTPS'08, pages 19–19, Berkeley, CA, USA,

2008. USENIX Association.

[13] H. Halpin, V. Robu, and H. Shepherd. The complex dynamics of collaborative tagging. In *Proceedings of the 16th international conference on World Wide Web*, WWW '07, pages 211–220, New York, NY, USA, 2007. ACM.

[14] S. Kairam, M. J. Brzozowski, D. Huffaker, and E. H. Chi. Talking in circles: Selective sharing in google+. In *Proceedings of the ACM Conference on Human Factors in Computing Systems (CHI '12)*, pages 1065–1074, New York, NY, 2012.

[15] C. T. Kello, G. D. A. Brown, R. Ferrer-i Cancho, J. G. Holden, K. Linkenkaer-Hansen, T. Rhodes, and G. C. Van Orden. Scaling laws in cognitive sciences. *Trends in Cognitive Sciences*, 14(5):223–232, May 2010.

[16] D. Kim, Y. Jo, I.-C. Moon, and A. Oh. Analysis of twitter lists as a potential source for discovering latent characteristics of users. In *Workshop on Microblogging at the ACM Conference on Human Factors in Computer Systems. (CHI 2010)*, 2010.

[17] R. Kumar and S. Vassilvitskii. Generalized distances between rankings. In *Proceedings of the 19th international conference on World wide web*, WWW '10, pages 571–580, New York, NY, USA, 2010. ACM.

[18] W. Li. Random texts exhibit zipf's-law-like word frequency distribution. *IEEE Transactions on Information Theory*, pages 1842–1845, 1992.

[19] G. Macgregor and E. McCulloch. Collaborative tagging as a knowledge organisation and resource discovery tool. *Library Review*, 55(5), in press.

[20] C. Marlow, M. Naaman, D. Boyd, and M. Davis. Ht06, tagging paper, taxonomy, flickr, academic article, to read. In *Proceedings of the seventeenth conference on Hypertext and hypermedia*, HYPERTEXT '06, pages 31–40, New York, NY, USA, 2006. ACM.

[21] A. Mathes. Folksonomies: Cooperative classification and communication through shared metadata. `http://www.adammathes.com/academic/computer-mediated-communication/folksonomies.html`, June 2004. Accessed: 2013-07-11.

[22] M. J. Muller. Comparing tagging vocabularies among four enterprise tag-based services. In *Proceedings of the 2007 international ACM conference on Supporting group work*, GROUP '07, pages 341–350, New York, NY, USA, 2007. ACM.

[23] A. Rappoport. *Zipf's law revisited.* Studienverlag Bockmeyer, 1982.

[24] C. Wagner, S. Asur, and J. Hailpern. Religious politicians and creative photographers: Automatic user categorization in twitter. In *we will see*, 2013.

[25] W. Webber, A. Moffat, and J. Zobel. A similarity measure for indefinite rankings. *ACM Trans. Inf. Syst.*, 28(4):20:1–20:38, Nov. 2010.

[26] G. U. Yule. A Mathematical Theory of Evolution, Based on the Conclusions of Dr. J. C. Willis, F.R.S. 213(402-410):21–87, Jan. 1925.

[27] G. K. Zipf. *Human behavior and the principle of least effort.* Addison-Wesley Press, 1949.

3.3 Emergent Semantic

Since social streams lack explicit semantics, adding semantics to streams is crucial for increasing the searchability and navigability of social streams. The objective of the following research question is to explore to what extent structural metadata and usage metadata can be exploited for semantically annotating social streams and to what extent novel social stream mining methods which go beyond existing text mining approaches could benefit from them.

3.3.1 RQ 2: To what extent may structural metadata and usage metadata contribute to acquiring emerging semantics from streams and annotating streams with semantics?

The following five publications address different aspects of this research questions and focus on user streams and hashtag streams. According to the *tweetonomy* model which was introduced in Chapter 3.2, a user stream $S(U')$ is defined as a tuple $S(U') = (U, M, R, Y', ft)$, where $Y' = \{(u, m, r) \,|\, u \in U' \vee u' \in U', \tilde{m} \in M, r \in R : (u', \tilde{m}, r) \in Y\}$ and $U' \subseteq U$ and $Y' \subseteq Y$. U define a set of users, U' is a specific subset of U, M is a set of messages, and Y defines a ternary relation $Y \subseteq U \times M \times R$ between U, M, and R. ft is a function which assigns to each Y a temporal marker, $ft : Y \rightarrow N$. In words, a user stream contains all messages which are related with a certain set of users $u \in U'$ (in our case the set consists of one selected user) and all resources and further users which are related with these messages. A hashtag stream $S(R_h)$ is defined as follows: $S(R_h) = (U, M, R, Y', ft)$, where $Y' = \{(u, m, r) \,|\, r \in \{R_h\} \vee \exists r' \in \{R_h\}, \tilde{m} \in M, u \in U : (u, \tilde{m}, r') \in Y\}$ where $R_h \subseteq R$ and $Y' \subseteq Y$. R_h defines a set of hashtags, M a set of messages, U a set of users and Y a ternary relation between U, M, and R. In words, a hashtag stream consists of all messages containing a specific hashtag and all resources and users related with these messages.

In the first of the following publications [Wagner and Strohmaier, 2010] we focus on creating semantic annotations of user streams. In this work we show that for predicting the topics of a future message in a user stream knowing who communicated with whom is useful, since messages authored by a user $u1$ are more likely to be about similar topics as messages authored by users with whom user $u1$ has communicated in the past. Therefore, we empirically show that structural metadata which emerge from users' communication activities in social streams may indeed be useful for creating semantic annotations of user streams.

Based on this observation, one may hypothesize that the background knowledge of the audience might be useful for creating semantic annotations of social stream messages. In the second publication [Wagner et al., 2011] we test this hypothesis and explore the value of the background knowledge of the audience for the task of semantically annotating social media messages. In this work we show that the audience of certain types of social stream possesses knowledge which is indeed useful for interpreting the meaning of social streams' messages.

Since users may conduct various types of user activities in online social environments such as Twitter, we present a comparative user study as well as a prediction experiment and analyze the potential of different types of data (generated through different user activities) for informing human's expertise judgements as well as computational expertise models in the third publication [Wagner et al., 2012a]. Our results show that different types of data generated through different types of user activities are useful for different computational and cognitive tasks, and the task of expertise modeling benefits most from information contained in user lists as opposed to tweet, retweet or bio information.

In the fourth publications of this section [Wagner et al., 2013a], we again compare a large variety of features for classifying Twitter users according to their professions and online personality. Our results show that linguistic-style and social-semantic features are very efficient for classifying users according to their personality related attributes and professions. On the other hand, we found that activity features were not useful across all classes and semantic features did not perform well on random users which suggest that they do not generalize well. This indicates that *how a user says something* (linguistic style) and *what others say about/to a user* (social-semantic) tend to be most useful for identifying users' professional areas and personality related attributes. *What a user says* (semantics) and *how a user behaves online* (activity patterns) tend to reveal less information about his professional areas and personality.

In addition to user streams, we also investigate the semantic annotation of hashtag streams [Posch et al., 2013]. In the fifth and final publication of this section we show that different semantic categories of hashtag streams reveal significantly different usage patterns and consequently reveal significantly different structural properties which can be used to predict the semantic categories of hashtag streams.

Pragmatic metadata matters: How data about the usage of data effects semantic user models

Claudia Wagner[1], Markus Strohmaier[2], and Yulan He[3]

[1] JOANNEUM RESEARCH, Institute for Information and Communication Technologies
Steyrergasse 17, 8010 Graz, Austria
claudia.wagner@joanneum.at
[2] Graz University of Technology and Know-Center
Inffeldgasse 21a, 8010 Graz, Austria
markus.strohmaier@tugraz.at
[3] The Open University, KMi
Walton Hall, Milton Keynes MK7 6AA, UK
yhe@open.ac.uk

Abstract. Online social media such as wikis, blogs or message boards enable large groups of users to generate and socialize around content. With increasing adoption of such media, the number of users interacting with user-generated content grows and as a result also the amount of *pragmatic metadata* - i.e. data about the usage of content - grows.

The aim of this work is to compare different methods for learning topical user profiles from Social Web data and to explore if and how pragmatic metadata has an effect on the quality of semantic user models. Since accurate topical user profiles are required by many applications such as recommender systems or expert search engines, learning such models by observing content and activities around content is an appealing idea.

To the best of our knowledge, this is the first work that demonstrates an effect between pragmatic metadata on one hand, and the quality of semantic user models based on user-generated content on the other. Our results suggest that *not all types of pragmatic metadata are equally useful* for acquiring *accurate* semantic user models, and some types of pragmatic metadata can even have detrimental effects.

Keywords: Semantic Analysis, Social Web, Topic Models, User Models

1 Introduction

Online social media such as Twitter, wikis, blogs or message boards enable large groups of users to create content and socialize around content. When a large group of users interact and socialize around content, *pragmatic metadata* is produced as a side product. While *semantic metadata* is often characterized as *data about the meaning of data*, we define *pragmatic metadata* as *data about the usage of data.* Thereby, pragmatic metadata captures how data/content is used by individuals or groups of users - such as who authored a given message, who replied to messages, who “liked” a message, etc. Although the amount of pragmatic metadata is growing, we still know little about how these metadata can be exploited for understanding the topics users engage with.

Many applications, such as recommender systems or intelligent tutoring systems, require good user models, where ”‘good”’ means that the model accurately reflects user‘s interest and behavior and is able to predict future content and activities of users. In this work we explore to what extent and how pragmatic metadata may contribute to semantic models of users and their content and compare different methods for learning topical user profiles from Social Web data.

To this end, we use data from an online message board. We incorporate different types of pragmatic metadata into different topic modeling algorithms and use them to learn topics and to annotate users with topics. We evaluate the quality of different semantic user models by comparing their predictive performance on future posts of user. Our evaluation is based on the assumption that “better” user models will be able to predict future content of users more accurately and will need less time and training data.

Generative probabilistic models are a state of the art technique for unsupervised learning. In such models, observed and latent variables are represented as random variables and probability calculus is used to describe the connections that are assumed to exist between these variables. Only if the assumptions made by the model are correct, Bayesian inference can be used to answer questions about the data. Generative probabilistic models have been successfully applied to large document collections (see e.g. [1]). Since for many documents one can also observe metadata, several generative probabilistic models have been developed which allow exploiting special types of metadata (see e.g., the Author Topic model [10], the Author-Recipient Topic model [8], the Group Topic model [14] or the Citation Influence Topic model [2]). However, previous research [10] has also shown that incorporating metadata into the topic modeling process may lead to model assumptions which are too strict and might overfit the data. This means that incorporating metadata does not necessarily lead to “better” topic models, where “better” means, for example, that the model is able to predict future user-generated content more accurately and needs less trainings data to fit the model.

Our work aims to advance our understanding about the effects of pragmatics on semantics emerging from user-generated content and specifically aims to answer the following questions:

1. Does incorporating pragmatic metadata into topic modeling algorithms lead to more accurate models of users and their content and if yes, what types of pragmatic metadata are more useful?
2. Does incorporating behavioral user similarities help acquiring more accurate models of users and their content and if yes, which types of behavioral user similarity are more useful?

The remainder of the paper is organized as follows: Section 2 gives an overview of the related work, while Section 3 describes our experimental setup. In Section 4 we report our results, followed by a discussion of our findings in Section 5.

2 Related Work

From a machine learning perspective, social web applications such as Boards.ie provide a huge amount of unlabeled training data for which usually many types of metadata can be observed. Several generative probabilistic models have been developed which allow exploiting special types of metadata (such as the Author Topic model [10], the Author-Recipient Topic model [8], the Group Topic model [14] or the Citation Influence Topic model [2]). In contrast to previous work where researchers focused on creating new topic models for each type of metadata, [9] presents a new family of topic models, Dirichlet-Multinomial Regression (DMR) topic models, which allow incorporating arbitrary types of observed features . Our work builds on the DMR topic model and aims to explore the extent to which different types of pragmatic metadata contribute to learning topic models from user generated content.

In addition to research on advancing topic modeling algorithms, the usefulness of topic models has been studied in different contexts, including social media. For example, [5] explored different schemes for fitting topic models to Twitter data and compared these schemes by using the fitted topic model for two classification tasks. As we do in our work, they also point out that models trained with a "'User"' scheme (i.e., using post aggregations of users as documents) perform better than models trained with a "'Post"' scheme. However, in contrast to our work they only explore relatively simple topic models and do not take any pragmatic metadata (except authorship information) into account when learning their models.

In our own previous work, we have studied the relationship between pragmatics and semantics in the context of social tagging systems. We have found that, for example, the pragmatics of tagging (users' behavior and motivation in social tagging systems [11, 6, 4]) exert an influence on the usefulness of emergent semantic structures [7]. In social awareness streams, we have shown that different types of Twitter stream aggregations can significantly influence the result of semantic analysis of tweets [12]. In this paper, we extend this line of research by (i) applying general topic models and (ii) using a dataset that offers rich pragmatic metadata.

3 Experimental Setup

The aim of our experiments is to explore to what extent and how pragmatic metadata can be exploited when semantically analyzing user generated content.

3.1 Dataset

The dataset used for our experiments and analysis was provided by Boards.ie,[4] an Irish community message board that has been in existence since 1998. We used all messages published during the first week of February 2006 (02/01/2006 - 02/07/2006) and the last week of February 2006 (02/21/2006 - 02/28/2006). We only used messages authored by users who published more than 5 messages and replied to more than 5 messages during this week. While we performed our experiments on both datasets, the results are similar. Consequently, we focus on reporting results obtained on the first dataset which consists of 1401 users and 27525 posts which were authored by these users and got replies.

To assess the predictive performance of different topic models we estimate how well they are able to predict the content (i.e. the actual words) of future posts. We generated a test corpus of 4007 held out posts in the following way: for each of the 1401 user in our training corpus we crawled 3 future posts which were authored by them and to which at least one user of our training corpus has replied. From here on, we refer to this data has *hold-out* data.

3.2 Methodology

In this section we first introduce the topic modeling algorithms (LDA, AT-model and DMR topic model) on which our work is based and then proceed to describe the topic models which we fitted to our training data, their model assumptions and how we compared and evaluated them.

Latent Dirichlet Allocation (LDA) The idea behind LDA is to model documents as mixtures of topics and force documents to favor few topics. Therefore, each document exhibits different topic proportions and each topic is defined as a distribution over a fixed vocabulary of terms. That means the generation of a collection of documents is modeled as a three step process: First, for each document d a distribution over topics θ_d is sampled from a Dirichlet distribution α. Second, for each word w_d in the document d, a single topic z is chosen according to this distribution θ_d. Finally, each word w_d is sampled from a multinomial distribution over words ϕ_z which is specific for the sampled topic z.

The Author Topic (AT) model The Author Topic model [10] is an extension of LDA, which learns topics conditioned on the mixture of authors that composed the documents. The assumption of the AT model is that each document is generated from a topic distribution which is specific to the set of authors of the document. The observed set of variables are the words per document (similar as in LDA) and the authors per document. The latent variables which are learned by fitting the model, are the topic distribution per author (rather than the topic distribution per document as in LDA) and the word distribution per topic.

[4] `http://www.boards.ie/`

We implemented the AT-model based on Dirichlet-multinomial Regression (DMR) Models (explained in the next section). While the original AT-model uses multinomial distribution (which are all drawn from the same Dirichlet) to represent an author-specific topic distributions, the DMR-model based implementation uses a "fresh" Dirichlet prior for each author from which then the topic distribution is drawn.

Dirichlet-multinomial Regression (DMR) Models Dirichlet-multinomial regression (DMR) topic models [9] assume not only that documents are generated by a latent mixture of topics but also that mixtures of topics are influenced by an additional factor which is specific to each document. This factor is materialized via observed features (in our case pragmatic metadata such as authorship or reply user information) and induce some correlation across individual documents in the same group. This means that e.g. documents which have been authored by the same user (i.e., they belong to one group) are more likely to chose the same topics. Formally, the prior distribution over topics α is a function of observed document features, and is therefore specific to each distinct combination of feature values. In addition to the observed features we add a default feature to each document, to account for the mean value of each topic.

Fitting Topic Models In this section we describe the different topic models which we fitted to our training datasets (see table 1 and 2). Each topic model makes different assumptions on what a document is (see column 3), takes different types of pragmatic metadata into account (see column 4) and makes different assumptions on the document-specific topic distributions θ which generates each documents (see column 5).

For all models, we chose the standard hyperparameters which are optimized during the fitting process: $\alpha = 50/T$ (prior of the topic distributions), $\beta = 0.01$ (prior of the word distributions) and $\sigma^2 = 0.5$ (variance of the prior on the parameter values of the Dirichlet distribution α). For the default features $\sigma^2 = 10$. Based on the empirical findings of [13], we decided to place an asymmetric Dirichlet prior over the topic distributions and a symmetric prior over the distribution of words. All models share the assumption that the total number of topics used to describe all documents of our collection is limited and fixed (via hyperparameter T) and that each topic must favor few words (as denoted by hyperparameter β which defines the Dirichlet distribution from which the word distributions are drawn - the higher β the less distinct the drawn word distributions).

Following the model selection approach described in [3], we selected the optimal number of topics for our training corpus by evaluating the probability of held out data for various values of T (keeping $\beta = 0.01$ fixed). For both datasets (each represents one week boards.ie data), a model trained on the "'Post"' scheme (i.e., using each post as a document) gives on average (over 10 runs) the highest probability to held out documents if $T = 240$ and model trained on the "'User"' scheme (i.e., using all posts authored by one user as a document) gives on average (over 10 runs) the highest probability to held out documents if $T = 120$. We kept T fixed for all our experiments.

Evaluation of Topic Models To compare different topic models we use perplexity which is a standard measure for estimating the performance of a probabilistic model. Perplexity measures the ability of a model to predict words on held out documents. In our case a low perplexity score may indicate that a model is able to accurately predict the content of future posts authored by a user. The perplexity measure is defined as followed:

$$perplexity(d) = exp[-\frac{\sum_{i=0}^{N_d} lnP(w_i|\hat{\phi}, \alpha)}{N_d}] \tag{1}$$

In words, the perplexity of a held out post d is defined as the exponential of the negative normalized predictive likelihood of the words w_i of the held out post d (where N_d is the total number of words in d) conditioned on the fitted model.

ID	Alg	Doc	Metadata	Model Assumption
M1	LDA	Post	-	A post is generated by a mixture of topics and has to favor few topics.
M2	LDA	User	-	All posts of one user are generated by a mixture of topics and have to favor few topics.
M3	DMR	Post	author	A post is generated by a user‘s authoring-specific mixture of topics and a user has to favor few topics he usually writes about.
M4	DMR	User	author	All posts of one user are generated by a user‘s authoring-specific mixture of topics and a user has to favor few topics he usually writes about.
M5	DMR	Post	user who replied	A post is generated by a user‘s replying-specific mixture of topics and a user has to favor few topics he usually replies to.
M6	DMR	User	user who replied	All posts of one user are generated by a user‘s replying-specific mixture of topics and a user has to favor few topics he usually replies to.
M7	DMR	Post	related user	A post is generated by a user‘s authoring- or replying-specific mixture of topics and a user has to favor few topics he usually replies to and he usually writes about.

M8	DMR	User	related user	All posts of one user are generated by a user‘s authoring- or replying-specific mixture of topics and a user has to favor few topics he usually replies to and he usually writes about.

Table 1: Overview about different topic models which incorporate different types of pragmatic metadata.

ID	Alg	Doc	Metadata	Model Assumption
M9	DMR	Post	top 10 forums of author	A post is generated by a mixture of topics which is specific to users who show a similar forum usage behavior as the author of the post.
M10	DMR	User	top 10 forums of author	All posts are generated by a mixture of topics which is specific to users who show a similar forum usage behavior as the author of the post-aggregation.
M11	DMR	Post	top 10 communication partner of author	A post is generated by a mixture of topics which is specific to users who show a similar communication behavior as the author of the post.
M12	DMR	User	top 10 communication partner of author	All posts are generated by a mixture of topics which is specific to users who show a similar communication behavior as the author of the post-aggregation.

Table 2: Overview about different topic models which incorporate different types of smooth pragmatic metadata based on behavioral user similarities.

4 Experimental Results

Our experiments were set up to answer the following questions:

1. Does incorporating pragmatic metadata into topic modeling algorithms lead to more accurate models of users and their content and if yes, what types of pragmatic metadata are more useful?

To answer this question, we fit different models to our training corpus and tested their predictive performance on future posts authored by our trainings users.

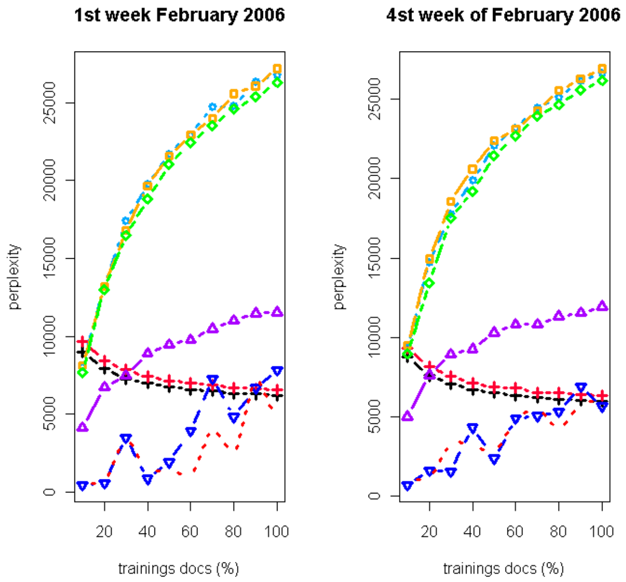


Fig. 1. Comparison of the predictive performance of different topic models on held out posts. The y-axis shows the average perplexity (over 10 runs) and the x-axis indicates the percentage of whole dataset used as training data. As baseline we use 2 versions of LDA (M1and M2).

Figure 1 shows that the predictive performance of semantic models of users which are either solely based on the users (i.e., aggregations of users‘ posts) to whom these users replied (M6) or which take in addition also the content authored by these users (M8) into account, is best. Therefore, our results suggest that it is beneficial to take user‘s reply behavior into account when learning topical user profiles from user generated content.

We also noted that all models which use the “User” training scheme (M4, M6 and M8) perform better than the models which use the “Post” training scheme (M3, M5 and M7). One possible explanation for this is the sparsity of posts which consist of only 66 tokens on average.

Since we were interested in how the predictive performance of different models change depending on the amount of data and time used for training, we split our training dataset randomly into smaller buckets and fitted the model on different proportions of the whole training corpus. One would expect that as the percentage of training data increases the predictive power of each model would improve as it adapts to the dataset. Figure 1 however shows that this is only true for our baseline models M1 and M2 which ignore all metadata of posts. The model M3 which corresponds to the Author Topic model exhibits a behavior that is similar to the behavior reported in [10]: When observing only few training data, M3 makes more accurate predictions on held-out posts than our baseline models. But the predictive performance of the model is limited by the strong assumptions that future posts of one author are about the same topics as past posts of the same author. Like M3, also M5 (and M7) seem to over-fit the data by making the assumptions that future posts of a user will be about the same topics as posts he replied to in the past (and posts he authored in the past).

To address these over-fitting problems we decided to incorporate smoother pragmatic metadata into the modeling process which we get by exploiting behavioral user similarities. The pragmatic metadata we used so far capture information about the usage behavior of individuals (e.g., who authored a document), while our smoother variants of pragmatic metadata capture information about the usage behavior of groups of users which share some common characteristics (e.g., what are the forums in which the author of this document is most active). Our intuition behind incorporating these smoother pragmatic metadata which are based on user similarities is that users which behave similar tend to talk about similar topics.

2. Does incorporating behavioral user similarities help acquiring more accurate models of users and their content and if yes, which types of behavioral user similarity are more useful?

From Figure 2 one can see that indeed all models which incorporate behavioral user similarity exhibit lower perplexity than our baseline models, especially if only few training samples are available. The model M12, which is based on the assumption that users who talk to the same users talk about the same topics, exhibits the lowest perplexity and outperforms our baseline models in terms of their predictive performance on held out posts.

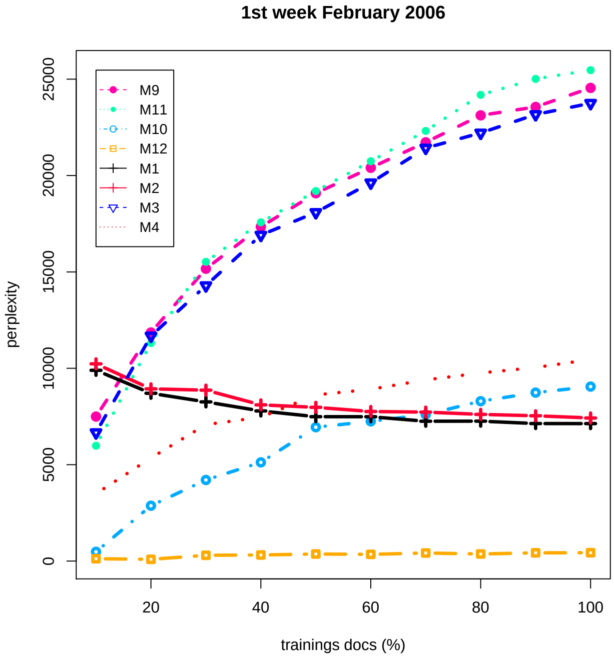


Fig. 2. Comparison of the predictive performance of topic models which take smooth pragmatic metadata into account by exploiting user similarities. The y-axis shows the average perplexity (over 10 runs) and the x-axis indicates the percentage of whole dataset used as training data. As baseline we use 2 versions of LDA (M1and M2).

For the model M10 which assumes that users who tend to post to the same forums talk about the same topics, we can only observe a lower perplexity than our baseline models when only few trainings data are available, but it still outperforms other state of the art topic models such as the Author topic model.

5 Discussion of Results and Conclusion

While it is intuitive to assume that incorporating metadata about the pragmatic nature of content leads to better learning algorithms, our results show that not all types of pragmatic metadata contribute in the same way. Our results confirm previous research which showed that topic models which incorporate pragmatic metadata such as the author topic model tend to over-fit data. That means incorporating metadata into a topic model can lead to model assumptions which are too strict and which yield the model to perform worse.

Summarizing, our results suggest that:

- **Pragmatics of content influence its semantics:** Integrating pragmatic metadata information into semantic user models influences the quality of resulting models.
- **Communication behavior matters:** Taking user‘s reply behavior into account when learning topical user profiles is beneficial. Content of users to which a user replied seems to be even more relevant for learning topical user profiles than content authored by a user.
- **Behavioral user similarities improve user models:** Smoother versions of metadata based topic models which take user similarity into account always seem to improve the models.
- **Communication behavior based similarities matter:** Different types of proxies for behavioral user similarity (e.g., number of forums they both posted to, number of shared communication partners) lead to different results. User who have a similar communication behavior seem to be more likely to talk about the same topics, than users who post to similar forums.

Acknowledgments. The authors want to thank Boards.ie for providing the dataset used in our experiments and Matthew Rowe for pre-processing the data. Furthermore we want to thank David Mimno for answering questions about the DMR topic model and Sofia Angelouta for fruitful discussions. Claudia Wagner is a recipient of a DOC-fForte fellowship of the Austrian Academy of Science.

References

1. Blei, D.M., Ng, A.Y., Jordan, M.I.: Latent dirichlet allocation. J. Mach. Learn. Res. 3, 993–1022 (2003)
2. Dietz, L., Bickel, S., Scheffer, T.: Unsupervised prediction of citation influences. In: International Conference on Machine Learning. pp. 233–240 (2007)
3. Griffiths, T.L., Steyvers, M.: Finding scientific topics. Proceedings of the National Academy of Sciences 101(Suppl. 1), 5228–5235 (April 2004)

4. Helic, D., Trattner, C., Strohmaier, M., Andrews, K.: On the navigability of social tagging systems. In: The 2nd IEEE International Conference on Social Computing (SocialCom 2010), Minneapolis, Minnesota, USA. pp. 161–168 (2010)
5. Hong, L., Davison, B.D.: Empirical study of topic modeling in twitter. In: Proceedings of the First Workshop on Social Media Analytics. pp. 80–88. SOMA '10, ACM, New York, NY, USA (2010), `http://doi.acm.org/10.1145/1964858.1964870`
6. Koerner, C., Kern, R., Grahsl, H.P., Strohmaier, M.: Of categorizers and describers: An evaluation of quantitative measures for tagging motivation. In: 21st ACM SIGWEB Conference on Hypertext and Hypermedia (HT 2010), Toronto, Canada, ACM. ACM, New York, NY, USA (June 2010)
7. Koerner, C., Benz, D., Strohmaier, M., Hotho, A., Stumme, G.: Stop thinking, start tagging - tag semantics emerge from collaborative verbosity. In: Proc. of the 19th International World Wide Web Conference (WWW 2010). ACM, Raleigh, NC, USA (Apr 2010), `http://www.kde.cs.uni-kassel.de/benz/papers/2010/koerner2010thinking.pdf`
8. Mccallum, A., Corrada-Emmanuel, A., Wang, X.: The author-recipient-topic model for topic and role discovery in social networks: Experiments with enron and academic email. Tech. rep., UMass CS (December 2004)
9. Mimno, D., McCallum, A.: Topic Models Conditioned on Arbitrary Features with Dirichlet-multinomial Regression. In: Proceedings of the 24th Conference on Uncertainty in Artificial Intelligence (UAI '08) (2008), `http://citeseerx.ist.psu.edu/viewdoc/summary?doi=10.1.1.140.6925`
10. Rosen-Zvi, M., Griffiths, T., Steyvers, M., Smyth, P.: The author-topic model for authors and documents. In: Proceedings of the 20th conference on Uncertainty in artificial intelligence. pp. 487–494. UAI '04, AUAI Press, Arlington, Virginia, United States (2004), `http://portal.acm.org/citation.cfm?id=1036843.1036902`
11. Strohmaier, M., Koerner, C., Kern, R.: Why do users tag? Detecting users' motivation for tagging in social tagging systems. In: International AAAI Conference on Weblogs and Social Media (ICWSM2010), Washington, DC, USA, May 23-26. AAAI, Menlo Park, CA, USA (2010)
12. Wagner, C., Strohmaier, M.: Exploring the wisdom of the tweets: Knowledge acquisition from social awareness streams. In: Proceedings of the Semantic Search 2010 Workshop (SemSearch2010), in conjunction with the 19th International World Wide Web Conference (WWW2010), Raleigh, NC, USA, April 26-30, ACM (2010)
13. Wallach, H.M., Mimno, D., McCallum, A.: Rethinking LDA: Why priors matter. In: Proceedings of NIPS (2009), `http://books.nips.cc/papers/files/nips22/NIPS2009\_0929.pdf`
14. Wang, X., Mohanty, N., Mccallum, A.: Group and topic discovery from relations and text. In: In Proc. 3rd international workshop on Link discovery. pp. 28–35. ACM (2005)

The Wisdom of the Audience: An Empirical Study of Social Semantics in Twitter Streams

Claudia Wagner[1], Philipp Singer[2], Lisa Posch[2], and Markus Strohmaier[2]

[1] JOANNEUM RESEARCH, Institute for Information and Communication Technologies
Steyrergasse 17, 8010 Graz, Austria
[2] Graz University of Technology, Knowledge Management Institute
Inffeldgasse 13, 8010 Graz, Austria

Abstract. Interpreting the meaning of a document represents a fundamental challenge for current semantic analysis methods. One interesting aspect mostly neglected by existing methods is that authors of a document usually assume certain background knowledge of their intended audience. Based on this knowledge, authors usually decide what to communicate and how to communicate it. Traditionally, this kind of knowledge has been elusive to semantic analysis methods. However, with the rise of social media such as Twitter, background knowledge of intended audiences (i.e., the community of potential readers) has become explicit to some extents, i.e., it can be modeled and estimated. In this paper, we (i) systematically compare different methods for estimating background knowledge of different audiences on Twitter and (ii) investigate to what extent the background knowledge of audiences is useful for interpreting the meaning of social media messages. We find that estimating the background knowledge of social media audiences may indeed be useful for interpreting the meaning of social media messages, but that its utility depends on manifested structural characteristics of message streams.

1 Introduction

In many social semantic web scenarios, understanding the meaning of social media documents is a crucial task. While existing semantic analysis methods can be used to understand and model the semantics of individual social media messages to some extent, the real time nature and the length of individual messages make it challenging to understand and model their semantics (Inches, Carman, & Crestani, 2010).

One drawback of existing methods is that they are limited to analyzing content, i.e. they do not have access to the background knowledge of potential readers. But as we know from communication theory, e.g., the Maxim of Quantity by Grice (Grice, 1975) or from Speech Act Theory (Searle, 1975), authors of messages usually make their messages as informative as required but do not provide more information than necessary. This suggests that the background knowledge of an intended audience for a given message can contribute to a semantic analysis task.

This paper sets out to study this hypothesis. We use three datasets obtained from Twitter, a popular microblogging service. Since information consumption on Twitter is mainly driven by explicitly defined social networks, we approximate the potential audience of a stream using the social network of a given author. In addition, we estimate the collective background knowledge of an audience by using the content published by the members of the audience. While the aim of this work is not to predict who will read a message, we want to approximate the collective background knowledge of a set of users who are likely to be exposed to a message and might have the background knowledge to interpret it. We do that to assess the value of background knowledge for interpreting the semantics of microblog messages. More specifically, this work addresses following research questions:

RQ1: To what extent is the background knowledge of the audience useful for guessing the meaning of social media messages? To investigate this question, we conduct a classification experiment in which we aim to classify messages into hashtag categories. As shown in (Laniado & Mika, 2010), hashtags can in part be considered as a manually constructed semantic grounding of individual microblog messages. In this work, we are going to assume that an audience which can guess the hashtag of a given message more accurately can also interprete the meaning of the message more accurately. We will use messages authored by the audience of a stream for training the classifier and we will test the performance on actual messages of a stream.

RQ2: What are the characteristics of an audience which possesses useful background knowledge for interpreting the meaning of a stream's messages and which types of streams tend to have useful audiences? To answer this question, we introduce several measures describing structural characteristics of an audience and its corresponding social stream. Then, we measure the correlation between these characteristics and the corresponding classification performance analyzed in RQ1. This shows the extent to which useful audiences can be identified based on structural characteristics.

The results of our experiments demonstrate that the background knowledge of a stream's audience is useful for the task of interpreting the meaning of microblog messages, but that the performance depends on structural characteristics of the audience and the underlying social stream. To our best knowledge, this is the first work which explores *to what extent* and *how* the background knowledge of an audience can be used to understand and model the semantics of individual microblog messages. Our work is relevant for researchers interested in learning semantic models from text and researchers interested in annotating social streams with semantics.

This paper is structured as follows: In Section 3 we give an overview about related research. Section 4 describes our experimental setup, including our methodology and a description of our datasets. Section 5 presents our experiments and empirical results. In Section 6 we discuss our results and conclude our work in Section 7.

2 Terminology

We define a *social stream* as a stream of data or content which is produced through users' activities conducted in an online social environment like Twitter where others see the manifestation of these activities. We assume that no explicitly defined rules for coordination in such environments exist. In this work we explore one special type of social streams, i.e., *hashtag streams*. A hashtag stream is a special type of a resource stream (Wagner & Strohmaier, 2010) and is defined as a tuple $S(R') = (U, M, R, Y', ft)$, where $Y' = \{(u, m, r) \mid r \in R' \vee \exists r' \in R', \tilde{m} \in M, u \in U : (u, \tilde{m}, r') \in Y\}$ and $R' \subseteq R$ and $Y' \subseteq Y$. In words, a hashtag stream consists of all messages containing one or several specific hashtags $r' \in R'$ and all resources (e.g., other hahstags, URLs or keywords) and users related to these messages.

In social online environments, information consumption is driven by explicitly defined social networks and therefore we can estimate the *audience* of a social stream by analyzing the incoming and outgoing links of the authors who created the stream. We call a user U_1 a *follower* of user U_2 if U_1 has established a unidirectional link with U_2 (in contrast user U_2 is a *followee* of user U_1), while we call a user U_3 a *friend* of user U_1 if U_1 has established a link with U_3 and vice versa. In this work, we assume that the union of the friends of all authors of a given hashtag constitute a hashtag stream's *audience*.

3 Related Work

Understanding and modeling the semantics of individual messages is important in order to support user in consuming social streams efficiently – e.g., via filtering social streams by users' interests or recommending tweets to users. Using topic relevance is an established approach to compute recommendations (Balabanović & Shoham, 1997) (Melville, Mooney, & Nagarajan, 2001) (Mooney & Roy, 2000).

However, the sparsity of microblog messages (i.e., the limited length of messages) makes it challenging to assess the topics of individual messages. Hence, researchers got interested in exploring the limitations of state-of-the-art text mining approaches in the context of microblogs and other short texts and develop methods for overcoming them. Two commonly used strategies for improving short text classification are: (a) improving the classifier or feature representation and (b) using background knowledge for enriching sparse textual data.

Improving the classifier or feature representation: Sriram et al. (Sriram, Fuhry, Demir, Ferhatosmanoglu, & Demirbas, 2010) present a comparison of different text mining methods applied on individual Twitter messages. Similar to our work, they use a message classification task to evaluate the quality of the outcome of each text mining approach. Limitations of their work are that they only use 5 broad categories (news, opinions, deals, events and private message) in which they classify tweets. Further, they perform their experiments on a very small set of tweets (only 5407 tweets) which were manually assigned to the aforementioned categories. Their results show that the authorship plays a crucial role since authors generally adhere to a specific tweeting pattern i.e., a majority of tweets from the same author tend to be within a limited set of categories. However, their authorship feature requires that tweets of the same authors occur in the trainings and test dataset.

Latent semantic models such as topic models provide a method to overcome data sparsity by introducing a latent semantic layer on top of individual documents. Hong et al. (Hong & Davison, 2010) compare the quality and effectiveness of different standard topic models in the context of social streams and examine different training strategies. To assess the quality and effectiveness of different topic models and training strategies the authors use them in two classification tasks: a user and message classification task. Their results show that the overall accuracy for classifying messages into 16 general Twitter suggest categories (e.g., Health, Food&Drinks, Books) when using topics as features is almost twice as accurate as raw TF-IDF features. Further their results suggest that the best performance can be achieved by training a topic model on aggregated messages per user. One drawback of their work is that they only use 274 users from 16 selected Twitter suggest directories[3]. These users are selected by a Twitter algorithm and it is therefore very likely that these users mainly post messages about the topic they are assigned to and that they are very popular.

In (Tang, Wang, Gao, Hu, & Liu, n.d.) the authors present an efficient approach that enriches data representation by employing machine translation to increase the number of features from different languages. Concretely the authors present a novel framework which performs multi-language knowledge integration and feature reduction simultaneously through matrix factorization techniques. The proposed approach is evaluated in terms of effectiveness on two social media datasets from Facebook and Twitter. For both Facebook and Twitter datasets, the authors construct a ground truth by selecting 30 topics from Google Trends, and retrieve the most relevant personal status or tweets via their APIs. Their results suggest that their proposed approach significantly improves the short text clustering performance.

Enriching sparse textual data with background knowledge: Based on the type of background knowledge being used, prior work can be categorized into one of the following three categories: thesaurus, web knowledge, and both of them.

Web Knowledge: Text categorization performance is improved by augmenting the bag of word representation with new features from ODP and Wikipedia as shown in (Gabrilovich & Markovitch, 2005) and (Gabrilovich & Markovitch, 2006). In (P. Wang & Domeniconi, 2008) the authors embed background knowledge derived from Wikipedia into a semantic kernel, which is then used to enrich the representation of documents. Their empirical evaluation with real data sets demonstrates that their approach successfully achieves improved classification accuracy with respect to the bag of words approach. Banerjee et al. (Banerjee, Ramanathan, & Gupta, 2007) show that clustering performance of Google news items at the feed reader end can be improved by incorporating titles of the top-

[3] http://twitter.com/invitations/suggestions

relevant Wikipedia articles as extra features. In (Phan, Nguyen, & Horiguchi, 2008) the authors present a general framework to build classifiers for short and sparse text data by using hidden topics discovered from huge text and Web collections. Their empirical results show that exploiting those hidden topics improves the accuracy significantly within two tasks: "Web search domain disambiguation" and "disease categorization for medical text".

Thesaurus or Dictionary: group words according to their similarity of meaning. Hotho et al. (Hotho, Staab, & Stumme, 2003) present an extensive study on the usage of background knowledge from WordNet for enriching documents and show that most enrichment strategies can indeed improve the document clustering accuracy. However, it is unclear if their results generalize to the social media domain since the vocabulary mismatch between WordNet and Twitter might be bigger than between WordNet and news articles.

Yoo et al. (Yoo, Hu, & Song, 2006) mapped terms in a document into MeSH concepts through the MeSH thesaurus and found that this strategy can improve the performance of text clustering. In (Shen et al., 2005) the authors use WordNet to reduce the vocabulary mismatch between the categories in the space of a search engine and the space of KDDCUP categories.

Thesaurus and Web Knowledge: For example, Hu et al. (Hu, Sun, Zhang, & Chua, 2009) cluster short texts (i.e., Google snippets) by first extracting the important phrases and expanding the feature space by adding semantically close terms or phrases from WordNet andWikipedia. Their proposed method employs a hierarchical three-level structure to tackle the data sparsity problem of original short texts and reconstruct the corresponding feature space with the integration of multiple semantic knowledge bases Wikipedia and WordNet. Empirical evaluation with Reuters and real web dataset demonstrates that their approach is able to achieve significant improvement as compared to the state-of-the-art methods.

Ontologies: include the Is-A hierarchy as well as non-taxonomic relations between entities (such as hasWonPrize).

In (Bloehdorn, Cimiano, Hotho, & Staab, 2005) the authors present an approach that uses text mining to learn the target ontology from text documents and uses then the same target ontology in order to improve the effectiveness of both supervised and unsupervised text categorization. Using Boosting as actual learning algorithm and both, term stems and concepts as features, the authors were able to achieve consistent improvements of the categorization results (1% 3% range for the Reuters-21578 corpus and in the 2.5% 7% range for the OHSUMED corpus).

In (B. B. Wang, Mckay, Abbass, & Barlow, 2002) the authors present a novel method to search for the optimal representation of a document in a domain ontology hierarchical structure to reflect concepts. Experiments have shown this is a feasible method to reduce the dimensionality of the document vector space effectively and reasonably and consequently improves the accuracy of the classifier while decreasing the computational costs. Further experiments with conceptual feature representations for supervised text categorization are presented in

(B. B. Wang, Mckay, Abbass, & Barlow, 2003) and suggest as well that concept-feature representations often outperform bag of word features.

Incorporating Background Knowledge: Hotho et al. (Hotho et al., 2003) compare several methods (add, replace, only) for incorporating background knowledge into the Bag of Words approach. The method *add* adds concepts to the word vector, while the method *replace* substitutes words with corresponding concepts. The method *only* uses only the concept vector. Hotho et al. also present different approaches for relating concepts with words. Those methods range from simple string matching to more complex word-context based disambiguation methods.

Latent semantic models such as topic models allow to incorporate background knowledge directly into the model learning step. For example, (?, ?) present approach that allows incorporating domain knowledge (in form of which words should have high or low probability in various topics) using a novel Dirichlet Forest prior in a Latent Dirichlet Allocation framework.

While (?, ?) suggest to represent background knowledge as prior probabilities of words for given topics, (?, ?) allow representing background knowledge as hierarchies of semantic concepts. In (?, ?) the authors present a probabilistic framework for combining human-defined background knowledge (represented via a hierarchy of semantic concepts) with a statistical topic model to seek the best of both worlds. Results indicate that this combination leads to systematic improvements in generalization performance.

Hashtags on Twitter: Since we use hashtags as semantic categories in which we aim to classify messages in our experiment, also research about users' hashtagging behavior is relevant for our work. In (Yang, Sun, Zhang, & Mei, 2012) the authors show that hashtags have a dual role – they are on the one hand used as topical or context marker of messages and on the other hand they are used as a symbol of community membership. The work by (Huang, Thornton, & Efthimiadis, 2010) suggests that hashtags are more commonly used to join public discussions than to organize content for future retrieval. The work of (Laniado & Mika, 2010) explores to what extent hashtags can be used as strong identifiers like URIs are used in the Semantic Web. Using manual annotations, they find that about half of the hashtags can be mapped to Freebase concepts with a high agreement between assessors. The authors make the assumption that hashtags are mainly used to ground tweets.

Summary: Recent research has shown promising steps towards improving short text classification by enhancing classifiers and feature representation or by using background knowledge from external sources such as Thesauri or the Web, to expand sparse textual data. However - to the best of our knowledge - using the background knowledge of intended audiences to interpret the meaning of social media messages represents a novel approach that has not been studied before. The general usefulness of such an approach is thus unknown.

4 Experimental Setup

The aim of our experiments is to explore different approaches for modeling and understanding the semantics or the main theme of microblog messages using different kinds of background knowledge. Since the audience of a microblog message are the users who are most likely to interpret (or to be able to interpret) the message, we hypothesize that the background knowledge of the audience of such messages might help to understand what a single message is about. In the following we describe our datasets and methodology.

4.1 Datasets

In this work we use three Twitter datasets each consisting of a temporal snapshot of the selected hashtag streams, the social network of stream's authors, their follower and followees and the tweets authored by the selected followers and followees (see Figure 1). We generate a diverse sample of hashtag streams as follows: In (Romero, Meeder, & Kleinberg, 2011) the authors created a classification of frequently used Twitter hashtags by category, identifying eight broad categories: celebrity, games, idioms, movies/TV, music, political, sports, and technology. We decided to reuse these categories and sample from each category 10 hashtags. We bias our random sample towards active hashtag streams by resampling hashtags for which we found less than 1,000 messages when crawling (4. March 2012). For those categories for which we could not find 10 hashtags which had more than 1,000 messages (games and celebrity) we select the most active hashtags per category (i.e., the hashtags for which we found the most messages). Since two hashtags (#bsb and #mj) appeared in the sample twice (i.e., in two different categories), we ended up having a sample of 78 different hashtags.

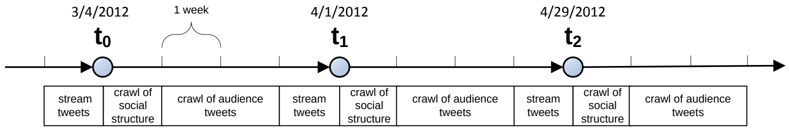


Fig. 1. Timeline of the crawling process.

Each dataset corresponds to one timeframe. The starting dates of the timeframes are March 4th (t_0), April 1st (t_1) and April 29th, 2012 (t_2). We crawled the most recent English tweets for each hashtag of our selection using Twitter's public search API on the first day of each timeframe and retrieved tweets that were authored within the last week. During the first week of each timeframe the user IDs of the followers and followees of streams's authors were crawled. Finally, we also crawled the most recent 3,200 tweets (or less if less were available) of

Table 1. Randomly selected hashtags per category (ordered alphabetically).

technology	idioms	sports	political	games	music	celebrity	movies
blackberry	factaboutme	f1	climate	e3	bsb	ashleytisdale	avatar
ebay	followfriday	football	gaza	games	eurovision	brazilmissesdemi	bbcqt
facebook	dontyouhate	golf	healthcare	gaming	lastfm	bsb	bones
flickr	iloveitwhen	nascar	iran	mafiawars	listeningto	michaeljackson	chuck
google	iwish	nba	mmot	mobsterworld	mj	mj	glee
iphone	nevertrust	nhl	noh8	mw2	music	niley	glennbeck
microsoft	omgfacts	redsox	obama	ps3	musicmonday	regis	movies
photoshop	oneofmyfollowers	soccer	politics	spymaster	nowplaying	teamtaylor	supernatural
socialmedia	rememberwhen	sports	teaparty	uncharted2	paramore	tilatequila	tv
twitter	wheniwaslittle	yankees	tehran	wow	snsd	weloveyoumiley	xfactor

all users who belong either to the top hundred authors or audience users of each hashtag stream. We ranked authors by the number of tweets they contributed to the stream and ranked audience users by the number of stream's authors with whom they have established a bidirectional follow relation. Figure 1 illustrates this process. Table 2 depicts the number of tweets and relations between users that we crawled during each timeframe.

Table 2. Description of the datasets.

	t_0	t_1	t_2
Stream Tweets	94,634	94,984	95,105
Audience Tweets	29,144,641	29,126,487	28,513,876
Stream Authors	53,593	54,099	53,750
Followers	56,685,755	58,822,119	66,450,378
Followees	34,025,961	34,263,129	37,674,363
Friends	21,696,134	21,914,947	24,449,705
Mean Followers per Author	1,057.71	1,087.31	1,236.29
Mean Followees per Author	634.90	633.34	700.92
Mean Friends per Author	404.83	405.09	454.88

4.2 Modeling Twitter Audiences and Background Knowledge

Audience Selection: Since the audience of a stream is potentially very large, we ranked the members of the audience according to the number of authors per stream an audience user is friend with. This allows us to determine key audience members per hashtag stream (see figure 2). We experimented with different thresholds (i.e., we used the top 10, 50 and top 100 friends) and got similar results. In the remainder of the paper, we only report the results for the best thresholds (c.f., table 3).

Background Knowledge Estimation: Beside selecting an audience of a stream, we also needed to estimate their knowledge. Hence, we compared four different methods for estimating the knowledge of a stream's audience:

- The first method (*recent*) assumes that the background knowledge of an audience can be estimated from the most recent messages authored by the audience users of a stream.

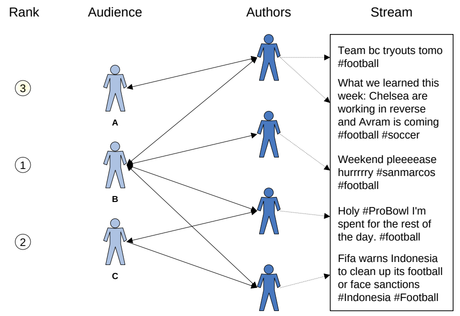

Rank
Audience
Authors
Stream
3
1
2
A
B
C
Team bc tryouts tomo #football
What we learned this week: Chelsea are working in reverse and Avram is coming #football #soccer
Weekend pleeeease hurrrrry #sanmarcos #football
Holy #ProBowl I'm spent for the rest of the day. #football
Fifa warns Indonesia to clean up its football or face sanctions #Indonesia #Football

Fig. 2. To estimate the audience of a hashtag stream, we ranked the friends of the stream's authors by the number of authors they are related with. In this example, the hashtag stream #football has four authors. User B is a friend of all four authors of the stream and is therefore most likely to be exposed to the messages of the stream and to be able to interpret them. Consequently, user B receives the highest rank. User C is a friend of two authors and receives the second highest rank. The user with the lowest rank (user A) is only the friend of one author of the stream.

- The second method (*top links*) assumes that the background knowledge of the audience can be estimated from the messages authored by the audience which contain one of the top links of that audience – i.e., the links which were recently published by most audience-users of that stream. Since messages including links tend to contain only few words due to the character limitations of Twitter messages (140 characters), we test two variants of this method. In the first variant we represented the knowledge of the audience via the plain messages which contain one of the top links (*top links plain*). In the second variant (*top links enriched*) we resolved the links and enriched the messages with keywords and title information which we got from the meta-tags of the html page the links are pointing to.
- Finally, the last method (*top tags*) assumes that the knowledge of the audience can be estimated via the messages authored by the audience which contain one of the top hashtags of that audience – i.e., the hashtags which were recently used by most audience users of that stream.

4.3 Methods

In this section we present the text mining methods we used to extract content features from raw text messages. In a preprocessing step we removed all English stopwords, URLs and Twitter usernames from the content of our microblog messages. We also removed Twitter syntax such as *RT* or *via*. For stemming we used Porter Stemming. In the following part of this section we describe the text mining methods we used for producing semantic annotations of microblog messages.

Bag-of-Words Model: Vector-based methods allow us to represent each microblog message as a vector of terms. Different methods exist to weight these terms – e.g., term frequency (*TF*), inverse document frequency (*IDF*) and term frequency-inverse document frequency (*TF-IDF*). We have used different weighting approaches and have achieved the best results by using TF-IDF. Therefore, we only report results obtained from the TF-IDF weighting schema in this paper.

Topic Models: Topic models are a powerful suite of algorithms which allow discovering the hidden semantic structure in large collection of documents. The idea behind topic models is to model documents as arising from multiple topics, where each document has to favor few topics. Therefore, each document exhibits different topic proportions and each topic is defined as a distribution over a fixed vocabulary of terms, where few words are favored.

The most basic topic modeling algorithm is Latent Dirichlet Allocation (LDA) (Blei, Ng, & Jordan, 2003). In our experiments we used MALLET's (McCallum, 2002) LDA implementation and fitted an LDA model to our tweet corpus using individual tweets as trainings document. We chose the default hyperparameters ($\alpha = 50/T$, $\beta = 0.01$) and optimized them during training by using Wallach's fixed point iteration method (Wallach, 2008). We chose the number of topics $T=$ 500 empirically by estimating the log likelihood of a model with $T=$ 300, 500 and 700 on held out data. Given enough iterations (we used 2000) the Markov chain (which consists of topic assignments z for each token in the training corpus) has potentially converged and we can get estimates of the word distribution of topics ($\hat{\phi}$) and the topic distribution of documents ($\hat{\theta}$) by drawing samples from the chain. The estimated distributions $\hat{\phi}$ and $\hat{\theta}$ are predictive distributions and are later used to infer the topics of social stream messages.

4.4 Message Classification Task

To evaluate the quality and utility of audience's background knowledge for interpreting the meaning of microblog message, we conducted a message classification task using hashtags as classes (i.e., we had a multi-class classification problem with 78 classes). We assume that an audience which is better in guessing the hashtag of a Twitter message is better in interpreting the meaning of the message. For each hashtag stream, we created a baseline by picking the audience of another stream at random and compared the performance of the random audience with the real stream's audience. Our baseline tests how well a randomly selected audience can interpret the meaning of stream's messages. One needs to note that a simple random guesser baseline would be a weaker baseline than the one described above and would lead to a performance of 1/78.

We extracted content features (via the aforementioned methods) from messages authored by the audience of a stream before $t1$ and used them to train a classifier. That means messages of the audience of a stream were used as training samples to learn a semantic representation of messages in each hashtag class. We tested the performance of the classifier on actual messages of a stream which were published after $t1$. In following such an approach, we ensured that our classifier does not benefit from any future information (e.g., messages published in the future or social relations which were created in the future). Out of several classification algorithms applicable for text classification such as Logistic Regression, Stochastic Gradient Descent, Multinomial Naive Bayes or Linear SVC, we could achieve the best results using a Linear SVC[4]. As evaluation metric we chose the weighted average *F1-score* which is the average of the harmonic means of precision and recall of each class weighted by the number of test samples from each class.

4.5 Structural Stream Measures

To assess the association between structural characteristics of a social stream and the usefulness of its audience (see RQ2), we introduce the following measures which describe structural aspects of those streams. We differ between static measures which only use information from one time point and dynamic measures which combine information from several time points.

Static Measures

- **Coverage Measures:** The coverage measures characterize a hashtag stream via the nature of its messages. For example the *informational coverage* measure indicates how many messages of a stream have an informational purpose - i.e., contain a link. The *conversational coverage* measures the mean number of messages of a stream that have a conversational purpose - i.e., those messages that are directed to one or several specific users. The *retweet coverage* measures the percentage of messages which are retweets. The *hashtag coverage* measures the mean number of hashtags per message in a stream.
- **Entropy Measures:** We use normalized entropy measures to capture the randomness of stream's authors and their followers, followees and friends. We rank for each hashtag stream the authors by the number of tweets they authored and the followers, followees and friends by the number of authors they are related with. A high *author entropy* indicates that the stream is created in a democratic way since all authors contribute equally much. A high *follower entropy* and *friend entropy* indicate that the followers and friends do not focus their attention towards few authors but distribute it equally across all authors. A high *followee entropy* and *friend entropy* indicate that the authors do not focus their attention on a selected part of their audience.
- **Overlap Measures:** The overlap measures describe the overlap between the authors and the followers (*Author-Follower Overlap*), followees (*Author-Followee Overlap*) or friends (*Author-Friend Overlap*) of a hashtag stream. If these overlaps are one, the stream is consumed and produced by the same users who are interconnected. A high overlap suggests that the community around the hashtag is rather closed, while a low overlap indicates that the

[4] `http://www.csie.ntu.edu.tw/~cjlin/liblinear/`

community is more open and that the active and passive part of the community do not extensively overlap.

Dynamic Measures To explore how the social structure of a hashtag stream changes over time we measure the distance between the tweet-frequency distributions of stream's authors at different time points and the author-frequency distributions of stream's followers, followees or friends at different time points. We use a symmetric version of the *Kullback-Leibler (KL) divergence* which represents a natural distance measure between two probability distributions and is defined as follows: $\frac{1}{2}D_{KL}(A||B)+\frac{1}{2}D_{KL}(B||A)$. The KL divergence is *zero* if the two distributions A and B are identical and approaches infinity as they differ more and more. We measure the KL divergence for the distributions of authors, followers, followees and friends.

5 Experiments

The aim of our experiments is to explore different methods for modeling and understanding the semantics of Twitter messages using background knowledge of different kinds of audiences. Due to space restrictions we only report results obtained when training our model on the dataset t_0 and testing it on the dataset t_1. We got comparable results when training on the dataset t_1 and testing on dataset t_2.

5.1 RQ1: To what extent is the background knowledge of the audience useful for guessing the meaning of social media messages?

To answer this question we compared the performance of a classification model using messages authored by the audience of a stream (i.e., the top friends of a hashtag stream's authors) as training samples with the performance of a classification model using messages of a randomly selected audience (a baseline, i.e. the top friends of the authors of a randomly selected hashtag stream) as training samples. If the audience of a stream does not possess more knowledge about the semantics of the stream's messages than a randomly selected baseline audience, the results from both classification models should not differ significantly.

Our results show that all classifiers trained on messages authored by the audience of a hashtag stream clearly outperform a classifier trained on messages authored by a randomly selected audience. This indicates that the messages authored by the audience of a hashtag stream indeed contain important information. Our results also show that a TF-IDF based feature representation slightly outperforms a topical feature representation.

The comparison of the four different background knowledge estimation methods (see section 4.2) shows that the best results can be achieved when using the most recent messages authored by the top 10 audience users and when using messages authored by the top 100 audience users containing one of the top hashtags of the audience (see table 3). Tweets containing one of the top links of the audience (no matter if enriched or not) are less useful than messages containing one of the top hashtags of the audience. Surprisingly, our message link enrichment strategies did not show a large boost in performance. A manual inspection of a small sample of links showed that the top links of an audience often point to multimedia sharing sites such as youtube[5], instagr.am[6] or twitpic[7]. Unfortunately, title and keywords which can be extracted from the meta information of those sites often contain information which is not descriptive.

Table 3. Average weighted F1-Scores of different classification models trained on data crawled at t_0 and tested on data crawled at t_1. We either used words weighted via TF-IDF or topics inferred via LDA as features for a message. The table shows that all audience-based classification models outperformed a random baseline. For the random baseline, we randomly swapped audiences and hashtag streams. A classifier trained on the most recent messages of the top 10 friends of a hashtag stream yields the best performance.

Classification Model	**F1 (TF-IDF)**	**F1 (LDA)**
Baseline (Random audience: top 10 friends, Messages: recent)	**0.01**	**0.01**
Audience: top 10 friends, Messages: recent	**0.25**	**0.23**
Audience: top 100 users, Messages: top links enriched	0.13	0.10
Audience: top 100 users, Messages: top links plain	0.12	0.10
Audience: top 100 users, Messages: top tags	**0.24**	**0.21**

To gain further insights into the usefulness of an audience's background knowledge, we compared the average weighted F1-Score of the eight hashtag categories from which our hashtags were initially drawn (see Table 4). Our results show that for certain categories such as sports and politics the knowledge of the audience clearly helps to learn the semantics of hashtag streams' messages, while for other streams – such as those belonging to the categories celebrities and idioms – background knowledge of the audience seems to be less useful. This suggests that only certain types of social streams are amenable to the idea of exploiting the background knowledge of stream audiences. Our intuition is that audiences of streams that are about fast-changing topics are *less useful.* We think that these audiences are only loosely associated to the topics of the stream, and therefore their background knowledge does not add much to a semantic analysis task. Analogously, we hypothesize audiences of streams that are narrow and stable are *more useful.* It seems that a community of tightly knit users is built around a topic and a common knowledge is developed over time. This seems to provide useful background knowledge to a semantic analysis task. Next, we want to understand the characteristics that distinguish audiences that are useful from audiences that are less useful.

[5] `http://www.youtube.com`
[6] `http://instagram.com/`
[7] `http://twitpic.com/`

Table 4. Average weighted F1-Score per category of the best audience-based classifier using recent messages (represented via TF-IDF weighted words or topic proportions) authored by the top ten audience users of a hashtag stream. The support represents the number of test messages for each class. We got the most accurate classification results for the category *sports* and the least accurate classification results for the category *idioms*.

		TFIDF		**LDA**	
category	**support**	**mean F1**	**variance F1**	**mean F1**	**variance F1**
celebrity	4384	0.17	0.08	0.15	0.16
games	6858	0.25	0.33	0.22	0.31
idioms	14562	**0.09**	0.14	**0.05**	0.05
movies	14482	0.22	0.19	0.18	0.18
music	13734	0.23	0.25	0.18	0.26
political	13200	0.36	0.22	0.33	0.21
sports	13960	**0.45**	0.19	**0.42**	0.21
technology	13878	0.22	0.20	0.22	0.2

5.2 RQ2: What are the characteristics of an audience which possesses useful knowledge for interpreting the meaning of stream's messages and which types of streams tend to have useful audiences?

To understand whether the structure of a stream has an effect on the usefulness of its audience for interpreting the meaning of its messages, we perform a correlation analysis and investigate to what extent the ability of an audience to interpret the meaning of messages correlates with structural stream properties. We use the F1-scores of the best audience based classifiers (using TFIDF and LDA) as a proxy measure for the audience's ability to interpret the meaning of stream's messages.

Figure 3a shows the strength of correlation between the F1-scores and the structural properties of streams across all categories. An inspection of the first two columns of the correlation matrix reveals interesting correlations between structural stream properties and the F1-scores of the audience-based classifiers. We further report all significant *Spearman rank correlation coefficients* ($p < 0.05$) across all categories in table 3b.

Figure 3a and table 3b show that across all categories, the measures which capture the overlap between the authors and the followers, friends and followees shows the highest positive correlation with the F1-scores. That means, the higher the overlap between authors of a stream and the followers, friends and followees of the stream, the better an audience-based classifier performs. This is not surprising since it indicates that the audience which is best in interpreting stream messages is an active audience, which also contributes to the creation of the stream itself (high author friend overlap). Further, our results suggest that the audience of a stream possesses useful knowledge for interpreting stream's messages if the authors of a stream follow each other (high author follower and author followee overlap). This means that the stream is produced and consumed by a community of users who are tightly interconnected. The only significant coverage measure is the conversational coverage measure. It indicates that the audiences of conversational streams are better in interpreting the meaning of stream's messages. This suggests that it is not only important that a community exists around a stream, but also that the community is communicative.

All entropy measures show significant negative correlations with the F1-Scores. This shows that the more focused the author-, follower-, followee- and/or friend-distribution of a stream is (i.e., lower entropy), the higher the F1-Scores of an audience-based classification model are. The entropy measures the randomness of a random variable. For example, the author-entropy describes how random the tweeting process in a hashtag stream is – i.e., how well one can predict who will author the next message. The friend-entropy describes how random the friends of hashtag stream's authors are – i.e., how well one can predict who will be a friend of most hashtag stream's authors. Our results suggest that streams tend to have a better audience if their authors and author's followers, followees and friends are less random.

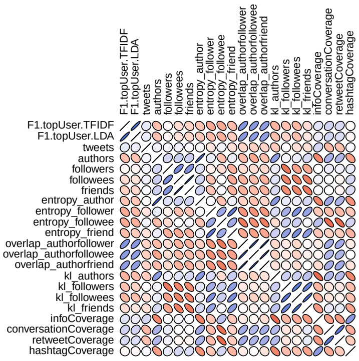
Finally, the KL divergences of the author-, follower-, and followee-distributions show a significant negative correlation with the F1-Scores. This indicates that the more stable the author, follower and followee distribution is over time, the better the audience of a stream is. If for example the followee distribution of a stream changes heavily over time, authors are shifting their social focus. If the author distribution of a stream has a high KL divergence, this indicates that the set of authors of stream are changing over time.

In summary, our results suggest that *streams which have a useful audience tend to be created and consumed by a stable and communicative community* – i.e., a group of users who are interconnected and have few core users to whom almost everyone is connected.

6 Discussion of Results

The results of this work show that messages authored by the audience of a hashtag stream indeed represent background knowledge that can help interpreting the meaning of streams' messages. We showed that the usefulness of an audience's background knowledge depends on the applied content selection strategies (i.e., how the potential background knowledge of an audience is estimated). However, since the audience of a hashtag stream is potentially very large, picking the right threshold for selecting the best subset of the audience is an issue. In our experiments we empirically picked the best threshold but did not conduct extensive experiments on this issue. Surprisingly, more sophisticated content selection strategies such as top links or top hashtags were only as good or even worse than the simplest strategy which used the most recent messages (up to 3,200) of each top audience user.

Our work shows that not all streams exhibit audiences which possess knowledge useful for interpreting the meaning of stream's messages (e.g., streams in certain categories like celebrities or especially idioms). Our results suggest that the utility of a stream's audience is significantly associated with structural characteristics of the stream.



(a)

feature	cor with F1 (TF-IDF)	cor with F1 (LDA)
overlap_authorfollower	0.675	0.655
overlap_authorfollowee	0.642	0.628
overlap_authorfriend	0.612	0.602
conversation_coverage	0.256	0.256
kl_followers	-0.281	–
kl_followees	-0.343	-0.302
kl_authors	-0.359	-0.307
entropy_author	-0.270	-0.400
entropy_friend	-0.307	–
entropy_follower	-0.400	-0.319
entropy_followee	-0.401	-0.368

(b)

Fig. 3. Figure 3a shows the Spearman rank correlation strength between structural stream properties and F1-Scores of two audience-based classification models averaged across all categories. The color and form of the ellipse indicate the correlation strength. Red means negative and blue means positive correlation. The rounder the ellipse the lower the correlation. The inspection of the first two columns of the correlation matrix reveals that several structural measures are correlated with the F1-Scores and table 3b shows which of those are indeed statistical significant.

Finally, our work has certain limitations. Recent research on users' hashtagging behavior (Yang et al., 2012) suggests that hashtags are not only used as topical or context marker of messages but can also be used as a symbol of community membership. In this work, we have mostly neglected the social function of hashtags. Although the content of a message may not be the only factor which influences which hashtag a user choses, we assume a “better” semantic model might be able to predict hashtags more accurately.

7 Conclusions and Future Work

This work explored whether the background knowledge of intended Twitter audiences can help in identifying the meaning of social media messages. We introduced different approaches for estimating the background knowledge of a stream's audience and presented empirical results on the usefulness of this background knowledge for interpreting the meaning of social media documents.

The main findings of our work are:

- The audience of a social stream possesses knowledge which may indeed help to interpret the meaning of stream's messages.
- The audience of a social stream is most useful for interpreting the meaning of stream's messages if the stream is created and consumed by a stable and communicative community – i.e., a group of users who are interconnected and have few core users to whom almost everyone is connected.

In our future work we want to explore further methods for estimating the potential background knowledge of an audience (e.g., using user lists or bio information rather than tweets). Combining latent and explicit semantic methods for estimating audience's background knowledge and exploiting it for interpreting the main theme of social media messages are promising avenues for future research.

Acknowledgments

This work was supported in part by a DOC-fForte fellowship of the Austrian Academy of Science to Claudia Wagner and by the FWF Austrian Science Fund Grant I677 and the Know-Center Graz.

References

Andrzejewski, D., Zhu, X., & Craven, M. (2009). Incorporating domain knowledge into topic modeling via dirichlet forest priors. In *Proceedings of the 26th annual international conference on machine learning* (pp. 25–32). New York, NY, USA: ACM. Available from `http://doi.acm.org/10.1145/1553374.1553378`

Balabanović, M., & Shoham, Y. (1997, March). Fab: content-based, collaborative recommendation. *Commun. ACM*, *40*(3), 66–72. Available from `http://doi.acm.org/10.1145/245108.245124`

Banerjee, S., Ramanathan, K., & Gupta, A. (2007). Clustering short texts using wikipedia. In *Proceedings of the 30th annual international acm sigir conference on research and development in information retrieval* (pp. 787–788). New York, NY, USA: ACM. Available from `http://doi.acm.org/10.1145/1277741.1277909`

Blei, D. M., Ng, A. Y., & Jordan, M. I. (2003). Latent dirichlet allocation. *J. Mach. Learn. Res.*, *3*, 993–1022.

Bloehdorn, S., Cimiano, P., Hotho, A., & Staab, S. (2005, May). An ontology-based framework for text mining. *LDV Forum - GLDV Journal for Computational Linguistics and Language Technology*, *20*(1), 87-112.

Gabrilovich, E., & Markovitch, S. (2005). Feature generation for text categorization using world knowledge. In *Proceedings of the 19th international joint conference on artificial intelligence* (pp. 1048–1053). San Francisco, CA, USA: Morgan Kaufmann Publishers Inc. Available from `http://dl.acm.org/citation.cfm?id=1642293.1642461`

Gabrilovich, E., & Markovitch, S. (2006). Overcoming the brittleness bottleneck using wikipedia: enhancing text categorization with encyclopedic knowledge. In *proceedings of the 21st national conference on artificial intelligence - volume 2* (pp. 1301–1306). AAAI Press. Available from `http://dl.acm.org/citation.cfm?id=1597348.1597395`

Grice, H. P. (1975). Logic and conversation. In P. Cole (Ed.), *Speech acts* (Vol. 3, pp. 41–58). New York: Academic Press.

Hong, L., & Davison, B. D. (2010). Empirical study of topic modeling in twitter. In *Proceedings of the igkdd workshop on social media analytics (soma),.*

Hotho, A., Staab, S., & Stumme, G. (2003). Wordnet improves text document clustering. In *In proc. of the sigir 2003 semantic web workshop* (pp. 541–544).

Hu, X., Sun, N., Zhang, C., & Chua, T.-S. (2009). Exploiting internal and external semantics for the clustering of short texts using world knowledge. In *Proceedings of the 18th acm conference on information and knowledge management* (pp. 919–928). New York, NY, USA: ACM. Available from `http://doi.acm.org/10.1145/1645953.1646071`

Huang, J., Thornton, K. M., & Efthimiadis, E. N. (2010). Conversational tagging in twitter. In *Proceedings of the 21st acm conference on hypertext and hypermedia* (pp. 173–178). New York, NY, USA: ACM. Available from `http://doi.acm.org/10.1145/1810617.1810647`

Inches, G., Carman, M., & Crestani, F. (2010). Statistics of online user-generated short documents. *Advances in Information Retrieval*, 649–652. Available from `http://dx.doi.org/10.1007/978-3-642-12275-0_68`

Laniado, D., & Mika, P. (2010). Making sense of twitter. In P. F. Patel-Schneider et al. (Eds.), *International semantic web conference (1)* (Vol. 6496, p. 470-485). Springer. Available from `http://dblp.uni-trier.de/db/conf/semweb/iswc2010-1.html#LaniadoM10`

McCallum, A. K. (2002). *Mallet: A machine learning for language toolkit.* (http://mallet.cs.umass.edu)

Melville, P., Mooney, R. J., & Nagarajan, R. (2001). Content-boosted collaborative filtering. In *In proceedings of the 2001 sigir workshop on recommender systems.*

Mooney, R. J., & Roy, L. (2000). Content-based book recommending using learning for text categorization. In *Proceedings of the fifth acm conference on digital libraries* (pp. 195–204). New York, NY, USA: ACM. Available from `http://doi.acm.org/10.1145/336597.336662`

Phan, X.-H., Nguyen, L.-M., & Horiguchi, S. (2008). Learning to classify short and sparse text & web with hidden topics from large-scale data collections. In *Proceedings of the 17th international conference on world wide web* (pp. 91–100). New York, NY, USA: ACM. Available from `http://doi.acm.org/10.1145/1367497.1367510`

Romero, D. M., Meeder, B., & Kleinberg, J. (2011). Differences in the mechanics of information diffusion across topics: idioms, political hashtags, and complex contagion on twitter. In *Proceedings of the 20th international conference on world wide web* (pp. 695–704). New York, NY, USA: ACM. Available from `http://doi.acm.org/10.1145/1963405.1963503`

Searle, J. (1975). A taxonomy of illocutionary acts. In K. Gunderson (Ed.), *Minnesota studies in the philosophy of language* (pp. 334–369). Minneapolis: University of Minnesota Press.

Shen, D., Pan, R., Sun, J.-T., Pan, J. J., Wu, K., Yin, J., et al. (2005, December). Q2c@ust: our winning solution to query classification in kddcup 2005. *SIGKDD Explor. Newsl.*, *7*(2), 100–110. Available from `http://doi.acm.org/10.1145/1117454.1117467`

Sriram, B., Fuhry, D., Demir, E., Ferhatosmanoglu, H., & Demirbas, M. (2010). Short text classification in twitter to improve information filtering. In *Proceedings of the 33rd international acm sigir conference on research and development in information retrieval* (pp. 841–842). New York, NY, USA: ACM. Available from `http://doi.acm.org/10.1145/1835449.1835643`

Steyvers, M., Smyth, A. P., & Chemuduganta, B. C. (2011). Combining Background Knowledge and Learned Topics. *Topics in Cognitive Science*, *3*(18–47). Available from `http://citeseerx.ist.psu.edu/viewdoc/summary?doi=10.1.1.165.7316`

Tang, J., Wang, X., Gao, H., Hu, X., & Liu, H. (n.d.). Enriching short texts representation in microblog for clustering. *Frontiers of Computer Science*.

Wagner, C., & Strohmaier, M. (2010). The wisdom in tweetonomies: Acquiring latent conceptual structures from social awareness streams. In *Semantic search workshop at www2010.* Available from `http://www.student.tugraz.at/claudia.wagner/publications/wagner_semsearch2010.pdf`

Wallach, H. M. (2008). *Structured topic models for language.* Unpublished doctoral dissertation, University of Cambridge.

Wang, B. B., Mckay, R. I. (bob, Abbass, H. A., & Barlow, M. (2002). Learning text classifier using the domain concept hierarchy. In *In proceedings of international conference on communications, circuits and systems 2002* (pp. 1230–1234). Press.

Wang, B. B., Mckay, R. I. B., Abbass, H. A., & Barlow, M. (2003). A comparative study for domain ontology guided feature extraction. In *Proceedings of the 26th australasian computer science conference - volume 16* (pp. 69–78). Darlinghurst, Australia, Australia: Australian Computer Society, Inc. Available from `http://dl.acm.org/citation.cfm?id=783106.783115`

Wang, P., & Domeniconi, C. (2008). Building semantic kernels for text classification using wikipedia. In *Proceedings of the 14th acm sigkdd international conference on knowledge discovery and data mining* (pp. 713–721). New York, NY, USA: ACM. Available from `http://doi.acm.org/10.1145/1401890.1401976`

Yang, L., Sun, T., Zhang, M., & Mei, Q. (2012). We know what @you #tag: does the dual role affect hashtag adoption? In *Proceedings of the 21st international conference on world wide web* (pp. 261–270). New York, NY, USA: ACM. Available from `http://doi.acm.org/10.1145/2187836.2187872`

Yoo, I., Hu, X., & Song, I.-Y. (2006). Integration of semantic-based bipartite graph representation and mutual refinement strategy for biomedical literature clustering. In *Proceedings of the 12th acm sigkdd international conference on knowledge discovery and data mining* (pp. 791–796). New York, NY, USA: ACM. Available from `http://doi.acm.org/10.1145/1150402.1150505`

It's not in their tweets: Modeling topical expertise of Twitter users

Claudia Wagner*, Vera Liao†, Peter Pirolli‡, Les Nelson‡ and Markus Strohmaier§
*Institute of Information and Communication Technologies, JOANNEUM RESEARCH, Graz, Austria
Email: claudia.wagner@joanneum.at
†Department of Computer Science, University of Illinois at Urbana-Champaign, Urbana, Illinois
Email: liao28@illinois.edu
‡Palo Alto Research Center, Palo Alto, California
Email: pirolli@parc.com, lnelson@parc.com
§ Knowledge Management Institute and Know-Center, Graz University of Technology,Graz, Austria
Email: markus.strohmaier@tugraz.at

***Abstract*—One of the key challenges for users of social media is judging the topical expertise of other users in order to select trustful information sources about specific topics and to judge credibility of content produced by others. In this paper, we explore the usefulness of different types of user-related data for making sense about the topical expertise of Twitter users. Types of user-related data include messages a user authored or re-published, biographical information a user published on his/her profile page and information about user lists to which a user belongs. We conducted a user study that explores how useful different types of data are for informing human's expertise judgements. We then used topic modeling based on different types of data to build and assess computational expertise models of Twitter users. We use Wefollow directories as a proxy measurement for perceived expertise in this assessment.**

Our findings show that different types of user-related data indeed differ substantially in their ability to inform computational expertise models and humans's expertise judgements. Tweets and retweets — which are often used in literature for gauging the expertise area of users — are surprisingly useless for inferring the expertise topics of their authors and are outperformed by other types of user-related data such as information about users' list memberships. Our results have implications for algorithms, user interfaces and methods that focus on capturing expertise of social media users.

***Index Terms*—expertise, user profiling, expert search, Twitter, microblogs**

I. Introduction

On social media applications such as Twitter, information consumption is mainly driven by social networks. Therefore, judging topical expertise of other users is a key challenge in maximizing the credibility and quality of information received. Recent research on users' perception of tweet credibility indicates that information about the authors is most important for informing credibility judgments of tweets [1]. This highlights that judging the credibility and expertise of Twitter users is crucial for maximizing the credibility and quality of information received. However, the plethora of information on a Twitter page makes it challenging to assess users' expertise accurately. In addition to the messages a user authored (short *tweets*) and re-published (short *retweets*), there is additional information on the Twitter interface that could potentially inform expertise judgements. For example, with fewer than 160 characters, the biographical section (short *bio*) may contain important information that indicates users' expertise level, such as his/her self summarized interests, career information, and links to his/her personal web page. Another feature of Twitter that could potentially be useful for assessing users' level of expertise is the support of user lists (short *lists*). User lists allow users to organize people they are following into labeled groups and aggregate their tweets by groups. If a user is added to a list, the list label and short description of the list will appear on his/her Twitter page. Unlike bio information, which may contain self-reported expertise indication, users' list memberships can reflect external expertise indications, i.e., followers' judgements about one's expertise. However, little is known about the motivations of users for adding other users to lists and the type of information which is revealed by users' list memberships, their bio section and tweet and retweets published by them.

This paper aims to shed some light on the usefulness of different types of user-related data (concretely we use *tweets*, *retweets*, *bio* and *list* data) for making sense of the domain expertise of Twitter users. We use Wefollow[1] directories as a proxy measurement for perceived expertise in this assessment. Wefollow is an application that allows Twitter users to register themselves in a maximum of 3 topical directories. Although Wefollow directories may not provide perfect ground truth for perceived expertise, canonical ranking and social judgments by peers are commonplace for identifying expertise [2]. We assume the way that Wefollow functions, by ranking users according to the number of followers in the same field, is a reflection of such social judgment. Our assumption is supported by previous research which has shown that the majority of the top 20 Wefollow users for selected directories were perceived as experts for the corresponding topic [3] and that experts tend to agree that users with high Wefollow rank are more knowledgeable than users with low or no Wefollow rank [4]. We leverage these findings for our study which aims

[1] http://wefollow.com

to address the following research questions:

1) *What type of user-related social media data is most useful for informing human's expertise judgements about Twitter users?*
2) *Do different types of user-related social media data lead to similar topical expertise profiles of Twitter users or are these profiles substantially different?*
3) *What type of user-related social media data is most useful for creating topical expertise profiles of Twitter users?*

We approached this question from two complementary perspectives. First, we conducted a user study to explore how useful different types of data are for *informing participants' expertise judgements*. Second, we investigated how useful different types of user-related data are for informing *computational expertise models of users*, which represent each user as a set of topic-weight pairs where a user is most knowledgeable in the topic with the highest weight. We used standard topic modeling algorithms to learn topics and annotate users with topics inferred from their *tweets*, *retweet*, *bio* and *list* memberships, and compared those topic annotations via information theoretic measures and a classification task.

Our findings reveal significant differences between various types of user-related data from an expertise perspective. The results provide implications that are not only relevant for expert recommender algorithms in the context of social media applications, but also for user interface designer of such applications. Although our experiments are solely based on Twitter, we believe that our results may also apply to other micro-blogging applications, and more broadly, to applications that allow users to create and organize their social network and share content with them.

The remainder of this paper is structured as follows: In Section 2 we discuss related work on modeling expertise of social media users. Section 3 describes our user study on how humans perceive and judge domain expertise of Twitter users. In Section 4 we present our experiments on modeling perceived expertise of Twitter users. We discuss our results in Section 5 and highlight implications of our work in Section 6. Section 7 describes limitations of our work and discusses ideas for future work. Finally, we conclude our work in Section 8.

II. Related Work

A widely used approach for identifying domain experts is peer-reviews [2]. Many state of the art expertise retrieval algorithms rely on this idea and often use content of documents people create, the relations between people, or a combination of both. For example, in [5] the authors use generative language models to identify experts among authors of documents. In [6] the authors explore topic-based models for finding experts in academic fields. The work presented in [7] uses network analysis tools to identify experts based on the documents or email messages they create within their organizations. In [8] the authors propose a probabilistic algorithm to find experts on a given topic by using local information about a person (e.g., profile info and publications) and co-authorship relationships between people. While previous research often neglects the variety of different types of data that can be observed for social media users, we focus on comparing different types of user-related data from an expertise perspective.

One of the key challenges of expert search algorithms is to accurately identify domains or topics related with users. Topic models are a state of art method for learning latent topics from document collections and allow annotating single documents with topics. Standard topic models such as LDA [9] can also be used to annotate users with topics e.g. by representing each user as an aggregation of all documents he/she authored. More sophisticated topic models, such as the Author Topic (AT) model [10] assume that each document is generated by a mixture of its authors' topic distributions. The Author Persona Topic (APT) model [11] introduces several personas per author because authors often have expertise in several domains and therefore also publish papers about different topics. The Author Interest Model [12] is similar to the APT model except that the personas are not local (i.e. not every user has an individual local set of personas) but global (i.e. all users share a common set of personas). In our work we do not introduce a new topic model, but empirically study how, and to what extent, existing topic modeling algorithms can be used to model the perceived expertise of Twitter users. Although our work is not the first work which applies topic models to Twitter (see e.g. [13] or [14]), previous topic modeling research on Twitter only took tweets into account, while we systematically compare different types of data that can be observed for Twitter users.

Recently, researchers started exploring different approaches for identifying experts on Twitter. For example, in [15] the authors present TwitterRank, an adapted version of the topic sensitive PageRank, which allows identifying topical influential Twitter users based on follow relations and content similarities. In [16] the authors compare different network-based features and content/topical features to find authoritative users. To evaluate their approach they conducted a user study and asked participants to rate how interesting and authoritative they found the author and his/her tweets. The work of [3] presents an approach to find topical relevant Twitter users by combining standard Twitter text search mechanism with information about the social relationships in the network and evaluate their approach via Amazon Mechanical Turk. Previous research agrees on the fact that one needs both, content and structural network features, for creating a powerful expert retrieval algorithm. However, to our best knowledge most existing expert retrieval work on Twitter limits their content features to tweets, while our results suggest that tweets are inferior for making sense of the expertise of Twitter users compared to other types of user-related data.

The issue of how users perceive the credibility of microblog updates is only just beginning to receive attention. In [1] the authors present results from two controlled experiments which were designed to measure the impact of three features (user image, user name and message content) on users' assessment of tweet credibility. Unlike our work, Morris et al. explore which factors influence users' perception of the credibility of a tweet, while we focus on users' perception of other users expertise. Further, our study sets out to gain empirical insight into the usefulness of different types of data (such as tweets, retweets, user lists and bio information) for informing expertise or credibility judgements of users, while their experiments aim to identify the factors which influence such judgments. That means, while Morris et al. manipulate data (i.e., tweets, user images and user names) within their experiment to measure the impact of their manipulation on users' judgments, we do not manipulate any user-related data, but manipulate the type and amount of data we show. Similar to our results their results indicate that users have difficulty discerning trustfulness based on content alone. In [17] the authors do not examine expertise or credibility per se. In their study they asked users to rate how "interesting" a tweet was and how "authoritative" its author was, manipulating whether or not they showed the author's user name. In our work we decided not to show user names at all amongst others for the following reasons: first, showing user names may add uncontrolled noise to our experiment since participants may recognize some of the users to judge. Therefore their expertise judgments would be based on their background knowledge rather than on the information which is shown to them during the user study. Second, algorithms and automated methods can not exploit user names but will require further information related with those names to gauge users' potential expertise. Since our aim was to create expertise models of users, our experiment set out to evaluate only information which can be accessed and exploited by humans and automated methods.

III. User Study

We conducted a user study to explore how useful different types of user-related social media data are for *informing humans' expertise judgements* about Twitter users. To that end, we compare the ability of participants to correctly judge the expertise of Twitter users when the judgement is based on the contents they published (*tweets and retweets*), self-reported and externally-reported contextual information (*bio* and *user lists*), or both contents and contextual information.

A. Participants

We chose "semantic web" to be the topic in the experiment. We recruited a group of 16 participants consisting of users with rather basic and high knowledge about the topic semantic web. We recruited 8 participants by contacting the faculties and students of the International Summer School on Semantic Computing 2011 held at UC Berkeley and 8 participants from a university town in the United States. Participants' age ranged from 20 to 34.

B. Design and Procedure

We used Wefollow[2] to select candidate Twitter users to be judged. Wefollow is a user powered Twitter directory where users can sign up for at most 3 directories. Wefollow ranks all users based on a proprietary algorithm which takes amongst others into account how many users in a certain directory follow a user. Users who are followed by more users signed up for a topic directory get a higher rank of the particular topic. At the time we crawled Wefollow (July 2011), the Wefollow directory of the topic "semantic web" suggested 276 Twitter users relevant to the topic. For candidates to represent high level of expertise, we randomly selected six users from rank 1–20 and six users from rank 93–113. For candidates of low expertise, we randomly selected six users from rank 185–205 and six users from the public Twitter timeline who did not show any relation to the topic. To validate the manipulation, we also conducted a pilot study by asking 3 raters to compare the expertise of 50 pairs of candidates randomly selected from the high and low expertise group. The results showed that all of them had 95% or higher agreement with our manipulation, and the inter-rater agreement was 0.94. This result proved that our expertise level manipulation was successful.

Our experiment tested three conditions: 1) participants saw the latest 30 messages published by a user (i.e., the user's most recent tweets and retweets) and contextual information including the user's bio information and his/her latest 30 user list memberships ; 2) participants saw only the latest 30 tweets and retweets of a user; 3) participants saw only the bio and the latest 30 list memberships (or all list memberships if fewer than 30 were available). Each of the 24 pages which we selected in step one was randomly assigned to one of the three conditions. In other words, for each condition, we had four Twitter user candidates of high expertise and four Twitter user candidates of low expertise. To tease out the influence of the Twitter interface and further uncontrolled variables such as user images or user names, we presented only the plain textual information in a table. The users' names, profile pictures and list creators' names were removed to avoid the influence of uncontrolled variables. For condition 1 the table had two randomly ordered columns to present tweets and contextual information separately. For condition 2 and 3 the table only had one column to present everything.

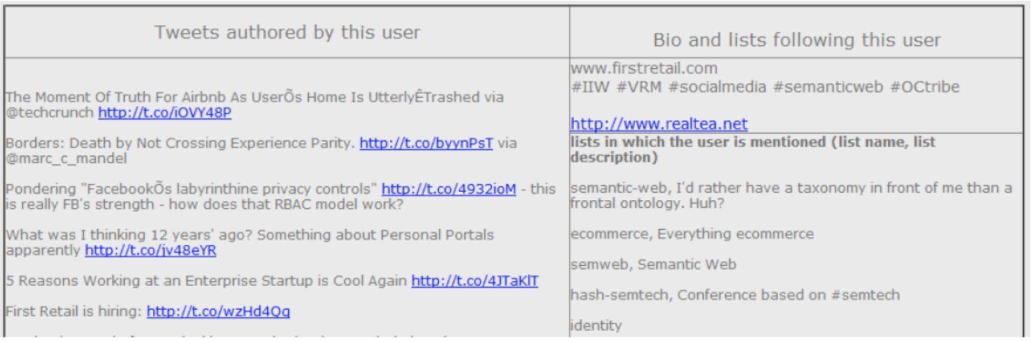
Before the task, participants were asked to answer demographical questions and complete a knowledge test. Then they were presented with 24 evaluation tasks (three conditions, eight pages for each condition) in sequence. They were told that the information in the table was derived from a real Twitter user, and asked to rate how much this person knew about the topic, semantic web, on a one (least) to five (most) scale. The tasks took about 30-40 minutes.

C. Results

We analyzed participants' expertise ratings by performing two-way repeated measure ANOVA with Twitter user expertise (high/low) and conditions (content and contextual information/only content/only contextual information) as within subjects variables.

Interestingly, there is an interaction between conditions and Twitter user expertise ($F(2,30) = 8.326$, $p < 0.01$). It means there exists significant differences in users' ability

[2] www.wefollow.com



Tweets authored by this user	Bio and lists following this user
The Moment Of Truth For Airbnb As UserÕs Home Is UtterlyÊTrashed via @techcrunch http://t.co/iOVY48P Borders: Death by Not Crossing Experience Parity. http://t.co/byynPsT via @marc_c_mandel Pondering "FacebookÕs labyrinthine privacy controls" http://t.co/4932ioM - this is really FB's strength - how does that RBAC model work? What was I thinking 12 years' ago? Something about Personal Portals apparently http://t.co/jv48eYR 5 Reasons Working at an Enterprise Startup is Cool Again http://t.co/4JTaKlT First Retail is hiring: http://t.co/wzHd4Qq	www.firstretail.com #IIW #VRM #socialmedia #semanticweb #OCtribe http://www.realtea.net **lists in which the user is mentioned (list name, list description)** semantic-web, I'd rather have a taxonomy in front of me than a frontal ontology. Huh? ecommerce, Everything ecommerce semweb, Semantic Web hash-semtech, Conference based on #semtech identity

Fig. 1. Example of the experimental task under condition 1. Randomly ordered tables and plain text without pictures and usernames were used to present different types of user-related data to participants.

of differentiating high and low expertise across these three conditions. To understand the difference, we compared each pair of conditions by performing the same ANOVA test. When comparing between condition 1, where participants saw both content and contextual information, and condition 2, where participants' expertise judgments were only informed by content, participants were significantly more able to make the correct judgment in condition 1 ($F(1, 15) = 23.39$, $p < 0.01$). When comparing condition 3, where participants' judgments were informed by contextual information, to condition 2, where participant's expertise judgments were only informed by content, participants made significantly better judgments in condition 3 ($F(1, 15) = 5.91$, $p = 0.03$). There was no significant difference observed between condition 1 and condition 3 ($F(1, 15) = 2.19$, $p = 0.16$). These results indicated that participants made the worst expertise judgments when the judgments were based on tweets and retweets only. Interestingly, participants' expertise judgments, when only based on contextual information (i.e., information about users' bio and list memberships), were almost as good as judgments based on both content and contextual information. To illustrate the interaction, we plot participants' average ratings in different conditions in Figure 2. The slopes in Figure 2 reflect the ability of participants to differentiate between Twitter users of high and low expertise in different conditions.

Our findings highlight the low quality of topical expertise judgement based solely on tweets' and retweets' contents. It implies that there is a large variance of information in what people tweet and retweet about. Experts of a particular topic do not necessarily publish or re-published content about the topic all the time, if any. In contrast, contextual information such as bio and user list memberships provides salient and straightforward cues for expertise judgements since these cues often provide descriptive information about the person himself, such as personal interests, professional experience, community the person belongs to, etc.

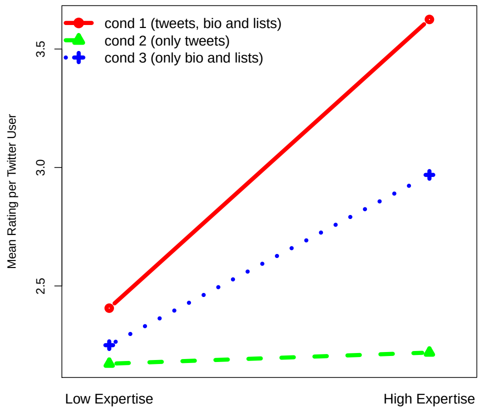


Fig. 2. Average expertise ratings given to Twitter users with high/low expertise by participants in each condition. The slope of each line indicates the ability of participants to differentiate between experts and novices.

IV. Experiments

Since our user study supported our hypothesis that different types of user-related data differ in their ability to inform *humans' expertise judgments* we further aim to compare how useful different types of data are for learning *computational expertise models* of Twitter users by using topic modeling. Therefore, we first compare topic distributions of users inferred from different types of user-related data, namely *tweets*, *retweets*, *bio* and *user list* data and study if those topic distributions differ substantially on average. Second, we explore to what extent different topic distributions reflect users' perceived expertise categories by using information theoretic measures and by casting our problem as a user classification task.

A. Dataset

For our experiments we selected the following 10 topics (including rather general and rather specific topics and topics with high and low polarity): semanticweb, biking, wine, democrat, republican, medicine, surfing, dogs, nutrition and diabetes. For each topic we selected the top 150 users from the corresponding Wefollow directory (i.e., the 150 user with the highest rank). We excluded users whose required information (i.e. tweets, retweets, lists memberships and biographical information) were not available to crawl. We also excluded users who appeared in more than one of the 10 Wefollow directories and users who mainly do not tweet in English. For all remaining 1145 users we crawled at maximum their last 1000 tweets and retweets, the last 300 user lists to which they were added and their bio info. Tweets, retweets and bio information often contain URLs. Since information on Twitter is sparse, we enriched all URLs with additional information (title and keywords) obtained from the meta-tags in the headers of webpages they are pointing to. User list names and descriptions usually do not contain URLs, but list names can be used as search query terms to find web documents which reveal further information about the potential meaning of list labels. We used the top 5 search query result snippets obtained from Yahoo Boss[3] to enrich list information. After enriching our dataset, we removed standard English stopwords and performed stemming using Porter's algorithm [18].

B. Topic Models

Topic models are a powerful suite of algorithms which allow discovering the hidden semantic structure in large collection of documents. The idea behind topic models is to model documents as arising from multiple topics, where each document has to favor few topics. Therefore, each document exhibits different topic proportions and each topic is defined as a distribution over a fixed vocabulary of terms and has as well to favor few words.

Topic models treat our data as arising from a generative process that includes hidden variables. This generative process defines a joint probability distribution over both the observed and hidden random variables. Given this joint distribution one can compute the conditional distribution of the hidden variables given the observed variables. This conditional distribution is also called the posterior distribution.

The most basic topic modeling algorithm, Latent Dirichlet Allocation (LDA) [9], encodes the following generative process: First, for each document d a distribution over topics θ is sampled from a Dirichlet distribution α. Second, for each word w in the document d, a single topic z is chosen according to its document specific topic distribution θ. Finally, each word w is sampled from a multinomial distribution over words ϕ which is specific for the sampled topic z.

Fitting an LDA model to a collection of training documents requires finding the parameters which maximize the posterior distribution $P(\phi, \theta, z|\alpha, \beta, w, \dot{)}$ which specifies a number of dependencies that are encoded in the statistical assumptions behind the generative process. In our experiments we used MALLET's [19] LDA implementation and aggregated all user-related data into artificial user-documents which we used to train the model. We chose the default hyperparameters ($\alpha = 50/T$, $\beta = 0.01$ and the number of topics T= 10, 30, 50, 100, 200, 300, 400, 500, 600 and 700) and optimized them during training by using Wallach's fixed point iteration method [20]. Based on the empirical findings of [21], we decided to place an asymmetric Dirichlet prior over the topic distributions and a symmetric prior over the distribution of words. Given enough iterations (we used 1500) the Markov chain (which consists of topic assignments z for each token in the training corpus) has potentially converged and we can get estimates of the word distribution of topics ($\hat{\phi}$) and the topic distribution of documents ($\hat{\theta}$) by drawing samples from the chain. The estimated distributions $\hat{\phi}$ and $\hat{\theta}$ are predictive distributions and are later used to infer the topics of users via different types of user-related data. Figure 3 shows some randomly selected sample topics learned via LDA when the number of topics was 50 ($T = 50$).

C. Evaluation Metrics

To answer whether different types of data related to a single user lead to substantially different topic annotations, we compare the average Jensen-Shannon (JS) divergence between pairs of topic annotations inferred from different types of data related with a single user. We always use the average topic distributions inferred via 10 independent runs of a Markov chain as topic annotations. The JS divergence is a symmetric measure of the similarity between two distributions. The JS divergence is 0 if the two distributions are identical and approaches infinity as they differ more and more. The JS divergence is defined as follows:

$$D_{JS} = \frac{1}{2}D_{KL}(A||B) + \frac{1}{2}D_{KL}(B||A) \tag{1}$$

where $D_{KL}(A||B)$ represents the KL divergence between random variable A and B. The KL divergence is calculated as follows:

$$D_{KL}(A||B) = \sum_{i} A(i) \log \frac{A(i)}{B(i)} \tag{2}$$

To address the question which user-related data are more suitable for creating topical expertise profiles of users, we aim to estimate the degree to which different types of users' topic annotations reflect their perceived expertise. Since we know the ground truth label of all 1145 users in our dataset, we can compare the quality of different topic annotations by measuring how likely the topics agree with our true expertise category labels. Here, we use Normalized Mutual Information (NMI) between users' topic distribution (θ_{user}) and the topic distribution of users' Wefollow directories (θ_{label}) which is defined as the average topic distribution of all users in that directory.

[3] http://boss.yahoo.com

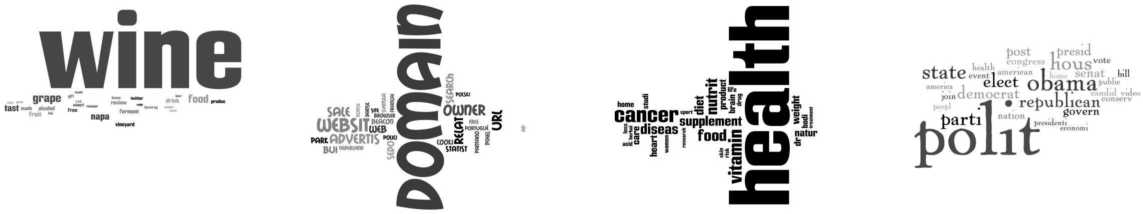


Fig. 3. Top 20 stemmed words of 4 randomly selected topics learned via LDA with number of topics $T = 50$.

$$NMI(\theta_{label}, \theta_{user}) = \frac{I(\theta_{label}, \theta_{user})}{[H(\theta_{label}) + H(\theta_{user})]/2} \quad (3)$$

$I(\theta_{label}, \theta_{user})$ refers to the Mutual Information (MI), $H(\theta_{label})$ refers to the entropy of the Wefollow-directory-specific topic distribution and $H(\theta_{user})$ refers to a user-specific topic distribution which is inferred based on each of the four different types of user-related data.

$$I(\theta_{label}, \theta_{user}) = H(\theta_{user}) - H(\theta_{user}|\theta_{label}) \quad (4)$$

NMI is always between 0 and 1. A higher NMI value implies that a topic distribution more likely matches the underlying category information. Consequently, NMI is 1 if the two distributions are equal and 0 if the distributions are independent.

Finally, we aim to compare different types of topic annotations within a task-based evaluation. We consider the task of classifying users into topical categories (in our case Wefollow directories) and use tweet-, bio-, list-and retweet-based topic annotations as features to train a Partial Least Square (PLS) classifier[4]. We decided to use PLS, since our features are highly correlated and the number of features can be relative large (up to 700) compared to the number of observations for each trainings split (consisting of 916 users). PLS regression is particularly suited in such situations. Within a 5-fold-cross evaluation we compare the classification performance by standard evaluation measures such as Precision, Recall, F-Measure and Accuracy.

D. Results

In this section, we present our empirical evaluation of perceived expertise models of users based on different types of user activities and their outcomes. Firstly we investigate how similar topic distributions of an individual user inferred from different types of user-related data are on average. Secondly we explore how well different types of topic distributions capture the perceived expertise of users.

First, we aimed to explore whether the topic distributions of a single user inferred from different types of user-related data are differ substantially. Therefore, we compared the average JS divergence of different topic distributions inferred via different types of user-related social media data. Figure 4 shows that different types of user-related data lead to different topic annotations. Not surprisingly, we find that tweet- and retweet-based topic annotations are very similar. Further, bio- and tweet- and bio- and retweet-based topic distributions show high similarity, while list- and bio- and list- and tweet- and list- and retweet-based topic distributions are more distinct. This suggests that users with high Wefollow rank tend to tweet and retweet about similar topics and that they also mention these topics in their bio (or the other way around). Users' list memberships however do not necessarily reflect what users tweet or retweet about or the topics they mention in their bio, amongst others for the following three reasons: First, sometimes user lists describe how people feel about the list members (e.g., “great people”, “geeks”, “interesting twitterers”) or how they relate with them (e.g., “my family”, “colleagues”, “close friends”). Consequently, these list labels and descriptions do not reveal any information about the topics a user might be knowledgable about. Second, some user lists are topical lists and may therefore reveal information about

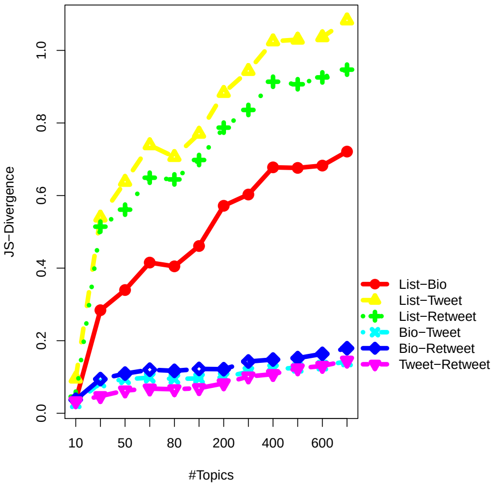


Fig. 4. Average JS-Divergence of 1145 Wefollow users' topic annotations inferred via their tweets, retweets, bio and list information.

[4] http://cran.r-project.org/web/packages/pls/

the topics other users associate with a given user. However, these topical associations can also be informed by exogenous factors, meaning a given user does not necessarily need to use Twitter to share information about a topic in order to be associated with that topic by other users. Third, since everyone can create user lists and add users to these list, spam can obviously be a problem, especially for popular users.

To get an initial impression of the nature of user list labels and descriptions, we randomly selected 455 lists memberships of 10 randomly selected users (out of our 1145 users) and we asked 3 human raters to judge whether a list label and its corresponding descriptions may reveal information about expertise domains or topics in general. To give an example: list labels such as "my friends" or "great people" do not reveal any information about the expertise of users in that list, while list labels such as "healthcare professionals" or "semanticweb" may help to gauge the expertise of users who are members of that lists. Our results suggest that 77.67% of user lists reveal indeed information about potential expertise topics of users with a fairly good inter-rater agreement ($\kappa = 0.615$).

Second, we explored how useful different types of user-related data are for inferring the perceived expertise of users by estimating how likely the topics agree with the true expertise category labels of users. So far we only know that it makes a difference which type of user-related data we use for inferring topic annotations of users. However, we don't know which types of data lead to "better" topic annotations of users, where better means that a topic distribution captures the perceived expertise of a user more accurately. Since we have a ground truth label of all users in our dataset (their Wefollow directories), we can estimate the quality of different topic annotations by measuring how likely the topics agree with the true category labels. Here, we used the Normalized Mutual Information (NMI) between users' topic distribution based on different types of data and the topic distribution of a users' Wefollow directory which is defined as the average topic distribution of all users in that directory. A higher NMI value implies that a topic distribution might more likely match the underlying category information. Figures 5 shows that list-based topic annotations tend to have higher NMI values than retweet-, tweet- and bio-based topic annotations. It suggests that list based topic annotations reflect the underlying category information best. In other words, users in a given Wefollow directory tend to be in topical similar lists, while the topics they tweet or retweet about or mention in their bio are more distinct. Firstly, this suggests that users assign other users to lists about topics which tend to reflect their self-view, because users have to register themselves for certain topics in Wefollow. Secondly, it indicates that users make these list assignments not only based on the content authored by the users they assign. They also seem to use background knowledge or other types of external information sources to inform their list assignments. As expected, the NMI values become lower with increasing number of topics.

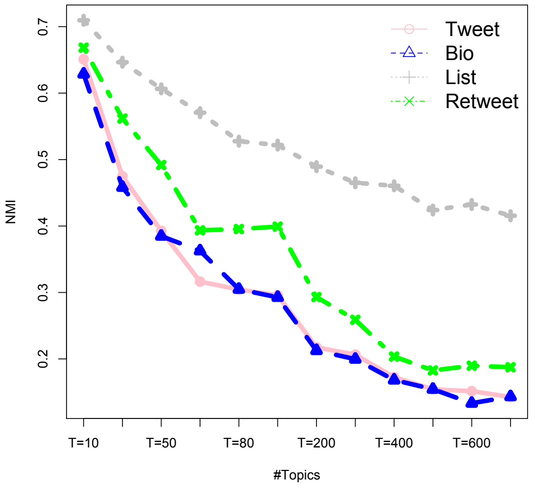


Fig. 5. Average Normalized Mutual Information (NMI) between 1145 users' tweet-, retweet-, list- and bio-based topic annotations and users' Wefollow directory

1) User Classification Experiment: To further quantify the ability of different types of topic annotations to reflect the underlying ground truth category information of users, we performed a task-based evaluation and considered the task of classifying users into topical categories such as Wefollow directories. We used tweet-, bio-, list- and retweet-based topic annotations as features, trained a Partial Least Square classifier and performed a 5-fold-cross validation to compare the performance of different trainings schemes. Note that since we used topic distributions as features rather than term vectors the number of features corresponds to the number of topics and does not depend on the length of different user-related data such as bio, tweet, retweet and list data. In other words, the number of features used for tweet-, bio-, list- and retweet-based classifiers were equal although different types of user-related data may differ in their content length.

Figure 6 shows the average F-measures and Accuracy of the classifier trained with different number of topics (T= 10, 30, 50, 70, 80, 100, 200, 300, 400, 500, 600 and 700) inferred via different types of user-related data. One can see from these figures that no matter how fine-grained topics are (i.e., how one chooses the number of topics), list-based topic annotations always outperform topic-annotations based on other types of user-related data.

We also compared the average classifier performance for individual Wefollow directories. Figure 6 shows the average F-measures and Accuracy of the classifier for each Wefollow directory. We averaged the classifier performance for each Wefollow directory over the results we got from the 5-fold cross validations of classifiers trained with different number topics inferred via different types of user-related data for each class. One can see from these figures that for all classes a classifier trained with list-based topic annotations performs best, i.e. yields to higher F-measures and Accuracy values than classifiers trained with other types of user-related data. However, for certain classes such as democrats or republicans the F-measures of all classifiers are very low, also if trained with list-based topic annotations. It suggests that although list-based topic annotations are best for classifying users into mutual exclusive topical expertise directories, for very similar topics information about users' list memberships might not be detailed enough. For example users which seem to have high knowledge about democrats or republicans, are all likely to be members of similar lists such as "politicians" or "politics".

To explore the classifiers' performance in more detail we also inspected their confusion matrices. Figure 7 shows the confusion matrices of a classifier trained with topic distributions over 30 topics (first row) and 300 topics (second row) inferred via different types of user-related data as features. Note that a perfect classifier would lead to a red map with a white diagonal. The confusion matrix for a classifier trained with list-based topic annotations (Figure 7) shows the closest match to the ideal and hence indicates least confusion. Again, on can see that confusion mainly happens for very similar classes such as democrats and republicans, since those users are likely to be members of similar lists.

V. Discussion

Judging expertise of social media users will continue to represent a relevant and challenging research problem and also an important task for social media users since judging topical expertise of other users is a key challenge in maximizing the credibility and quality of information received.

Through our experiments and our user study, we showed that different types of user-related data differ substantially in their ability to inform computational expertise models of Twitter users and expertise judgements of humans. We argue that these findings represent an important contribution to our research community since in past research topical user profiles are often learned based on an aggregation of all documents a user has authored or is related with, without taking the differences between various types of user activities and related outcomes into account.

Our experiments demonstrate that the aggregation of tweets authored or retweeted by a given user is less suitable for inferring the expertise topics of a user than information about users' list memberships. In addition, our user study clearly confirms that it is as well difficult for humans to identify experts based on their tweets and retweets. Further, our results show that topic annotations based on users' list memberships are most distinct from topic annotations based on other types of user-related data. Topic annotations based on bio information are however surprisingly similar to topic annotations based on the aggregation of tweets and retweets, which indicates that users tend to tweet and retweet messages about topics they mention in their bio or the other way around. This is interesting from a practical point of view, since it suggests that computational expertise models of users which just rely on their bio information achieve similar accuracy as models which are based on the aggregation of their tweets or retweets.

VI. Implications

Our experimental findings suggest that users' have difficulties in judging users' expertise based on their tweets and retweets only. Therefore, we suggest that user interface designer should take this into account when designing users' profile pages. We suspect that Twitter users' profile pages are amongst others used to inform users about the expertise, interests, authoritativeness or interestingness of a Twitter user. Therefore those type of information which facilitates these judgements should be most prominent.

Further, our results suggest that computational expertise models benefit from taking users' list memberships into account. Therefore, we argue that also expert-recommender systems and user-search systems should heavily rely on user list information. Further we argue that also social media provider and user interface designer might want to think of promoting and elaborating list features (or similar features which allow to tag or label other users) more, since user list information seems to be very useful for various tasks.

VII. Limitations and Future Work

The result of our user study is limited to a small subject population and one specific topic, semantic web. Readers who try to generalize our results beyond Twitter should also note that the motivation of users for using a system like Twitter in general and their motivation for creating user lists in specific, may impact how useful information about list memberships are for the expertise modeling task. On Twitter we found that indeed a large percentage of lists may potentially reveal information about the expertise of users assigned to the list. However, this can be different on other social media systems. Nevertheless, our results highlight the potential of user lists and if lists are used for different purpose automated methods can be applied in order to group lists by its purpose.

Our work highlights that different types of social media data reveal different types of information about users and therefore enable different implications. We will explore this avenue of work by investigating which implications different types of activities and related outcomes may enable and how they can be combined for creating probabilistic user models.

VIII. Conclusions

Information consumption on social media is mainly driven by social networks and credibility judgements of content are mainly informed by credibility judgements of authors [1]. Therefore, judging topical expertise of other users is a key challenge in maximizing the credibility and quality of information received. In this work we examined the usefulness of different types of user-related data (concretely we used *tweets*, *retweets*, *bio* and *user list memberships*) for making sense of the domain expertise of Twitter users. Our results

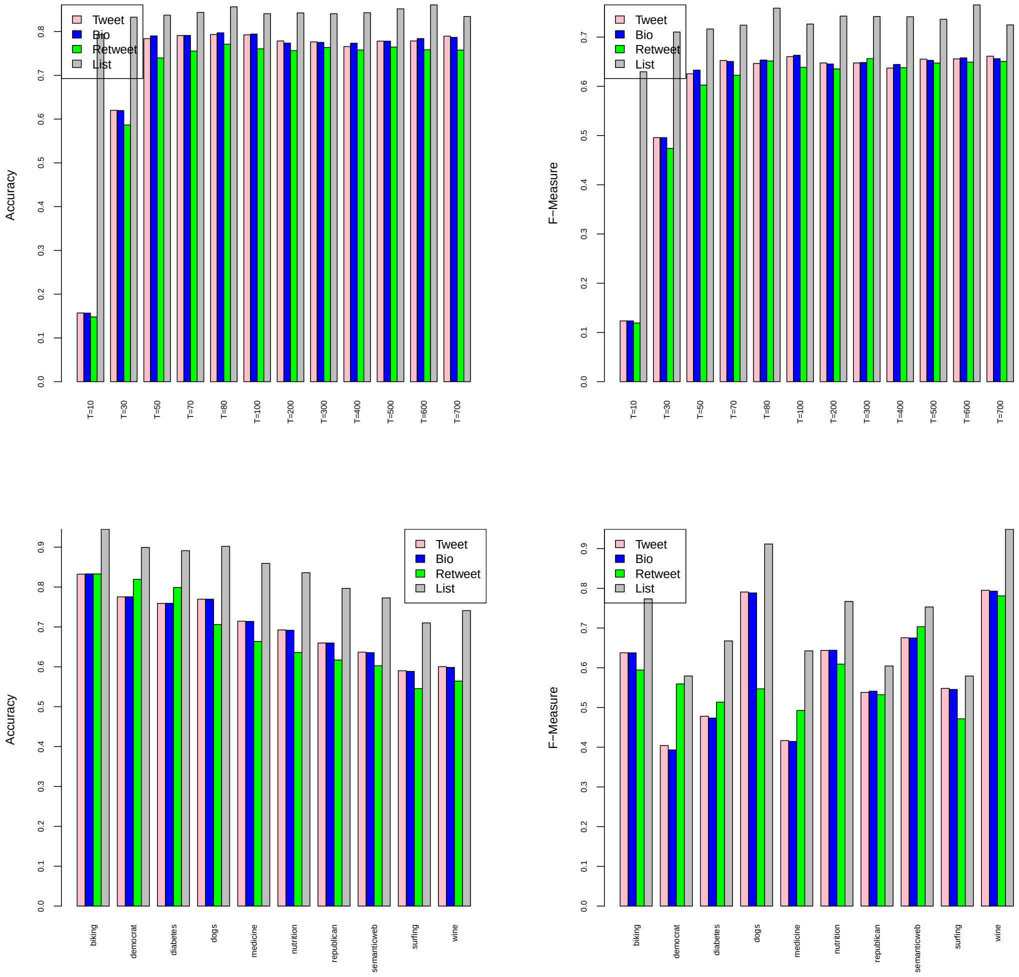


Fig. 6. Average Accuracy and F-measure of PLS classifier trained with bio-, list-, retweet-, and tweet-based topic distributions. The x-axes of the figures in the first row show the number of topics per distributions. The x-axes of the figures in the second row show the 10 Wefollow directories (biking, democrat, diabetes, dogs, medicine, nutrition, republican, semanticweb, surfing, and wine). The y-axes show the accuracy or F-measure of the classifier averaged over 5 folds and different numbers of topics (T=10, 30, 50, 70, 80, 100, 200, 300, 400, 500, 600, 700) or the 10 Wefollow directories.

suggests that different types user-related social media data are useful for different computational and cognitive tasks, and the task of expertise modeling benefits most from information contained in user lists as opposed to tweet, retweet or bio information. We hope our findings will inform the design of future algorithms, user interfaces and methods that focus on capturing expertise of social media users and stimulate research on making sense of different types of user-activities and related outcomes.

Acknowledgment

This work was supported in part by Office of Naval Research Contract No. N00014-08-C-0029 to Peter Pirolli and by a DOC-fForte fellowship of the Austrian Academy of Science to Claudia Wagner.

References

[1] M. R. Morris, S. Counts, A. Roseway, A. Hoff, and J. Schwarz, “Tweeting is believing?: understanding microblog credibility perceptions,” in *Proceedings of the ACM 2012 conference on Computer Supported Cooperative Work*, ser. CSCW ’12. New York, NY, USA: ACM, Feb. 2012, pp. 441–450.

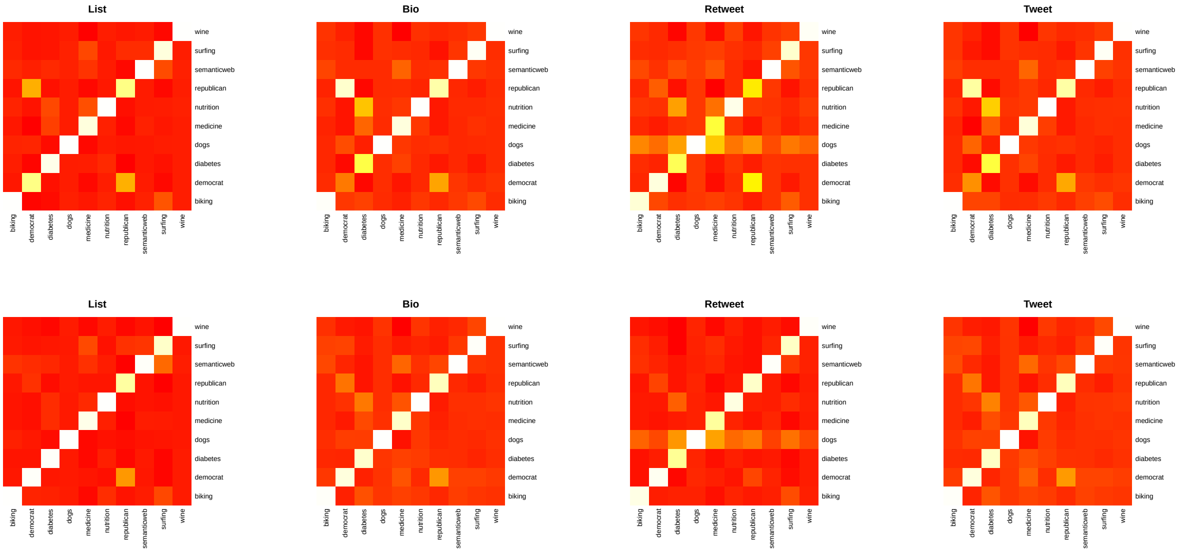


Fig. 7. Average confusion matrices (across 5 folds) of a classifier trained with bio-, list-, retweet-, and tweet-based topic annotations with 30 topics (first row) and 300 topics (second row). The x-axis of each confusion matrix shows the reference values and the y-axis shows the predictions for the10 Wefollow directories (biking, democrat, diabetes, dogs, medicine, nutrition, republican, semanticweb, surfing, and wine). The lighter the color the higher the value.

[2] K. A. Ericsson, *The Cambridge Handbook of Expertise and Expert Performance*. Cambridge: Cambridge University Press, 2006, ch. Protocol Analysis and Expert Thought: Concurrent Verbalizations of Thinking during Experts‘ Performance on Representative Tasks, pp. 223–241.

[3] K. R. Canini, B. Suh, and P. Pirolli, “Finding relevant sources in twitter based on content and social structure,” in *NIPS Workshop*, 2010.

[4] Q. Liao, C. Wagner, P. Pirolli, and W.-T. Fu, “Understanding experts’ and novices’ expertise judgment of twitter users,” in *Proceedings of the 30th ACM conference on Computer-Human Interaction(CHI)*, 2011.

[5] K. Balog, L. Azzopardi, and M. de Rijke, “Formal models for expert finding in enterprise corpora,” in *Proceedings of the 29th annual international ACM SIGIR conference on Research and development in information retrieval*, ser. SIGIR ’06. New York, NY, USA: ACM, 2006, pp. 43–50.

[6] H. Deng, I. King, and M. R. Lyu, “Formal models for expert finding on DBLP bibliography data,” in *Proceedings of International Conference on Data Mining*, 2008.

[7] J. Davitz, J. Yu, S. Basu, D. Gutelius, and A. Harris, “ilink: Search and routing in social networks,” in *Proceedings of The Thirteenth International Conference on Knowledge Discovery and Data Mining (KDD 2007)*, August 2007.

[8] J. Li, J. Tang, J. Zhang, Q. Luo, Y. Liu, and M. Hong, “Eos: expertise oriented search using social networks,” in *Proceedings of the 16th international conference on World Wide Web*, ser. WWW ’07. New York, NY, USA: ACM, 2007, pp. 1271–1272.

[9] D. M. Blei, A. Ng, and M. Jordan, “Latent dirichlet allocation,” *JMLR*, vol. 3, pp. 993–1022, 2003.

[10] M. Rosen-Zvi, T. Griffiths, M. Steyvers, and P. Smyth, “The author-topic model for authors and documents,” in *Proceedings of the 20th conference on Uncertainty in artificial intelligence*, ser. UAI ’04. Arlington, Virginia, United States: AUAI Press, 2004, pp. 487–494.

[11] D. Mimno and A. McCallum, “Expertise modeling for matching papers with reviewers,” in *KDD*, 2007.

[12] N. T. T. Comware, “Author interest topic model,” *Notes*, pp. 887–888, 2010.

[13] D. Ramage, S. Dumais, and D. Liebling, “Characterizing microblogs with topic models,” in *Proceedings of the International AAAI Conference on Weblogs and Social Media (ICWSM)*, 2010.

[14] L. Hong and B. D. Davison, “Empirical study of topic modeling in twitter,” in *Proceedings of the First Workshop on Social Media Analytics*, ser. SOMA ’10. New York, NY, USA: ACM, 2010, pp. 80–88.

[15] J. Weng, E.-P. Lim, J. Jiang, and Q. He, “Twitterrank: finding topic-sensitive influential twitterers,” in *Proceedings of the third ACM international conference on Web search and data mining*, ser. WSDM ’10. New York, NY, USA: ACM, 2010, pp. 261–270.

[16] A. Pal and S. Counts, “Identifying topical authorities in microblogs,” in *Proceedings of the fourth ACM international conference on Web search and data mining*, ser. WSDM ’11. New York, NY, USA: ACM, 2011, pp. 45–54.

[17] A. Paland and S. Counts, “What’s in a @name? how name value biases judgment of microblog authors.” in *Proceedings of the International AAAI Conference on Weblogs and Social Media (ICWSM)*. The AAAI Press, 2011.

[18] M. Porter, “An algorithm for suffix stripping,” *Program*, vol. 14, no. 3, pp. 130–137, 1980.

[19] A. K. McCallum, “Mallet: A machine learning for language toolkit,” 2002, http://mallet.cs.umass.edu.

[20] H. M. Wallach, “Structured topic models for language,” Ph.D. dissertation, University of Cambridge, 2008.

[21] H. M. Wallach, D. Mimno, and A. McCallum, “Rethinking LDA: Why priors matter,” in *Proceedings of NIPS*, 2009.

Religious Politicians and Creative Photographers: Automatic User Categorization in Twitter

Claudia Wagner*, Sitaram Asur† and Joshua Hailpern†
*Institute of Information and Communication Technologies, JOANNEUM RESEARCH, Graz, Austria
Email: claudia.wagner@joanneum.at
†HP labs, Palo Alto, California
Email: sitaram.asur@hp.com, joshua.hailpern@hp.com

***Abstract*—Finding the "right people" is a central aspect of social media systems. Twitter has millions of users who have varied interests, professions and personalities. For those in fields such as advertising and marketing, it is important to identify certain characteristics of users to target. However, Twitter users do not generally provide sufficient information about themselves on their profile which makes this task difficult. In response, this work sets out to automatically infer professions (e.g., musicians, health sector workers, technicians) and personality related attributes (e.g., creative, innovative, funny) for Twitter users based on features extracted from their content, their interaction networks, attributes of their friends and their activity patterns. We develop a comprehensive set of latent features that are then employed to perform efficient classification of users along these two dimensions (profession and personality). Our experiments on a large sample of Twitter users demonstrate both a high overall accuracy in detecting profession and personality related attributes as well as highlighting the benefits and pitfalls of various types of features for particular categories of users.**

I. Introduction

Manually built directory systems, such as *Wefollow*[1], have been created to allow end users to find new Twitter users to follow. In these ecosystems, users place themselves in explicitly defined categories (e.g. musician, doctor, or inspirational). While these categories are quite broad (e.g. geographical locations, brands, interests, personalities), their manual creation limits their scope. With an estimated 500 million users [1] on Twitter, only a small percentage of these handles are in systems like *Wefollow* – and therefore, the majority of Twitter users are not categorized. This information would be invaluable to both the end user, as well as advertisers.

This work directly addresses this need by focusing on automatic detection of attributes for users on Twitter. In particular, we focus on profession types/areas of Twitter users (e.g., musicians, entrepreneurs or politicians), and on online personality attributes of Twitter users (e.g., creative, innovative, funny). Because Twitter user metadata is highly limited, this work's analysis is based upon a comprehensive set of features that can be summarized into four key groups: *linguistic style*, *semantics*, *activity patterns* and *social-semantics*.

Our contributions in this paper are two-fold. First, we construct a comprehensive collection of features and examine their efficacy in classifying users on Twitter. Second, we consider two dimensions of user attributes, personality and profession, and show how efficient user classification can be performed on them. In order to create and explore these models we conducted extensive experiments using a corpus of more than 7k labeled Twitter users, crawled from *Wefollow*, and more than 3.5 million tweets. On this rich data set we trained Random Forest classifiers, and used correlation-based feature selection to prune correlated features.

Overall, the classifiers built on an optimal subset of features achieved an impressive accuracy ≥ 0.9 for most categories. When we examine the features independently, we observed that, not surprisingly, the feature groups differ in their accuracy across the various categories. For example, we found that social-semantic features are very efficient while classifying personality related attributes but achieve lower accuracy in the case of professions. Linguistic style features tend to work well with both personality and professions. However, not all features were universally useful for all categories. Specifically we found that *what a user says* (semantics) and *how a user behaves online* (activity patterns) tend to reveal less information about his professional areas and personality compared to *how a user says something* (linguistic style) and *what others say about/to another user* (social-semantic).

II. Related Work

Rao et al. [2] classify Twitter users according to a set of latent user attributes, including gender, age, regional origin, and political orientation. They show that message content is more valuable for inferring the gender, age, regional origin, and political orientation of a user than the structure or communication behavior of his/her social network. Rather than performing general user classification, [3] specifically models a user's political affiliation, ethnicity and affinity for one specific business, namely Starbucks. While their approach combines both user-centric features (profile, linguistic, behavioral, social), and social graph based features, their results suggest that user-centric features alone achieve good classification results, and social graph information has a negligible impact on the overall performance.

Our goal is to classify users along much broader planes of categories and we construct a comprehensive list of features for this purpose. Though the user attributes which we analyze in this work are substantially different from those analyzed in

[1] http://wefollow.com

[2] and [3], we also find content-based features more useful than activity features which capture amongst other structural similarities between users. However, unlike previous work we examine not only content which can directly be associated with a user (via an authorship relation) but also content which can indirectly be related with a user via other users.

Hong et al. [4] compare the quality and effectiveness of different standard topic models for analyzing social data. Their results suggest that topic features tend to be useful if the information to classify is sparse (message classification task), but if enough text is available (user classification task) simple TFIDF weighted words perform better. In our work we do not only compare the performance of TFIDF and topic models within the user classification task, but also explore the utility of explicit ontological concepts.

Unlike the above work which focused on individuals, [5] examine how networks emerging from user communication are closely replicated in the frequency of words used within these communities. In short, users who are strongly connected also talk about similar subjects and therefore use similar words. In addition, [5] also reveal that users who belong to one community tend to show similarities in the length of words they use or in their three letter word ending usage. This suggests that socio-linguistic features may help differentiate users in different communities. Therefore, we decided to incorporate linguistic style features which may have the potential to identify users who belong to the same community (e.g. group of users working in the same professional area or group of users sharing some personality characteristics).

Recently the prediction of personality related attributes of social media users gained interest in the research community [6] [7] [8], since characterizing users on this dimension would be useful for various applications such as recommender systems or online dating services. For example, [7] gathered personality data from 335 Twitter users by asking them to conduct a personality test and examined the relationship between personality and different types of Twitter users (popular users, influential users, listeners and highly read users). They identified those types of Twitter users by using publicly available counts of (what Twitter calls) “following,” “followers,” and “listed,” and by using existing social media ranks. In our work we do not aim to predict users’ personality based on the big five model of personality (since this requires users to complete a personality test first), but aim to predict self-reported personality related characteristics that form a user’s distinctive character on Twitter. This allows us to study different aspects of user’s online personality on a larger sample of Twitter users. Further, we explore a larger range of features which go far beyond the publicly available counts and ranks used in [7] and examine their utility for predicting self-reported personality characteristics.

III. Feature Extraction and Classification

In order to automatically categorize users on dimensions of interest and profession, a discriminative feature set must be extracted from Twitter for each user category. In this section, we describe various features that can capture user attributes and behavior on Twitter. We then show how the best features can be identified and used to build efficient classifiers.

A. Feature Engineering

1) Activity Features: Activity features capture various facets of user activities on Twitter including following, replying, favoriting, retweeting, tweeting, hashtagging and link sharing activities. The intuition behind this set of features is that *users who have similar online activity patterns are more likely to belong to the same category*. For example, people in advertising are more likely to reach out to other users. Celebrities such as musicians are likely to have many followers and follow fewer people.

a) Network-theoretic Features: Network-theoretic features describe user characteristics via their position in an activity network. Since we do not have access to the full social network of all users in our dataset, we construct three directed networks (*reply-*, *mention-*, and *retweet-network*) using information from the tweets of users. We use network-theoretic measures such as *In- and Out-Degree*, *Clustering Coefficient*, *Hub and Authority scores* [9], *Betweenness-, Eigenvector-, and Closeness-Centrality*.

b) Following, Retweeting and Favoriting: Since we do not have access to the full following network of users, we compute simple ratios and counts (*Follower Count, Followee Count, Follower-Followee Ratio, Follower-Tweet Ratio, Favorite-Message Ratio*) which expose how popular a user is and/or how valuable and interesting his content might be for other users.

c) Diversity of Activities: The next set of features capture the diversity in a user’s activity patterns. Our activity diversity features are based on Stirling’s diversity measure [10] which captures three qualities of diversity - *variety*, *balance*, and *similarity*.

Social/Hashtag/Link/Temporal Variety: The social variety of a user is defined as the ratio between the number of different users a user interacted with (U_i) and the total number of messages published by this user (M). A high social variety indicates that a user mainly uses Twitter for a social purpose. The hashtag, link and temporal varieties are defined in the same way as the social variety.

Social/Hashtag/Link/Temporal Balance: To quantify the social balance of a stream, we define an entropy-based measure, which indicates how evenly balanced the social interactions of a user are. If a user’s personal user stream has a high social balance, this indicates that the user interacts almost equally with a large set of users U_i. The hashtag, link and temporal balance are defined in the same way as the social balance as an entropy-based measure which quantifies how focused the hashtagging, the link sharing and the temporal tweeting activities of a user are.

Social/Hashtag/Link/Temporal Similarity: To measure the similarity between two users, we represent each user as a vector of users he interacted with, hashtags he used, links he used and time points he tweeted at. That means that we use the interaction partners, hashtags, links and time points as features and count their frequency.

2) Semantic Features: Next we present a set of features that can be used to characterize users semantically via the content of their messages, or the content of their personal description (bio information). The intuition behind this set of features is that *users who talk about similar things are more likely to belong to the same category*.

Bag of Words: We represent each user by the union of all the published messages, excluding stopwords, and use term frequency-inverse document frequency (*TFIDF*) as the weighting schema. TFIDF allows us to emphasize the words which are most discriminative for a document (where a document in this case is a user).

Latent Topics: Topic modeling approaches discover topics in large collections of documents. The most basic topic modeling algorithm is Latent Dirichlet Allocation (LDA) [11]. In this work we fit an LDA model to a stratified sample of 10% of our training documents where each document consists of all messages authored by one user. We choose the default hyperparameters ($\alpha = 50/T$, $\beta = 0.01$) and optimize them during training by using Wallach's fixed point iteration method [12]. We choose the number of topics $T = 200$ empirically by estimating the log likelihood of a model with T= 50, 100, 150, 200, 250, 300 on held out data.

Explicit Concepts: In [13] the authors evaluated several open APIs for extraction semantic concepts and entities from tweets. They found that the AlchemyAPI extracts the highest number of concepts, and has also the highest entity-concept mapping accuracy. We apply the concept extraction method provided by Alchemy[2] to the aggregation of a sample of users' messages and represent each user as a weighted vector of DBpedia concepts.

Possessives: Besides using topics, concepts and words as semantic features, we also employ words following personal pronouns as features (e.g. "my mac", "my wife", "my girlfriend") since previous work [2] has shown that self-reference pronouns are often useful for distinguishing certain properties of individuals.

3) Social Semantic Features: Beside textual content created by a given user, which can be explicitly attributed to the user, other users may also associate their textual content with the user. For example, Twitter allows users to create user lists. A user may use these lists to organize their contacts. Usually lists consist of a name and a short, optional description of the list. If a user adds another user to a list he/she directly associates the list name and description with this user (*List TFIDF*, *List Concepts*). Further, users can be indirectly associated with topics by extracting the topics the user's top friends are talking about (*Friend Concepts*). We determine the top friends of a user by analyzing how frequently he interacts with other users, since previous research has shown that communication intensity is second to communication intimacy in predicting tie strength [14]. This set of features examines what others are saying about/to particular users, and how that can aid in categorizing users.

4) Linguistic Style Features: The last set of features are designed to characterize users via their use of language. The motivation for this set of features is to consider not what the user is saying but *how he says it*. Given the short length of tweets, it would be interesting to observe if there are significant linguistic cues and how they vary over users of different categories.

We use LIWC [15] to classify words into 70 linguistic dimensions[3] which we used as features. Apart from LIWC we also use a Twitter-specific part-of-speech tagger [16] and compute how frequently a certain tag is used by a user on average. Tags include standard linguistic part of speech tags such as verbs, nouns, proper nouns, adjectives, but also include Twitter or social media specific tags such as emoticons, links, usernames or hashtags. Therefore we computed features such as the mean number of emoticons or the mean number of interjections (e.g., lol, haha, FTW, yea) a user is using. Finally, we assess how easily text (in our case the aggregation of all recent tweets authored by a user) can be read by using standard readability measures such as the *Flesch reading ease score*, the *Fog index*, and the *Flesch-Kincaid grade level score*.

B. Feature Selection

A number of feature selection metrics have been explored in text categorization, among which information gain (IG), chi-square (CHI), correlation coefficient (CC) and odds ratios (OR) are considered most effective. CC and OR are one-sided metrics while IG and CHI are two-sided. In this work we use the IG which measures the difference between the entropy of the class labels and the conditional entropy of the class labels given a feature and the CC which shows the worth of an attribute by measuring the Pearson correlation between it and the class.

C. Classification Models

Ensemble techniques such as Random Forests have an advantage in that they alleviate the small sample size problem and related overfitting problems by incorporating multiple classification models [17]. Random Forests grow a voting committee of decision trees by selecting a random set of $logM + 1$ features where M refers to the total number of features. Therefore, random forests are particularly useful for high-dimensional datasets because increased classification accuracy can be achieved by generating multiple prediction models, each with a different feature subset [18] [19].

However it is known that the performance of Random Forests depends on the correlation between trees as well as the prediction strength of each individual tree. Therefore, we decided to combine Random Forests with a greedy correlation based sub-feature-group selection method [20] which prefers subsets of features that are highly correlated with the class while having low intercorrelation. This ensures that the trees which are grown are strong and uncorrelated.

[2] http://www.alchemyapi.com

[3] http://www.liwc.net/descriptiontable1.php

To assess the performance of different classification models we first conduct a 5-fold cross validation and subsequently conduct a separate evaluation on a hold-out test dataset which consists of a random sample of users. We use the area under the ROC curve (AUC) as an evaluation measure. One advantage of ROC graphs is that they enable comparing classifiers performance without regard to class distributions which makes them very useful when working with unbalanced datasets. To have a realistic setup, we did not artificially balance our dataset and randomly chose three times more negative samples than positive ones for each class.

IV. Experimental Evaluation

In this section, we will first discuss the training and test datasets that we collected for this study. Then, we will describe our classification results in detail followed by a discussion of the implications of this study.

A. Datasets

1) Wefollow Top-500 Dataset: In order to construct our classification models, we need an established “ground truth” or gold standard. For this, we leveraged the manually curated *Wefollow* web directories. When a user wishes to place themselves within the *Wefollow* system, they either send a tweet with a maximum of 5 labels they wish to be associated with or register themselves via the *Wefollow* web application. It is important to note that users choose their own labels thus reflecting their opinion of themselves. While this therefore means that the labels are not guaranteed to reflect the consensus opinion of the user, it does mean that more hidden or subtle labels [21] are recorded. Each *Wefollow* directory corresponds to one label and users within each directory are ranked in a crowdsourced manner.

At the end of July 2012, we crawled the 100 largest *Wefollow* directories for Twitter handles. We then placed those directories into two broad dimensions: profession and personality[4]. For each relevant *Wefollow* directory, we extracted a list of users registered in this directory and their corresponding rank. *Wefollow* was using a proprietary algorithm to rank users at the time the data was crawled. According to their website, the algorithm took into account how many users in a directory follow another user in this directory. In order to ensure equal users for each class, we chose the top 500 users and mapped them to a dataset which was crawled between Sep 2009 and Apr 2011. To ensure that reasonable data was obtained, we excluded all users who had published less than 50 tweets in this time period, and for users with more than 3,000 tweets in our dataset we randomly sampled 3,000 of their tweets. After cleaning, the data amounted to 3,710,494 tweets from 7,121 users over these categories. It should be noted that 92% of the users provided a short bio description in their profile that we also extracted.

2) User Lists: An alternative to the self-assigned tags of *Wefollow* are user lists, which are categorizations users make of others on Twitter, and are public to view. Thus, for each of the 7,121 users in our dataset we crawled their 1,000 most recent list memberships. In our sample, which is obviously biased towards active and popular users, 96% of users were assigned to at least one list, with the median number of lists per user being 75, and the mean was 232.

Though the majority of user lists correspond to topical labels (e.g., “computer science” or “healthcare”), user lists may also describe how people feel about the list members (e.g., “great people”, “geeks”, “interesting twitterers”) and how they relate with them (e.g., “my family”, “colleagues”, “close friends”) [22]. Also, since user lists are created by the crowd they may be noisy, sparse or inappropriate.

3) Random Test Dataset: In addition to the *Wefollow* Top-500 dataset, which is biased towards users with high *Wefollow* rank, we crawled another random sample of *Wefollow* users which were not part of our original dataset. This, in theory, provides a broader base of users to sample from when testing our models (increasing generalizability), although it must be noted that, since these users were not highly ranked on *Wefollow* there is obviously a question regarding the reliability of their self-tags. This sample was collected by tracking new registrations made in April 2013 (to one of the above listed *Wefollow* directories). From this collection, we selected 100 random users from each directory.

B. Results

We trained multiple binary random forest classifiers for each category with different feature groups using the greedy correlation-based feature-group selection method [20]. The following section reports our results from the personality and profession classification tasks using cross fold validation and a separate test dataset of random users.

1) **Personality-related Categories:**

a) **WeFollow Top-500 Dataset:** Figure 1(a) shows that for all personality-related categories the best performance can be achieved when using a combination of all features. This provides an AUC score consistently ≥ 0.9 for 6 categories out of 8.

The highest performing individual feature group is the social-semantic group. These features achieve the highest AUC values for most categories (*advertising*, *creative*, *ecological* and *informational*). A separate performance comparison of the three different feature types of the social-semantic feature group (TFIDF based on user list memberships, concepts extracted from user list memberships and concepts extracted from the tweets of a users' top friends) shows that TFIDF based on user lists performs best ($AUC > 0.8$ for all categories except *inspirational* and *innovational*). This suggests that information about user list memberships is indeed useful for predicting personality-related user attributes. Also Table I which reveals the top five features for each category ranked via

[4] Not all the directories fit neatly into those dimensions. Directories that did not fit were excluded. The Wefollow directories *musician*, *dj*, *songwriter*, *singer* were merged into the category *musician*, the *developer*, *webdeveloper* and *computers* directory were merged into the category *IT*, the directories *business* and *entrepreneur* were merged into the category *business* and finally the directories *advertising* and *marketing* were merged into the category *marketing*. Each other *Wefollow* directory maps to exactly one category.

their Pearson correlation with the category label, shows that TFIDF weighted list names tend to be amongst the top features for all categories. For example *informational* users tend to be members of lists like *newspapers*, *outlets*, *newspapers*, *breaking* and *reporters*, while *ecological* users tend to be in lists called *eco*, *environmental* or *sustainable*.

Social-semantic features are closely followed by linguistic style features which achieve the highest AUC values for the category *funny* and *inspirational*. Ranking only features of the social-semantic feature group shows that *funny* users tend to use more swear words ($CC = 0.47$), body related words ($CC = 0.37$), negative emotions ($CC = 0.35$) and talk about themselves ($CC = 0.35$). On the other hand *inspirational* users have a high usage of the word "you" ($CC = 0.24$) and talk about positive emotions ($CC = 0.22$) and affective processes ($CC = 0.2$) – i.e., they use words which describe affection such as cried or abandon.

For the category *religious* semantic features achieve a slightly higher AUC value than linguistic style and social-semantic features. Ranking the features of the semantic features group reveals that religious (or more specific christian) Twitter users tend to talk about their religion and therefore use words such as worship ($CC = 0.44$), bible ($CC = 0.41$), church ($CC = 0.41$), praying ($CC = 0.39$) or god ($CC = 0.38$). Further Table I shows that *religious* users tend to use words which fall into the LIWC category of religious words ($CC = 0.48$) and tend to be mentioned in lists called *christian* or *church*. This indicates also that social-semantic and linguistic style features contribute to identifying religious users and Figure 1(a) shows that indeed *religious* users can be identified best when using an optimal combination of all features.

Activity features which describe users via their activity patterns tend to perform worse than social-semantic, semantic and linguistic style features for most categories except *advertising* and *informational*. Ranking features of the activity feature group by their information gain and correlation coefficient shows that users who are actively advertising tend to have a significantly higher temporal balance ($CC = 0.35$ and $IG = 0.09$) and temporal variance ($CC = 0.24$ and $IG = 0.08$) which indicates that they publish frequently and the same amount of messages. Informational users tend to have higher informational variety ($CC = 0.38$ and $IG = 0.09$) and informational balance ($CC = 0.2$ and $IG = 0.08$) which indicates that they tend to share a large variety of links but share each link only once – i.e. they provide more information.

Overall, the most difficult task was to classify *creative* users. One can see from Table I and Figure 1 that the best features for this category are social semantic features while all other feature groups perform pretty poor.

b) **Random Test Dataset:** Figure 1(b) shows that for the random test users, the overall accuracy is slightly lower than for the top-500 dataset. This can be attributed to the fact that these users are not the top-ranked users, and therefore there is question of reliability on their self-categorization. However, we observe that the AUC scores are still reasonably good over most of the categories (≥ 0.8) except for the category *creative*. The results on test users shows that social-semantic features and linguistic style features tend to be most useful for classifying random test users into personality categories. Again, activity features are only useful for the category *advertising* and *informational*. Interestingly, semantic features do not generalize well to random test users and are almost as useless as activity features. One possible explanation is that random test users may be less active on Twitter and may reveal less personal information when tweeting. Another possible explanation is that there might be a vocabulary mismatch between the train and test users which might become bigger if we reduce the feature space during training.

2) **Professional Areas:**

a) **WeFollow Top-500 Dataset:** One can see from Figure 2(a) that again using a optimal subset of all features provides excellent classification performance with AUC values ≥ 0.9 for all categories except *business*. The most useful feature groups for classifying users into professional areas are linguistic style and semantic features. It is interesting to note that social-semantic features which were most useful for identifying users' personality related attributes, are not as useful for identifying their professional areas. Particularly for identifying users interested in *business* or *health* and for identifying *politicians* and *writers* social-semantic features are pretty useless ($AUC < 0.6$).

Table II shows that indeed the features with the highest information gain tend to be semantic and linguistic style features. In the semantic feature group especially topical features and TFIDF weighted words were most useful which indicates that users working in different professional areas talk indeed about topics related to this area. For example, *photographers* tend to talk about photography and art and design in general, while *politicians* tend to talk about Obama, the republican party and health care.

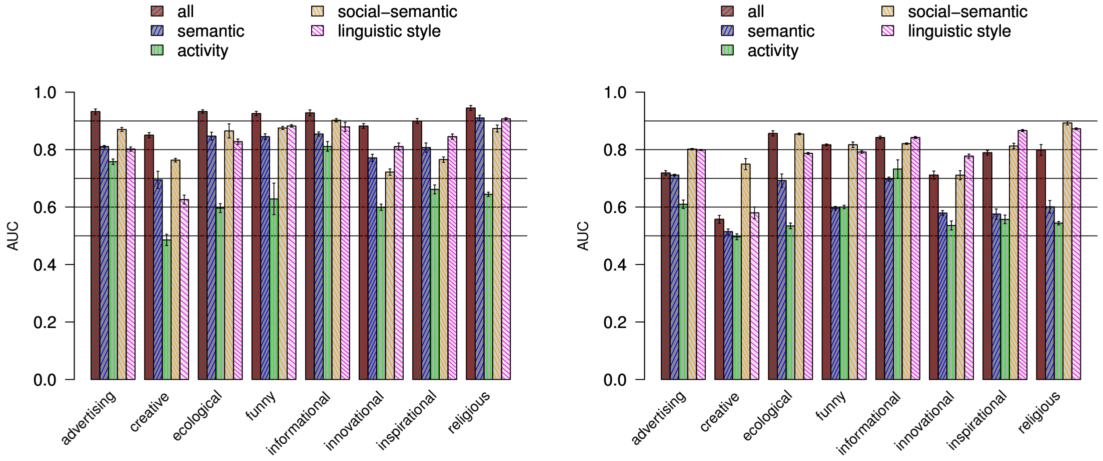
When comparing different types of semantic features we found for both tasks that concepts were pretty useless. One possible justification for this could be that the concept annotations tend to be too general. On the other hand, TFIDF weighted words work very well overall and TFIDF outperforms LDA on the random hold out dataset. This finding is in line with previous research [4] which shows as well that TF-IDF features perform almost twice as good as topic features within a user classification task.

In the linguistic style feature group LIWC features were most useful. One can see from Table II that for example *photographers* tend to focus on the perceptual process of seeing (i.e., they use words like *seeing* or *viewing*) while *musicians* focus on the perceptual process of hearing (i.e., they use words like *hearing* or *listening*). Not surprising, people working in the *health* sector tend to use words related with biological processes such as health-, body and ingestion-related words.

Finally, our results suggest that for some professional areas such as the movie industry, social activity features add value, since users who interact with key-players in their domain (such

TABLE I: Top features ranked by their Pearson correlation coefficient with each category. Topics are represented via their three most likely words. Feature Group: 🟧 social-semantic, 🟦 semantic, 🟪 linguistic-style

advertising	creative	ecological	funny	informational	innovational	inspirational	religious
tfidf_list: advertising (0.51)	tfidf_list: twibes-creative (0.3)	green, energy, climate (0.58)	liwc: swear (0.47)	tfidf_list: newspapers (0.59)	tfidf_list: twibes-innovation (0.46)	love, I, we (0.47)	tfidf_list: christian(0.53)
tfidf_list: ad (0.44)	tfidf_list: com/creative/ twitter-list (0.3)	tfidf_list: eco (0.51)	was, he, is (0.47)	tfidf_list: outlets	tfidf_list: com/innovation/ twitter-list (0.46)	tfidf_list: twibes-inspiration (0.45)	liwc: relig (0.48)
tfidf_list:marketers (0.44)	tfidf_list: twibes-creative (0.23)	tfidf_list: environmental (0.5)	shit, fuck, ass (0.44)	tfidf_list: newsnews (0.58)	business, research, model (0.41)	tfidf_list: com/inspiration/ twitter-list (0.45)	tfidf_list:church (0.47)
tfidf_list: marketing (0.42)	tfidf_list: outofbox (0.22)	tfidf_list: sustainable (0.5)	I, me, you(0.42)	tfidf_list: breaking (0.56)	tfidf_tweetbio: innovation (0.37)	you, is, it (0.44)	god, lord, jesus (0.47)
tfidf_tweetbio: marketing (0.4)	tfidf_list: ly/oaomgp (0.22)	tfidf_list: environment (0.5)	liwc: body (0.37)	tfidf_list: reporters (0.55)	tfidf_list: twibes-innovation (0.32)	tfidf_list: spirituality (0.28)	tfidf_list: christians (0.45)



(a) *Wefollow* Top-500 Dataset

(b) Random Test Dataset

Fig. 1: Binary classifiers for different personality related user attributes.

as the film critic writer Scott Weinberg and Peter Sciretta) are more likely to work in this industry.

b) **Random Test Dataset:** Once again the test dataset reduces accuracy but the values are still reasonably good considering the unreliability of the ground truth in this case (see Figure 2(b)).

Our results from the random test dataset show that linguistic style features work best, which means that they generalize very well compared to other feature groups. An exception of this general pattern is the category *photographer*. For this category linguistic style features are not very useful and social-semantic features work best.

Again we observe that semantic features perform well within the cross fold validation but do not generalize well to the random test users. Overall, *writers* were most difficult to classify and the best performance could be achieved for the category *health-related professions* and *musicians* ($AUC > 0.9$) by using linguistic style features.

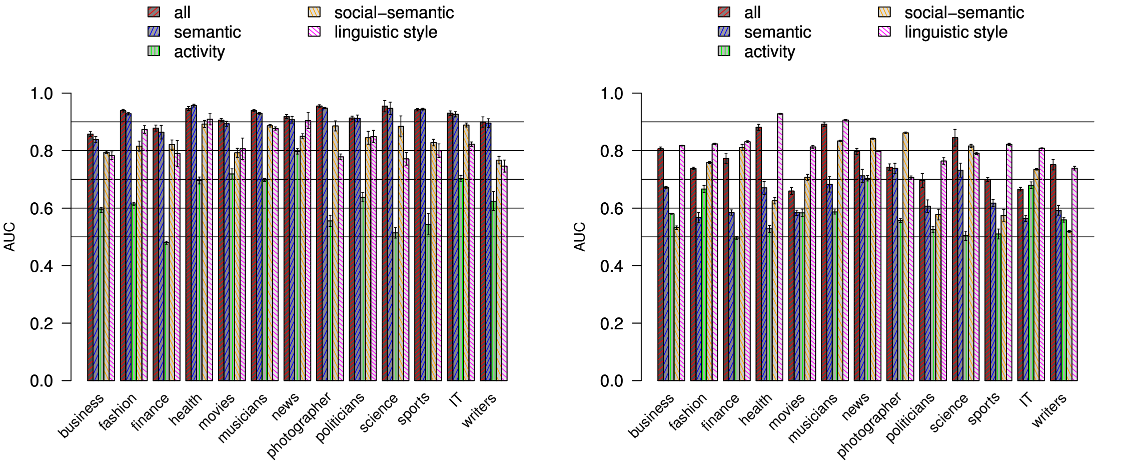
C. Discussion of Results

Our results show that random forests built on an optimal subset of the features demonstrate an impressive accuracy of above 0.8 (random test users) and 0.9 (top-500 *Wefollow* users), for most categories. For both tasks (personality and professional area classification) our results suggest that linguistic style features are most useful since they tend to generalize well on random test users. Further, the feature analysis reveals that LIWC based personal concern, linguistic and psychological process features, as well as our Twitter specific style and readability features, are useful for identifying users' professional areas and personality related attributes. This suggests that *not only what a user says but also how he expresses himself on Twitter* may reveal useful information about his professional areas and personality related attributes.

Further, we found that social-semantic features are very useful for predicting personality related attributes but less useful for predicting professional areas (especially for business, health, politics and writer, where the $AUC < 0.6$ when trained with social-semantic features). Since the best social-semantic

TABLE II: Top features ranked by their Information gain for each professional area. Topics are represented via their three most likely words. The cell color indicates to which group (🟦 semantic, 🟪 linguistic-style, 🟩 activity) a feature belongs.

business	fashion	finance	health	movies	music	news	photogr.	politician	science	sports	IT	writers
startup, startups, entrepreneurs (0.1)	dress, ebay, date (0.12)	dollar, #forex, u.s. (0.08)	liwc: health (0.1)	movie, trailer, review (0.11)	music, new, album (0.21)	u.s., obama, news (0.1)	photo, photography, photos (0.22)	obama, gop, palin (0.14)	science, research, data (0.12)	game, team, win (0.1)	code, web, project (0.24)	book, books, writing (0.08)
great, what, looking (0.09)	fashion, internship, intern (0.08)	liwc: money (0.06)	health, may, healthy (0.09)	video, movie, film (0.07)	liwc: hear (0.14)	death, two, men (0.09)	art, design, work (0.03)	obama, health, care (0.12)	space, nasa, science (0.06)	jets, jacksonville, nfl (0.06)	iphone, ipad, app (0.15)	book, ever, idea (0.05)
social, facebook, app (0.08)	so, love, oh (0.05)	$$, long, short (0.05)	liwc: bio (0.08)	mention slashfilm (0.03)	mix, remix, dj (0.12)	new, more, has (0.07)	liwc: see (0.03)	u.s., obama, news(0.07)	green, energy, climate (0.05)	bulls, lakers, nba (0.05)	new, just, site (0.11)	not, this, why (0.04)
twitter, social, media (0.06)	free, win, sale (0.05)	today, nice, $aapl (0.03)	health, patients, medical (0.07)	RT slashfilm (0.03)	liwc: work (0.1)	liwc: I (0.07)	new, more, has (0.02)	#tcot, #teaparty, #gop (0.06)	book, ever, idea (0.03)	yankees, baseball, mets (0.04)	google, twitter, apple (0.11)	also, actually, thing (0.03)
business, research, model (0.05)	show, girl, says (0.04)	business, apple, stock (0.03)	media, health, today (0.02)	RT scottEweinberg (0.03)	got, is, me (0.1)	liwc: assent (0.06)	rain, weather, snow (0.02)	new, more, has (0.05)	health, patients, medical (0.03)	world, cup, 2010 (0.03)	john, david, review (0.06)	#amwriting, las vegas, writing (0.03)



(a) *Wefollow* Top-500 Dataset

(b) Random Test Dataset

Fig. 2: Binary classifiers for different professional areas.

feature are TFIDF weighted list names, we can conclude that users' list memberships may indeed reveal information about users' personality related attributes (at least for those which were explored within this work). However, for professional areas the utility of social-semantic features may depend on the professional area.

Interestingly, semantic features perform well within the cross fold validation but do not generalize well to the random test users. One possible explanation is that there might be a vocabulary mismatch between the train and test users which is likely to become bigger if we reduce the feature space during training. Another potential explanation is that test users tend to be less active on Twitter and may reveal less personal information when tweeting. When comparing different types of semantic features we found that concepts did not provide much value. One possible justification could be that concept annotations tend to be too general. However, TFIDF weighted words work very well overall and TFIDF outperforms LDA on the random hold out dataset. This finding is in line with previous research [4] which shows as well that TFIDF features perform almost twice as good as topic features within a user classification task.

Consistently, we found that activity features are rather useless for most categories except those where users show very specific activity patterns (e.g., the category of informational users who tend to post much more links than others). This finding is inline with previous research [2] [3] which found that user-centric content features are more useful than features which capture structural relations and similarities between users.

One finding that was not inline with existing work was the utility of self-referring possessives (i.e., my followed by any word). Unlike [2], performance did not improve when self-referring possessives were added. It is important to note that the classification dimensions described in [2] are very different from those which we use in our work. For example, it is intuitive that self-referring possessives are useful for predicting the gender of a user since a user who talks e.g. about his wife (i.e., uses the bigram "my wife") is almost certainly male. For professional areas and personality related attributes we could not find self-referring possessives with similar predictive power.

One limitation of our work is that both datasets used consist of users who registered themselves at *Wefollow* and those users may not be representative for the Twitter population as a whole. Thus, as future work, we propose an in-depth investigation into the relationship and model performance between those users that explicitly promote themselves via services like *Wefollow* and those that do not use such services

V. Conclusions

In this work we have constructed a comprehensive collection of features (around 20k features) and examined their efficacy in classifying Twitter users according to two broad different dimensions: professions and personality. We showed that the large set of features can be pruned to around 100 features per category using a greedy correlation-based subset feature selection. Further, random forests built on the selected subset of features obtained an impressive accuracy of ≥ 0.9 for most categories using our top-500 *Wefollow* dataset, and an accuracy of around ≥ 0.8 for most categories using our random test user dataset. Based on the varying utility of the features across categories, we believe that in order to create new classifications, a large initial set of features is required that can be pruned based on the characteristics of each category. This ensures that the idiosyncrasies of different categories are captured well by the features. Overall, we observed in both tasks that using only linguistic style features lead to consistently good results.

References

[1] R. Holt. (2013) Half a billion people sign up for twitter. [Online]. Available: http://www.telegraph.co.uk/technology/9837525/Half-a-billion-people-sign-up-for-Twitter.html

[2] D. Rao, D. Yarowsky, A. Shreevats, and M. Gupta, "Classifying latent user attributes in twitter," in *Proceedings of the 2nd international workshop on Search and mining user-generated contents*, ser. SMUC '10. New York, NY, USA: ACM, 2010, pp. 37–44. [Online]. Available: http://doi.acm.org/10.1145/1871985.1871993

[3] M. Pennacchiotti and A.-M. Popescu, "Democrats, republicans and starbucks afficionados: user classification in twitter," in *Proceedings of the 17th ACM SIGKDD international conference on Knowledge discovery and data mining*, ser. KDD '11. New York, NY, USA: ACM, 2011, pp. 430–438. [Online]. Available: http://doi.acm.org/10.1145/2020408.2020477

[4] L. Hong and B. D. Davison, "Empirical study of topic modeling in twitter," in *Proceedings of the IGKDD Workshop on Social Media Analytics (SOMA),*, 2010.

[5] J. Bryden, S. Funk, and V. A. A. Jansen, "Word usage mirrors community structure in the online social network twitter," *EPJ Data Science*, vol. 2, no. 1, pp. 3+, 2013. [Online]. Available: http://dx.doi.org/10.1140/epjds15

[6] J. Golbeck, C. Robles, M. Edmondson, and K. Turner, "Predicting personality from twitter," in *SocialCom*. IEEE, 2011, pp. 149–156. [Online]. Available: http://dblp.uni-trier.de/db/conf/socialcom/socialcom2011.html#GolbeckRET11

[7] D. Quercia, M. Kosinski, D. Stillwell, and J. Crowcroft, "Our twitter profiles, our selves: predicting personality with twitter," in *SocialCom*, 2011.

[8] D. J. Hughes, M. Rowe, M. Batey, and A. Lee, "A tale of two sites: Twitter vs. facebook and the personality predictors of social media usage," *Comput. Hum. Behav.*, vol. 28, no. 2, pp. 561–569, Mar. 2012. [Online]. Available: http://dx.doi.org/10.1016/j.chb.2011.11.001

[9] J. M. Kleinberg, "Authoritative sources in a hyperlinked environment." in *SODA*, H. J. Karloff, Ed. ACM/SIAM, 1998, pp. 668–677. [Online]. Available: http://dblp.uni-trier.de/db/conf/soda/soda98.html#Kleinberg98

[10] A. Stirling, "A general framework for analysing diversity in science, technology and society." *Journal of the Royal Society*, vol. 4, no. 15, pp. 707–19, Aug 2007. [Online]. Available: http://rsif.royalsocietypublishing.org/cgi/content/abstract/4/15/707

[11] D. M. Blei, A. Ng, and M. Jordan, "Latent dirichlet allocation," *JMLR*, vol. 3, pp. 993–1022, 2003.

[12] H. M. Wallach, "Structured topic models for language," Ph.D. dissertation, University of Cambridge, 2008.

[13] H. Saif, Y. He, and H. Alani, "Semantic sentiment analysis of twitter." Boston, UA: Springer, 2012, pp. 508–524. [Online]. Available: http://ceur-ws.org/Vol-838/paper_01.pdf

[14] E. Gilbert and K. Karahalios, "Predicting tie strength with social media," in *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, ser. CHI '09. New York, NY, USA: ACM, 2009, pp. 211–220. [Online]. Available: http://doi.acm.org/10.1145/1518701.1518736

[15] Y. R. Tausczik and J. W. Pennebaker, "The psychological meaning of words: Liwc and computerized text analysis methods," 2010. [Online]. Available: http://homepage.psy.utexas.edu/homepage/students/Tausczik/Yla/index.html

[16] K. Gimpel, N. Schneider, B. O'Connor, D. Das, D. Mills, J. Eisenstein, M. Heilman, D. Yogatama, J. Flanigan, and N. A. Smith, "Part-of-speech tagging for twitter: annotation, features, and experiments," in *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies: short papers - Volume 2*, ser. HLT '11. Stroudsburg, PA, USA: Association for Computational Linguistics, 2011, pp. 42–47. [Online]. Available: http://dl.acm.org/citation.cfm?id=2002736.2002747

[17] T. G. Dietterich and D. Fisher, "An experimental comparison of three methods for constructing ensembles of decision trees," in *Bagging, boosting, and randomization. Machine Learning*, 2000, pp. 139–157.

[18] B. B. Z. Pengyi Yang, Yee Hwa Yang and A. Y. Zomaya, "A review of ensemble methods in bioinformatics," *Current Bioinformatics*, vol. 5, pp. 296–308, 2010.

[19] L. Breiman, "Random forests," in *Machine Learning*, 2001, pp. 5–32.

[20] M. A. Hall, "Correlation-based feature selection for machine learning," Tech. Rep., 1998.

[21] M. Kosinski, D. Stillwell, and T. Graepel, "Private traits and attributes are predictable from digital records of human behavior," 2013. [Online]. Available: http://www.pnas.org/cgi/doi/10.1073/pnas.1218772110

[22] C. Wagner, V. Liao, P. Pirolli, L. Nelson, and M. Strohmaier, "It's not in their tweets: Modeling topical expertise of twitter users," in *Proceedings ASE/IEEE International Conference on Social Computing (SocialCom2012)*, 2012.

Meaning as Collective Use: Predicting Semantic Hashtag Categories on Twitter

Lisa Posch
Graz University of Technology,
Knowledge Management
Institute
Inffeldgasse 13, 8010 Graz,
Austria
lposch@sbox.tugraz.at

Claudia Wagner
JOANNEUM RESEARCH,
Institute for Information and
Communication Technologies
Steyrergasse 17, 8010 Graz,
Austria
clauwa@sbox.tugraz.at

Philipp Singer
Graz University of Technology,
Knowledge Management
Institute
Inffeldgasse 13, 8010 Graz,
Austria
philipp.singer@tugraz.at

Markus Strohmaier
Graz University of Technology,
Knowledge Management
Institute
Inffeldgasse 13, 8010 Graz,
Austria
markus.strohmaier@tugraz.at

ABSTRACT

This paper sets out to explore whether data about the usage of hashtags on Twitter contains information about their semantics. Towards that end, we perform initial statistical hypothesis tests to quantify the association between usage patterns and semantics of hashtags. To assess the utility of pragmatic features – which describe how a hashtag is used over time – for semantic analysis of hashtags, we conduct various hashtag stream classification experiments and compare their utility with the utility of lexical features. Our results indicate that pragmatic features indeed contain valuable information for classifying hashtags into semantic categories. Although pragmatic features do not outperform lexical features in our experiments, we argue that pragmatic features are important and relevant for settings in which textual information might be sparse or absent (e.g., in social video streams).

Categories and Subject Descriptors

H.3.5 [**Information Storage and Retrieval**]: Online Information Services—*Web-based services*

Keywords

Twitter; hashtags; social structure; semantics

1. INTRODUCTION

A hashtag is a string of characters preceded by the hash (#) character and it is used on platforms like Twitter as descriptive label or to build communities around particular topics [15]. To outside observers, the meaning of hashtags is usually difficult to analyze, as they consist of short, often abbreviated or concatenated concepts (e.g., #MSM2013). Thus, new methods and techniques for analyzing the semantics of hashtags are definitely needed.

A simplistic view on Wittgenstein's work [17] suggests that *meaning is use*. This indicates that the meaning of a word is not defined by a reference to the object it denotes, but by the variety of uses to which the word is put. Therefore, one can use the narrow, lexical context of a word (i.e., its co-occurring words) to approximate its meaning. Our work builds on this observation, but focuses on the pragmatics of a word (i.e., how a word, or in our case a hashtag, is used by a large group of users) – rather than its narrow, lexical context.

The aim of this work is to investigate to what extent pragmatic characteristics of a hashtag (which capture how a large group of users uses a hashtag) may reveal information about its semantics. Specifically, our work addresses the following research questions:

- Do different semantic categories of hashtags reveal substantially different usage patterns?
- To what extent do pragmatic and lexical properties of hashtags help to predict the semantic category of a hashtag?

To address these research questions we conducted an empirical study on a broad range of diverse hashtag streams belonging to eight different semantic categories (such as *technology*, *sports* or *idioms*) which have been identified in previous research [12] and have shown to be useful for grouping hashtags. From each of the eight categories, we selected ten sample hashtags at random and collected temporal snapshots of messages containing at least one of these hashtags at three different points in time. To quantify how hashtags are used over time, we extended the set of pragmatic stream measures which we introduced in our previous work [16] and applied them to the hashtag streams in our dataset. These pragmatic measures capture not only the social structure of a hashtag at specific points in time, but also the changes in social structure over time.

WWW 2013 Companion, May 13–17, 2013, Rio de Janeiro, Brazil.
ACM 978-1-4503-2038-2/13/05.

To answer the first research question, we used statistical standard tests which allow to quantify the association between pragmatic characteristics of hashtag streams and their semantic categories. To tackle the second research question, we firstly computed lexical features using a a standard bag-of-words model with term frequency (TF). Then, we trained several classification models with lexical features only, pragmatic features only and a combination of both. We compared the performance of different classification models by using standard evaluation measures such as the F1-score (which is defined as the harmonic mean of precision and recall). To get a fair baseline for our classification models, we constructed a control dataset by randomly shuffling the category labels of the hashtag streams. That means we destroyed the original relationship between the pragmatic properties and the semantic categories of hashtags.

Our results show that pragmatic features indeed reveal information about hashtags' semantics and perform significantly better than the baseline. They can therefore be useful for the task of semantically annotating social media content. Not surprisingly, our results also show that lexical features are more suitable than pragmatic features for the task of semantically categorizing hashtag streams. However, an advantage of pragmatic features is that they are language- and text-independent. Pragmatic features can be applied to tasks where the creation of lexical features is not possible – such as multimedia streams. Also for scenarios where textual content is available, pragmatic features allow for more flexibility due to their independence of the language used in the corpus. Our results are relevant for social media and semantic web researchers who are interested in analyzing the semantics of hashtags in textual or non-textual social streams (e.g., social video streams).

This paper is structured as follows: Section 2 gives an overview of related research on analyzing the semantics of tags in social bookmarking systems and research on hashtagging on Twitter in general. In Section 3 we describe our experimental setup, including our datasets, feature engineering and evaluation approach. Our results are reported in Section 4 and further discussed in Section 5. Finally, we conclude our work in Section 6.

2. RELATED WORK

In the past, a considerable effort has been spent on studying the semantics of tags (e.g., tags in social bookmarking systems), but also hashtags in Twitter have received attention from the research community.

Semantics of tags: On the one hand, researchers explored to what extent semantics emerge from folksonomies by investigating different algorithms for extracting tag networks and hierarchies from such systems (see e.g., [1], [3] or [13]). The work of [14] evaluated three state-of-the-art folksonomy induction algorithms in the context of five social tagging systems. Their results show that those algorithms specifically developed to capture intuitions of social tagging systems outperform traditional hierarchical clustering techniques. Körner et al. [5] investigated how tagging usage patterns influence the quality of the emergent semantics. They found that 'verbose' taggers (*describers*) are more useful for the emergence of tag semantics than users who use a small set of tags (*categorizers*).

On the other hand, researchers investigated to what extent tags (and the resources they annotate) can be semantically grounded and classified into predefined semantic categories. For example, Noll and Meinel [8] presented a study of the characteristics of tags and determined their usefulness for web page classification [9]. Similar to our work, Overell et al. [10] presented an approach which allows classifying tags into semantic categories. They trained a classifier to classify Wikipedia articles into semantic categories, mapped Flickr tags to Wikipedia articles using anchor texts in Wikipedia and finally classified Flickr tags into semantic categories by using the previously trained classifier. Their results show that their ClassTag system increases the coverage of the vocabulary by 115% compared to a simple WordNet approach which classifies Flickr tags by mapping them to WordNet via string matching techniques. Unlike our work, they did not take into account how tags are used, but learn relations between tags and semantic categories via mapping them to Wikipedia articles.

Pragmatics and semantics of hashtags: On Twitter, users have developed a tagging culture by adding a hash symbol (#) in front of a short keyword. The first introduction of the usage of hashtags was provided by Chris Messina in a blog post [7]. Huang et al. [4] state that this kind of new tagging culture has created a completely new phenomenon, called *micro-meme*. The difference between such micro-memes and other social tagging systems is that the participation in micro-memes is an *a-priori* approach, while other social tagging systems follow an *a-posteriori* approach. This is due to the fact that users are influenced by the observation of the usage of micro-meme hashtags adopted by other users. The work of [4] suggests that hashtagging in Twitter is more commonly used to join public discussions than to organize content for future retrieval. The role of hashtags has also been investigated in [18]. Their study confirms that a hashtag serves both as a tag of content and a symbol of community membership. Laniado and Mika [6] explored to what extent hashtags can be used as strong identifiers like URIs are used in the Semantic Web. They measured the quality of hashtags as identifiers for the Semantic Web, defining several metrics to characterize hashtag usage on the dimensions of frequency, specificity, consistency, and stability over time. Their results indicate that the lexical usage of hashtags can indeed be used to identify hashtags which have the desirable properties of strong identifiers. Unlike our work, their work focuses on lexical usage patterns and measures to what extent those patterns contribute to the differentiation between strong and weak semantic identifiers (binary classification) while we use usage patterns to classify hashtags into semantic categories.

Recently, researchers have also started to explore the diffusion dynamics of hashtags - i.e., how hashtags spread in online communities. For example the work of [15] aims to predict the exposure of a hashtag in a given time frame while [12] are interested in the temporal spreading patterns of hashtags.

3. EXPERIMENTAL SETUP

Our experiments are designed to explore to what extent pragmatic properties of hashtag streams can be used to gauge the semantic category of a hashtag. We are not only interested in the idiosyncrasies of hashtag usage within one semantic category but also in the deltas between different semantic categories. In this section, we first introduce our dataset as well as the pragmatic and lexical measures which we used to describe hashtag streams. Then we present the methodology and evaluation approach which we used to answer our research questions.

3.1 Dataset

In this work we use data that we acquired from Twitter's API. Romero et al. [12] conducted a user study and a classification experiment and identified eight broad semantic categories of hashtags: *celebrity*, *games*, *idiom*, *movies/TV*, *music*, *political*, *sports* and *technology*. We used a list consisting of the 500 hashtags which were used by most users within their dataset and which were manually assigned to the eight categories as a starting point for creating our own dataset.

For each category, we chose ten hashtags at random (see Table 1). We biased our random sample towards active hashtag streams by re-sampling hashtags for which we found less than 1000 posts at the beginning of our data collection (March 4th, 2012). For those categories for which we could not find ten hashtags that had more than 1000 posts (i.e., *games* and *celebrity*), we selected the most active hashtags per category (i.e., the hashtags for which we found the most posts).

The dataset consists of three parts, each part representing a time frame of four weeks. The different time frames ensure that we can observe the usage of a hashtag over a given period of time. The time frames are independent of each other, i.e., the data collected at one time frame does not contain any information of the data collected at another time frame.

At the start of each time frame, we retrieved the most recent tweets in English for each hashtag using Twitter's public search API. Afterwards, we retrieved the followers and followees of each user who had authored at least one message in our hashtag stream dataset. Some pragmatic features capture information about who potentially consumes a hashtag stream (*followers*) or who potentially informs authors of a hashtag stream (*followees*) and therefore require the one-hop neighborhood of hashtag streams' authors. In this work, we call users who hold both of these roles (i.e., have established a bidirectional link with an author) *friends*. The starting dates of the time frames were March 4th (t_0), April 1st (t_1) and April 29th, 2012 (t_2). Table 2 depicts the number of tweets and relations between users that we collected during each time frame.

The stream tweets were retrieved on the first day of each time frame, fetching tweets that were authored a maximum of seven days previous to the date of retrieval. During the first week of each time frame, the user IDs of the followers and followees were collected. Figure 1 depicts this process.

Since we were interested in learning what types of characteristics are useful for describing a semantic hashtag category, we removed hashtag streams that belong to multiple categories (concretely, we removed the two hashtags #bsb and #mj). We also decided to remove inactive hashtag streams (those where less than 300 posts where retrieved) as estimating information theoretic measures is problematic if only few observations are available [11]. The most common solution is to restrict the measurements to situations where one has an adequate amount of data. We found four inactive hashtags in the category *games* and seven in the category *celebrity*. The removal of these hashtag streams resulted in the complete removal of the category *celebrity* as it was only left with one hashtag stream (#michaeljackson). A possible explanation for the low number of tweets in the hashtag streams for this category is that topics related to celebrities have a shorter life-span than topics related to other categories. Our final datasets consist of 64 hashtag streams and seven semantic categories which were sufficiently active during our observation period.

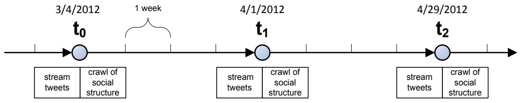


Figure 1: Timeline of the data collection process

Table 2: Description of the complete dataset

	t_0	t_1	t_2
Tweets	94,634	94,984	95,105
Authors	53,593	54,099	53,750
Followers	56,685,755	58,822,119	66,450,378
Followees	34,025,961	34,263,129	37,674,363
Friends	21,696,134	21,914,947	24,449,705
Mean Followers per Author	1,057.71	1,087.31	1,236.29
Mean Followees per Author	634.90	633.34	700.92
Mean Friends per Author	404.83	405.09	454.88

3.2 Feature Engineering

In the following, we define the pragmatic and lexical features which we designed to capture the different social and message based structures of hashtag streams. For our pragmatic features we further differentiate between static pragmatic features (which capture the social structure of a hashtag at a specific point in time) and dynamic pragmatic features (which combine information from several time points).

3.2.1 Static Pragmatic Measures:

Entropy Measures are used to measure the randomness of streams' authors and their followers, followees and friends. For each hashtag stream, we rank the authors by the number of messages they published in that stream (norm_entropy_author) and we rank the followers (norm_entropy_follower), followees (norm_entropy_followee) and

Table 1: Randomly selected hashtags per category (ordered alphabetically)

technology	idioms	sports	political	games	music	celebrity	movies
blackberry	factaboutme	f1	climate	e3	bsb	ashleytisdale	avatar
ebay	followfriday	football	gaza	games	eurovision	brazilmissesdemi	bbcqt
facebook	dontyouhate	golf	healthcare	gaming	lastfm	bsb	bones
flickr	iloveitwhen	nascar	iran	mafiawars	listeningto	michaeljackson	chuck
google	iwish	nba	mmot	mobsterworld	mj	mj	glee
iphone	nevertrust	nhl	noh8	mw2	music	niley	glennbeck
microsoft	omgfacts	redsox	obama	ps3	musicmonday	regis	movies
photoshop	oneofmyfollowers	soccer	politics	spymaster	nowplaying	teamtaylor	supernatural
socialmedia	rememberwhen	sports	teaparty	uncharted2	paramore	tilatequila	tv
twitter	wheniwaslittle	yankees	tehran	wow	snsd	weloveyoumiley	xfactor

friends (norm_entropy_friend) by the number of stream's authors they are related with. A high *author entropy* indicates that the stream is created in a democratic way since all authors contribute equally much. A high *follower entropy* and *friend entropy* indicate that the followers and friends do not focus their attention towards few authors but distribute it equally across all authors. A high *followee entropy* and *friend entropy* indicate that the authors do not focus their attention on a selected part of their audience.

Overlap Measures describe the overlap between the authors and the followers (overlap_authorfollower), followees (overlap_authorfollowee) or friends (overlap_authorfriend) of a hashtag stream. If overlap is *one*, all authors of a stream are also followers, followees or friends of stream authors. This indicates that the stream is consumed and produced by the same users. A high overlap suggests that the community around the hashtag is rather closed, while a low overlap indicates that the community is more open and that active and passive part of the community do not extensively overlap.

Coverage Measures characterize a hashtag stream via the nature of its messages. We introduce four coverage measures. The *informational coverage* measure (informational) indicates how many messages of a stream have an informational purpose - i.e., contain a link. The *conversational coverage* (conversational) measures the mean number of messages of a stream that have a conversational purpose - i.e., those messages that are directed to one or several specific users (e.g., through @replies). The *retweet coverage* (retweet) measures the percentage of messages which are retweets. The *hashtag coverage* (hashtag) measures the mean number of hashtags per message in a stream.

3.2.2 Dynamic Pragmatic Measures:

To explore how the social structure of a hashtag stream changes over time, we measure the distance between the tweet-frequency distributions of authors at different time points, and the author-frequency distributions of followers, followees or friends at different time points. The intuition behind these features is that certain semantic categories of hashtags may have a fast changing social structure since new people start and stop using those types of hashtags frequently, while other semantic categories may have a more stable community around them which changes less over time.

We use a symmetric variation of the *Kullback-Leibler divergence (D_{KL})* which represents a natural distance measure between two probability distributions (A and B) and is defined as follows: $\frac{1}{2}D_{KL}(A||B) + \frac{1}{2}D_{KL}(B||A)$. The KL divergence is also known as *relative entropy* or *information divergence*. The KL divergence is *zero* if the two distributions are identical and approaches infinity as they differ more and more. We measure the KL divergence for the distributions of authors (kl_authors), followers (kl_followers), followees (kl_followees) and friends (kl_friends).

Figure 2 visualizes the different time frames and their notation. t_0 only contains the static features computed from data collected at t_0. Consequently, t_1 and t_2 only contain the static features computed from data collected at t_1 or t_2, respectively. $t_{0\to1}$ includes static features computed on data collected at t_0 and the dynamic measures computed on data collected at t_0 and t_1. $t_{1\to0}$ includes static features computed on data collected at t_1 and the dynamic measures computed on data collected at t_0 and t_1. $t_{1\to2}$ and $t_{2\to1}$ are defined in the same way.

3.2.3 Lexical Measures:

We use vector-based methods which allow representing each microblog message as a vector of terms and use term frequency (*TF*) as weighting schema. In this work lexical measures are always computed for individual time points and are therefore static measures.

3.3 Usage Patterns of Hashtag Categories

Our first aim is to investigate whether different semantic categories of hashtags reveal substantially different usage patterns (such as that they are used and/or consumed by different sets of users or that they are used for different purpose). To compare the pragmatic fingerprints of hashtags belonging to different semantic categories and to quantify the differences between categories, we conducted a pairwise *Mann-Whitney-Wilcoxon-Test* which is a non-parametric statistical hypothesis test for assessing whether one of two samples of independent observations tends to have larger values than the other. We used a non-parametric test since the *Shapiro-Wilk-Test* revealed that not all features are normally distributed, even after applying arcsine transformation to ratio measures. *Holm-Bonferroni* method was used for adjusting the p-values and counteract the problem of multiple comparisons. For this experiment, we used the timeframes $t_{0\to1}$ and $t_{1\to2}$.

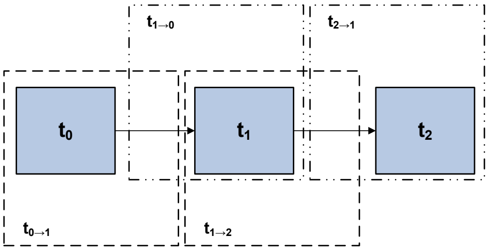


Figure 2: Illustration of our time frames

3.4 Hashtag Classification

Our second aim is to investigate to what extent pragmatic and lexical properties of hashtag streams contribute to classify them into their semantic categories. That means we aim to classify temporal snapshots of hashtag streams into their correct semantic categories (to which they were assigned in [12]) just by analyzing how they are used over time. We then compare the performance of the pragmatically informed classifier with the performance of a classifier informed by lexical features within a semantic multiclass classification task. We used the timeframes $t_{1\to0}$ and $t_{2\to1}$ for this experiment in order to avoid including information from the 'future' in our classification.

We performed grid search with varying hyperparameters using Support Vector Machine (linear and RBF kernels) and an ensemble method with extremely randomized trees. Since extremely randomized trees are a probabilistic method and perform slightly different in each run, we run them ten times and report the average scores. The features were standardized by subtracting the mean and scaling to unit variance. We used stratified 6-fold cross-validation (CV) to train and test each classification model.

Since we have two different types of pragmatic features, static and dynamic ones, we trained and tested three separate classification models which were only informed by pragmatic information:

Static Pragmatic Model: We trained and tested this classification model with static pragmatic features on data collected at t_1 using stratified 6-fold CV. The experiment was repeated on the data collected at t_2.

Dynamic Pragmatic Model: We trained and tested the classification model with dynamic pragmatic features on data collected at t_0 and t_1 using stratified 6-fold CV. The computation of our dynamic features requires at least two time points. We repeated this experiment on data collected at t_1 and t_2.

Combined Pragmatic Model: We combined the static and dynamic pragmatic features, and trained and tested the classification model on the data of $t_{1 \to 0}$ using stratified 6-fold CV. Again, we repeated the experiment on the data of $t_{2 \to 1}$.

We also performed our classification with a model using our lexical features (i.e., TF weighted words). Finally, we trained and tested a combined classification model using pragmatic and lexical features, which leads to the following classification models:

Lexical Model: We trained and tested the model on data from t_1 using stratified 6-fold CV and repeated the experiment on data collected at t_2.

Combined Pragmatic and Lexical Model: We trained and tested the mixed classifier on the data of $t_{1 \to 0}$ using stratified 6-fold CV, then repeated this experiment for $t_{2 \to 1}$. A simple concatenation of pragmatic and lexical features is not useful, since the vast amount of lexical features would overrule the pragmatic features. Therefore, we used a stacking method (see [2]) and performed firstly a classification using lexical features alone and Leave-One-Out cross-validation. We used a SVM with linear kernel for this classification since it worked best for these features. Secondly, we combined the pragmatic features with the resulting seven probability features which we got from the previous classification model and which describe how likely each semantic class is for a certain stream given its words.

To get a fair baseline for our experiment, we constructed a control dataset by randomly shuffling the category labels of the 64 hashtag streams. That means we destroyed the original relationship between the pragmatic properties and the semantic categories of hashtags and evaluated the performance of a classifier which tries to use pragmatic properties to classify hashtags into their shuffled categories within a 6-fold cross-validation. We repeated the random shuffling 100 times and used the resulting average F1-score as our baseline performance. For the baseline classifier we also used grid search to determine the optimal parameters prior to training. Our baseline classifier tests how well randomly assigned categories can be identified compared to our real semantic categories. One needs to note that a simple random guesser baseline would be a weaker baseline than the one described above and would lead to a performance of 1/7.

To gain further insights into the impact of individual properties, we analyzed their information gain (IG) with respect to the categories. The information gain measures how accurately a specific stream property P is able to predict stream's category C and is defined as follows: $IG(C, P) = H(C) - H(C \mid P)$ where H denotes the entropy.

4. RESULTS

In the following section, we present the results from our empirical study on usage patterns of different semantic categories of hashtags.

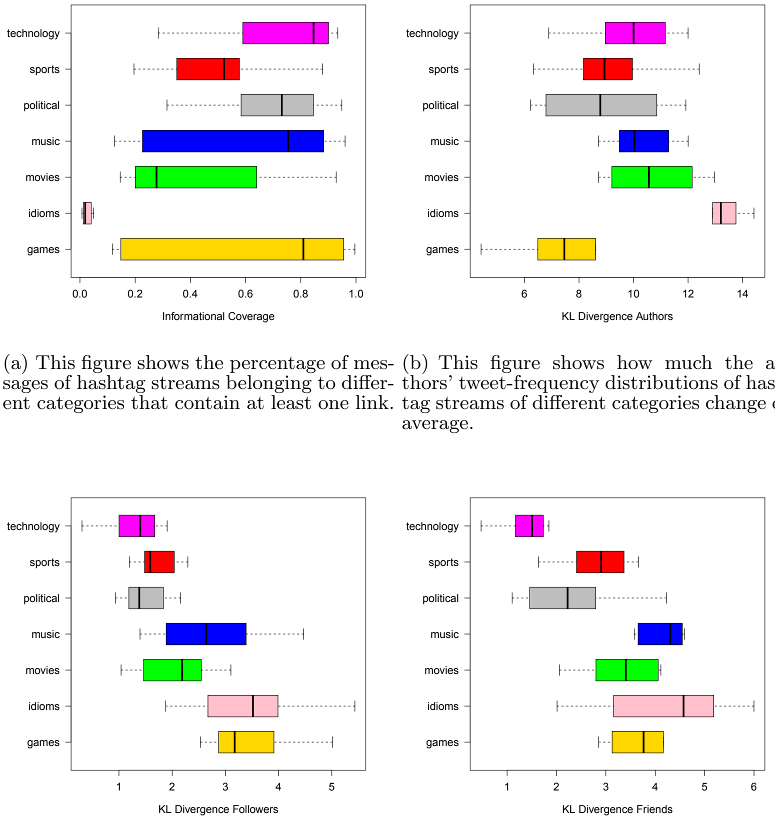
4.1 Usage Patterns of Hashtag Categories

To answer our first research question, we explored to what extent usage patterns of hashtag streams in different semantic categories are indeed significantly different.

Our results indicate that some pragmatic measures are indeed significantly different for distinct semantic categories. This indicates that hashtags of certain categories are used in a very specific way which may allow us to relate these hashtags with their semantic categories just by observing how users use them. Table 3 depicts the measures that show statistically significant ($p < 0.05$) differences in both $t_{0 \to 1}$ and $t_{1 \to 2}$. In total, 35 pragmatic category differences were found to be statistically significant (with $p < 0.05$) for $t_{0 \to 1}$ and 33 for $t_{1 \to 2}$. 26 pragmatic category differences were found to be significant for both $t_{0 \to 1}$ and $t_{1 \to 2}$ which suggests that results are independent of our choice of time frame.

Table 3: Features which showed a statistically significant difference (with $p < 0.05$) for each pair of categories in both $t_{0 \to 1}$ and $t_{1 \to 2}$

	games	**idioms**	**movies**	**music**	**political**	**sports**
idioms	informational retweet					
movies		informational				
music		informational				
political	kl_followers	kl_authors kl_followers kl_followees informational hashtag				
sports	kl_followers	kl_authors kl_followers informational				
technology	kl_followers	kl_authors kl_followers kl_followees kl_friends informational retweet hashtag	kl_friends	kl_friends	overlap_authorfollower overlap_authorfriend	



(a) This figure shows the percentage of messages of hashtag streams belonging to different categories that contain at least one link.

(b) This figure shows how much the authors' tweet-frequency distributions of hashtag streams of different categories change on average.

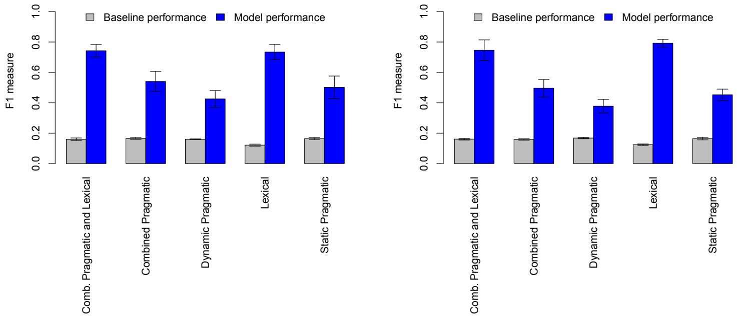
(c) This figure shows how much the followers' author-frequency distributions of hashtag streams of different categories change on average.

(d) This figure shows how much the friends' author-frequency distributions of hashtag streams of different categories change on average.

Figure 3: Each plot shows the feature distribution of different categories of one of the 4 best pragmatic features for $t_{0\rightarrow 1}$. We obtained similar results for $t_{1\rightarrow 2}$.

Not surprisingly, the category which shows the most specific usage patterns is *idioms* and therefore the hashtags of this category can be distinguished from all hashtags just by analyzing their pragmatic properties. Hashtag streams of the category *idioms* exhibit a significantly lower informational coverage than hashtag streams of all other categories (see Figure 3(a)) and a significantly higher symmetric KL divergence for author's tweet-frequency distributions (see Figure 3(b)). Also the followers' and friends' author-frequency distributions tend to have a higher symmetric KL divergence for idioms hashtags than for other hashtags (see Figures 3(c) and 3(d)). This indicates that the social structure of hashtag streams in the category *idioms* changes faster than hashtags of other categories. Furthermore, hashtag streams of this category are less informative - i.e., contain significantly less links per message on average.

The category *technology* can be distinguished from all other categories except *sports*, particularly because its followers' and friends' author-frequency distributions have significantly lower symmetric KL divergences than hashtags in the categories *games*, *idioms*, *movies* and *music* (see Figures 3(c) and 3(d)). This indicates that hashtag streams in the category *technology* have a stable social structure which changes less over time. This is not surprising since this semantic category denotes a topical area and users who are interested in such areas may consume and provide information on a regular base. It is especially interesting to note that the only pragmatic measures which allows distinguishing political and technological hashtag streams are the author-follower and author-friend overlaps since these overlaps are significantly lower for the category *technology* compared to the category *political.*



(a) $t_{1\to0}$ (b) $t_{2\to1}$

Figure 4: Weighted averaged F1-scores of different classification models trained and tested on $t_{1\to0}$ 4(a) and $t_{2\to1}$ 4(b) using 6-fold cross-validation

This indicates that the content of hashtag streams of the category *political* is far more likely to be produced and consumed by the same people than content of technological hashtag streams.

Comparing the individual measures reveals that the informational coverage (six category pairs) and the symmetric KL divergences of followers' author-frequency distributions (six category pairs), authors' tweet-frequency distributions (three pairs) and friends' follower-frequency distributions (three pairs) are the most discriminative measures. Figure 3 depicts the distributions of these four measures per category. Other measures that show significant differences in medians for both $t_{0\to1}$ and $t_{1\to2}$ are the symmetric KL divergence of followees' author-frequency distributions (two pairs), the author-follower and the author-friend overlap (one pair) as well as the retweet and hashtag coverage (two pairs).

Some measures like the conversational coverage measure did not show any significant differences for any of the category pairs, for any time frame. This indicates that in all hashtag streams an equal amount of conversational activities take place.

4.2 Hashtag Classification

In order to quantify the value of different pragmatic and lexical properties of hashtag streams for predicting their semantic category, we conducted a hashtag stream classification experiment and systematically compared the performance of various classification models trained with different sets of features.

Figure 4 shows the performance of the best classifier (extremely randomized trees) trained with different sets of features. One can see from this figure that in general lexical features perform better than pragmatic features, but also that pragmatic features (both static and dynamic) significantly outperform a random baseline. This indicates that pragmatic features indeed reveal information about a hashtag's meaning, even though they do not match the performance of lexical features in this case. In 4(a) we can see that for $t_{1\to0}$ the combination of lexical and pragmatic features performs slightly better than using lexical features alone.

Table 4: Top features for two different datasets ranked via Information Gain

Rank	$t_{1\to0}$	$t_{2\to1}$
1	informational	kl_followers
2	kl_followers	informational
3	kl_friends	hashtag
4	hashtag	kl_followees
5	norm_entropy_friend	kl_friends

4.2.1 Feature Ranking:

In addition to the overall classification performance which can be achieved solely based on analyzing the pragmatics of hashtags, we were also interested in gaining insights into the impact of individual pragmatic features. To evaluate the individual performance of the features we used information gain (with respect to the categories) as a ranking criterion. The ranking was performed on $t_{1\to0}$ and $t_{2\to1}$ with stratified 6-fold cross-validation. Table 4 shows that the top five features (i.e., the pragmatic features which reveal most about the semantic of hashtags) are features which capture the temporal dynamics of the social context of a hashtag (i.e., the temporal follower, followees and friends dynamics) as well as the informational and hashtag coverage. This indicates that the collective purpose for which a hashtag is used (i.e., if it used to share information rather than for other purposes) and the social dynamics around a hashtags – i.e., who uses a hashtag for whom – play a key role in understanding its semantics.

5. DISCUSSION OF RESULTS

Although our results show that lexical features work best within the semantic classification task, those features are text and language dependent. Therefore, their applicability is limited to settings where text is available. Pragmatic features on the other hand rely on usage information which is independent of the type of content which is shared in social streams and can therefore also be computed for social video or image streams.

We believe that pragmatic features can supplement lexical features if lexical features alone are not sufficient. In our experiments, we could see that the performance may slightly increase when combining pragmatic and lexical features. However, the effect was not significant. We think the reason for this is that in our setup lexical features alone already achieved good performance.

The classification results coincide with the results of the statistical significance tests. Ranking the properties by information gain showed that the most discriminative properties (the ones that showed a statistical significance in both $t_{0\rightarrow 1}$ and $t_{1\rightarrow 2}$ for the highest amount of category pairs) found in 4.1 were also the top ranked features (informational coverage and the KL divergences).

6. CONCLUSIONS AND IMPLICATIONS

Our work suggests that the collective usage of hashtags indeed reveals information about their semantics. However, further research is required to explore the relations between usage information and semantics, especially in domains where limited text is available. We hope that our research is a first step into this direction since it shows that hashtags of different semantic categories are indeed used in different ways.

Our work has implications for researchers and practitioners interested in investigating the semantics of social media content. Social media applications such as Twitter provide a huge amount of textual information. Beside the textual information, also usage information can be obtained from these platforms and our work shows how this information can be exploited for assigning semantic annotations to textual data streams.

7. ACKNOWLEDGEMENTS

This work was supported in part by a DOC-fForte fellowship of the Austrian Academy of Science to Claudia Wagner. Furthermore, this work is in part funded by the FWF Austrian Science Fund Grant I677 and the Know-Center Graz.

8. REFERENCES

[1] D. Benz, C. Körner, A. Hotho, G. Stumme, and M. Strohmaier. One tag to bind them all: Measuring term abstractness in social metadata. In *Proc. of 8th Extended Semantic Web Conference ESWC 2011, Heraklion, Crete, (May 2011)*, 2011.

[2] T. Hastie, R. Tibshirani, and J. Friedman. *The Elements of Statistical Learning*. Springer Series in Statistics. Springer New York Inc., New York, NY, USA, 2001.

[3] P. Heymann and H. Garcia-Molina. Collaborative creation of communal hierarchical taxonomies in social tagging systems. Technical Report 2006-10, Computer Science Department, April 2006.

[4] J. Huang, K. M. Thornton, and E. N. Efthimiadis. Conversational tagging in twitter. In *Proceedings of the 21st ACM conference on Hypertext and hypermedia*, HT '10, pages 173–178, New York, NY, USA, 2010. ACM.

[5] C. Körner, D. Benz, M. Strohmaier, A. Hotho, and G. Stumme. Stop thinking, start tagging - tag semantics emerge from collaborative verbosity. In *Proceedings of the 19th International World Wide Web Conference (WWW 2010)*, Raleigh, NC, USA, Apr. 2010. ACM.

[6] D. Laniado and P. Mika. Making sense of twitter. In P. F. Patel-Schneider, Y. Pan, P. Hitzler, P. Mika, L. Zhang, J. Z. Pan, I. Horrocks, and B. Glimm, editors, *International Semantic Web Conference (1)*, volume 6496 of *Lecture Notes in Computer Science*, pages 470–485. Springer, 2010.

[7] C. Messina. Groups for twitter; or a proposal for twitter tag channels. `http://factoryjoe.com/blog/2007/08/25/groups-for-twitter-or-a-proposal-for-twitter-tag-channels/`, 2007.

[8] M. G. Noll and C. Meinel. Exploring social annotations for web document classification. In *Proceedings of the 2008 ACM symposium on Applied computing*, SAC '08, pages 2315–2320, New York, NY, USA, 2008. ACM.

[9] M. G. Noll and C. Meinel. The metadata triumvirate: Social annotations, anchor texts and search queries. In *Web Intelligence*, pages 640–647. IEEE, 2008.

[10] S. Overell, B. Sigurbjörnsson, and R. van Zwol. Classifying tags using open content resources. In *Proceedings of the Second ACM International Conference on Web Search and Data Mining*, WSDM '09, pages 64–73, New York, NY, USA, 2009. ACM.

[11] S. Panzeri and A. Treves. Analytical estimates of limited sampling biases in different information measures. *Network: Computation in Neural Systems*, 7(1):87–107, 1996.

[12] D. M. Romero, B. Meeder, and J. Kleinberg. Differences in the mechanics of information diffusion across topics: idioms, political hashtags, and complex contagion on twitter. In *Proceedings of the 20th international conference on World wide web*, WWW '11, pages 695–704, New York, NY, USA, 2011. ACM.

[13] P. Schmitz. Inducing ontology from flickr tags. In *Proceedings of the Workshop on Collaborative Tagging at WWW2006*, Edinburgh, Scotland, May 2006.

[14] M. Strohmaier, D. Helic, D. Benz, C. Körner, and R. Kern. Evaluation of folksonomy induction algorithms. *Transactions on Intelligent Systems and Technology*, 2012.

[15] O. Tsur and A. Rappoport. What's in a hashtag?: content based prediction of the spread of ideas in microblogging communities. In *Proceedings of the fifth ACM international conference on Web search and data mining*, WSDM '12, pages 643–652, New York, NY, USA, 2012. ACM.

[16] C. Wagner and M. Strohmaier. The wisdom in tweetonomies: Acquiring latent conceptual structures from social awareness streams. In *Semantic Search Workshop at WWW2010*, 2010.

[17] L. Wittgenstein. *Philosophical Investigations.* Blackwell Publishers, London, United Kingdom, 1953. Republished 2001.

[18] L. Yang, T. Sun, M. Zhang, and Q. Mei. We know what @you #tag: does the dual role affect hashtag adoption? In *Proceedings of the 21st international conference on World Wide Web*, WWW '12, pages 261–270, New York, NY, USA, 2012. ACM.

3.4 Emergent Usage

This section presents three empirical studies which explore to what extent structural stream properties, usage metadata and semantic metadata may help anticipating users' future activities on a social stream and therefore allow predicting the evolution of a social stream.

Specifically, this thesis focuses on one specific user activity, namely users' reply behavior. In the following publications we aim to anticipate if anyone will reply to a message or not, how many users will reply to the message and who will reply to the message. Anticipating users' reply behavior is central for predicting the attention focus of a group of users or an individual and may therefore help to guide users towards content they might want to react on.

3.4.1 RQ 3: To what extent do structural, usage and semantic metadata help predicting future usage activities in social streams?

In the first publications [Wagner et al., 2012b] [Wagner et al., 2012c] of this section, we investigate to what extent information about the structure, the usage and the semantics of a user stream helps predicting how much attention a message of the user stream will receive. According to the *tweetonomy* model, which was introduced in Chapter 3.2, a user stream $S(U')$ is defined as a tuple $S(U') = (U, M, R, Y', ft)$, where $Y' = \{(u, m, r) \,|\, u \in U' \vee u' \in U', \tilde{m} \in M, r \in R : (u', \tilde{m}, r) \in Y\}$ and $U' \subseteq U$ and $Y' \subseteq Y$. U' and U define both a set of users, M a set of messages, and Y defines a ternary relation $Y \subseteq U \times M \times R$ between U, M, and R. *ft* is a function which assigns to each Y a temporal marker, $ft : Y \rightarrow N$. In words, a user stream contains all messages which are related with a certain set of users $u \in U'$ (which corresponds in the following publications to all active users of a certain forum and a random set of users) and all resources and further users which are related with these messages.

Specifically, we investigate if user streams of different topical communities reveal interesting differences in how they are used and which factors help to predict the future evolution of the stream. Concretely, we explore which semantic and usage metadata contribute most to the prediction of how many replies a message will get. Our findings show that message streams of different forums exhibit interesting differences in terms of how attention is generated. Our results show amongst others, that the purpose of a forum as well as the specificity of the topic of a forum impact which factors drive the reply behavior of a group of users. For example, user streams around very specific topics require messages to fit to the topical focus of the stream in order to attract attention while streams around more general topics do not have this requirement.

In the last publication [Schantl et al., 2013] of this section, we investigate the utility of structural features which captures users' socializing behavior and semantic features which capture users' topical interests for predicting who will reply to a message. Our work suggests that on Twitter conversations are more driven by structural social factors than by semantic factors which means that structural metadata are better features for predicting who will reply to a message on Twitter than semantic metadata.

Ignorance isn't Bliss: An Empirical Analysis of Attention Patterns in Online Communities

Claudia Wagner*, Matthew Rowe†, Markus Strohmaier‡, and Harith Alani†
*Institute of Information and Communication Technologies, JOANNEUM RESEARCH, Graz, Austria
Email: claudia.wagner@joanneum.at
†Knowledge Media Institute, The Open University, Milton Keynes, UK
Email: m.c.rowe@open.ac.uk, halani@open.ac.uk
‡ Knowledge Management Institute and Know-Center, Graz University of Technology,Graz, Austria
Email: markus.strohmaier@tugraz.at

***Abstract*—Online community managers work towards building and managing communities around a given brand or topic. A risk imposed on such managers is that their community may die out and its utility diminish to users. Understanding what drives attention to content and the dynamics of discussions in a given community informs the community manager and/or host with the factors that are associated with attention, allowing them to detect a reduction in such factors. In this paper we gain insights into the idiosyncrasies that individual community forums exhibit in their attention patterns and how the factors that impact activity differ. We glean such insights through a two-stage approach that functions by (i) differentiating between seed posts - i.e. posts that solicit a reply - and non-seed posts - i.e. posts that did not get any replies, and (ii) predicting the level of attention that seed posts will generate. We explore the effectiveness of a range of features for predicting discussions and analyse their potential impact on discussion initiation and progress.**

Our findings show that the discussion behaviour of different communities exhibit interesting differences in terms of how attention is generated. Our results show amongst others that the purpose of a community as well as the specificity of the topic of a community impact which factors drive the reply behaviour of a community. For example, communities around very specific topics require posts to fit to the topical focus of the community in order to attract attention while communities around more general topics do not have this requirement. We also found that the factors which impact the start of discussions in communities often differ from the factors which impact the length of discussions.

***Index Terms*—attention, online communities, discussion, popularity, user generated content**

I. Introduction

Social media applications such as blogs, video sharing sites or message boards allow users to share various types of content with a community of users. For the managers of such communities, the investment of time and money means that community utility is paramount. A reduction in activity could be detrimental to the appearance of the community to outside users, conveying an impression of a community that is no longer active and therefore of little utility. The different nature and intentions of online communities means that what drives attention to content in one community may differ from another. For example, what catches the attention of users in a question-answering or a support-oriented community may not have the same effect in conversation-driven or event-driven communities. In this paper we use the number of replies that a given post on a community message board yields as a measure of its attention.

To explore these and related questions, our paper sets out to study the following two research questions:

1) *Which factors impact the attention level a post gets in certain community forums?*
2) *How do these factors differ between individual community forums?*

Understanding what factors are associated with attention in different communities could inform managers and hosts of community forums with the know-how of what drives attention and what catches the attention of users in their community. Empowered with such information, managers could then detect changes in such factors that could potentially impact community activity and cause the utility of the community to alter.

We approach our research questions through an empirical study of attention patterns in 20 randomly selected forums on the Irish community message board Boards.ie.[1] Our study was facilitated through a two-stage approach that (i) differentiates between seed posts - i.e. thread starters on a community message board that got at least one reply - and non-seed posts - i.e. thread starters which did not get a single reply, and (ii) predicts the level of attention that seed posts will generate - i.e. the number of replies. Through the use of five distinct feature sets, containing a total of 28 features and including *user*, *focus*, *content*, *community* and *post title* features, we analysed how attention is generated in different community forums. We find interesting differences between these communities in terms of what drives users to reply to thread starters initially (through our *seed post identification experiment*) and what factors are associated with the length of discussions (through our *seed post activity level prediction experiment*). Our work

[1] http://www.boards.ie

is relevant for researchers interested in behavioural analysis of communities and analysts and community managers who aim to understand the factors that are associated with attention within a community.

The paper is structured as follows: section 2 describes related work within the fields of attention prediction on different social web platforms. Section 3 describes the dataset and Section 4 describes the features used in our analysis. Section 5 presents our experiments on identifying seed posts and anticipating their attention level in different communities. Section 6 discusses our findings and relates them to previous research. Section 7 concludes the paper with a summary of the key findings gleaned from our experiments and plans for future work.

II. Related Work

Attention on social media platforms can be gauged through assessing the number of replies that a piece of content or user receives. Within this context [1] consider the problem of reciprocity prediction and study this problem in a communication network extracted from Twitter. They essentially aim to predict whether a user A will reply to a message of user B by exploring various features which characterise user pairs and show that features that approximate the relative status of two nodes are good indicators of reciprocity. Our work differs from [1], since we do not aim to predict who will reply to a message, but consider the problem of identifying posts which will start a discussion and predicting the length of discussions. Further, we focus on exploring idiosyncrasies in the reply behaviour of different communities, while the above work studies communication networks on Twitter without differentiating between individual sub-communities which may use Twitter as a communication medium.

The work presented in [2] investigates factors that impact whether Twitter users reply to messages and explores if Twitter users selectively choose whom to reply to based on the topic or, otherwise, if they reply to anyone about anything. Their results suggest that the social aspect predominantly conditions users' interactions on Twitter. Work described in [3] considers the task of predicting discussions on Twitter, and found that certain features were associated with increased discussion activity - i.e. the greater the broadcast spectrum of the user, characterised by in-degree and list-degree levels, the greater the discussion activity. Further, in our previous work [4] we explored factors which may impact discussions on message boards and showed, amongst others, that content features are better indicators of seed posts than user features. Similar to our previous work [4] we also aim to predict discussions on message boards, but unlike past work, which aimed to identify global attention patterns, we focus on exploring and contrasting the discussion behaviour of individual communities.

Closely related to the problem of anticipating the reply-behaviour of social media users is the problem of predicting the popularity and virality of content. For example, the work described in [5] consider the task of predicting the rank of stories on Digg and found that the number of early comments and their quality and characteristics are useful indicators. Hong et al. [6] investigated the problem of predicting the popularity of messages on Twitter measured by the number of future retweets. One of their findings was that the likelihood that a portion of a user's followers will retweet a new message depends on how many followers the user has and that messages which only attract a small audience might be very different from the messages which receive huge numbers of retweets. Similar work by [7] explored the relation between the content properties of tweets and the likelihood of the tweets being retweeted. By analysing a logistic regression model's coefficients, Naveed et al. [7] found that the inclusion of a hyperlink and using terms of a negative valence increased the likelihood of the tweet being retweeted. The work of [8] explores the retweet behaviour of Twitter users by modeling individual micro-cosm behaviour rather than general macro-level processes. They present four retweeting models (general model, content model, homophily model, and recency model) and found that content based propagation models were better at explaining the majority of retweet behaviours in their data. Szabo et al. [9] studied content popularity on Digg and YouTube. They demonstrated that early access patterns of users can be used to forecast the popularity of content and showed that different platforms reveal different attention patterns. For example, while Digg stories saturate fairly quickly (about a day) to their respective reference popularities, YouTube videos keep attracting views throughout their lifetimes. In [10] the authors present a mutual dependency model to study the virality of hashtags in Twitter.

Although its is well-known that sub-communities of users can be identified on most social media applications, previous research did not explore differences in the attention patterns of such sub-communities. To the best of our knowledge, our work is the first to focus on exploring idiosyncrasies of communities' attention patterns by comparing the reply behaviour of different community forums. We also provide an extended set of features to assess the effects that community and focus features have on reply behaviour, something which has not been explored previously.

III. Dataset: Boards.ie

In this work, we analysed data from an Irish community message board, Boards.ie, which consists of 725 community forums ranging from communities around specific computer games or spiritual groups to communities around general topics such as films or music. Since our goal is to uncover the idiosyncrasies that individual community forums exhibit and the deltas between them, we selected 20 forums at random.

- *Forum 374 - Weather*: Community of users who have special interest in weather. This forum allows users to talk about the current, future and past weather all over the world and share information - e.g. weather pictures.
- *Forum 10 - Work & Jobs*: The community around this forum consists of users who are looking for jobs, offering jobs and/or are seeking advice in work-related things. This means that the community has, on the one hand,

a support and advice offering purpose and, on the other hand, is a marketplace for users who are in similar situations.

- *Forum 221 - Spanish*: Community of practice where users share a common long-term goal - namely to learn, improve or practice their Spanish.
- *Forum 343 - Golf*: Community of users who are interested in the sport Golf. In this forum users can discuss anything related with golf.
- *Forum 646 - adverts.ie Support*: A support oriented forum for adverts.ie, which is a community based marketplace where individuals can buy or sell items online.
- *Forum 235 - Rip Off Ireland*: Support-oriented forum which aims to help consumers in Ireland avoid being ripped off with the current spate of Euro price hikes.
- *Forum 865 - Home Entertainment (HE) Video Players & Recorders*: Community of users formed around a specific group of products namely HE Video Players and Recorders. In this forum users are seek advice and discuss issues related these products.
- *Forum 544 - Banking & Insurance & Pensions*: Support and advice oriented community of users who seek or provide advice about banking, insurance and pensions.
- *Forum 876 - Construction & Planning*: Forum where users can discuss topics related to construction and planning.
- *Forum 267 - Astronomy & Space*: Information and content-sharing community of users who are interested in astronomy and space.
- *Forum 669 - Google Earth* : Forum where users talk about Google Earth.
- *Forum 55 - Satellite*: Information and content-sharing community where users who are interested in satellite television can discuss this topic.
- *Forum 858 - Economics*: Community of users who have a special interest or expertise in economics.
- *Forum 44 - CTYI*: Community of users around the Centre for the Talented Youth of Ireland (CTYI) which is a youth programme for students between the ages of six and sixteen of high academic ability in Ireland.
- *Forum 538 - Japanese RPG*: Community of users playing Japanese role games.
- *Forum 227 - Television*: Discussion about television related topics such as TV series.
- *Forum 607 - Music Production*: Community of music producers and/or people interested in music and music production in general.
- *Forum 630 - Real-World Tournaments & Events*: Forum where users talk about events and tournaments - i.e. competitions involving a relatively large number of competitors, all participating in a sport, game or event.
- *Forum 190 - North West*: Forum around the North West of Ireland, where users who live in the North West or plan to visit the North West can discuss related questions.
- *Forum 625 - Greystones & Charlesland*: Forum where users talk about everything related with Charlesland and Greystones which are both located about 25 kilometres from Dublin city centre.

For our analysis we use all data published in one of these 20 forums in the year 2006. We use this year to enable comparisons of attention patterns with our previous work [4] over the same time period. Table I describes the properties of the dataset.

TABLE I
DESCRIPTION OF THE BOARDS.IE DATASET.

Forum	ID	Users	Posts	Threadstarter	Seeds
Work & Jobs	10	2371	13964	1741	1435
Music Production	607	308	2018	295	265
Golf	343	394	3361	415	364
Astronomy & Space	267	247	782	141	97
Weather	374	439	7598	233	209
HE Video Players & Recorders	865	134	294	61	52
Banking & Insurance & Pensions	544	956	3514	531	459
Google Earth	669	117	584	37	32
Satellite	55	1516	14704	1714	1620
Economics	858	73	260	28	26
Espanol (Spanish)	221	21	86	31	21
Rip Off Ireland	235	28	329	34	28
Construction & Planning	876	34	202	35	
CTYI	44	39	1505	42	39
Japanese RPG	538	71	1157	75	71
adverts.ie Support	646	304	1227	216	172
Television	227	2086	17442	1238	1139
North West	190	376	4866	291	271
Greystones & Charlesland	625	396	4930	418	382
Real-World Tournaments & Events	630	640	18551	1475	1172

IV. FEATURE ENGINEERING

Understanding what factors drive reply behaviour in online communities involves defining a collection of features and then assessing which are important and which are not. Within our approach setting we can identify the features that impact upon seeding a discussion - through our seed post identification experiments - and how features are associated with seed posts that generate the most attention.

For each thread starter post we computed the features by taking a 6-month window, based on work by [4], [11], prior to when the post was made. That means, we used all the author's past posts within that window to construct the necessary features - i.e. constructing a social network for the user features, assessing the forums in which the posts were made for the focus features and inferring topic distributions per user based the content of posts he/she authored within the previous 6 month. For the features that relied on topic models, we first fit a Latent Dirichlet Allocation [12] model which we use later for inferring users' topic distributions. For training the LDA model we aggregated all posts authored by one user in 2005 into an artificial user document and chose the default hyperparameters ($\alpha = 50/T$, $\beta = 0.01$ and $T = 50$) which we optimised during training by using Wallach's fixed point iteration method [13]. Based on the empirical findings of [14], we decided to place an asymmetric Dirichlet prior over the topic distributions and a symmetric prior over the distribution of words. We used the trained model to infer the average topic distributions (averaged over 10 independent runs of a Markov chain) of a user at a certain point in time by using all posts he/she authored within the last 6 months.

We define five feature sets: user features, focus features, content features, community features and title features, as follows.

A. User Features

User features describe the author of a post via his/her past behaviour, seeking to identify key behavioural attributes that are associated with seed and non-seed posts. For example, a post may only start a lengthy discussion if published by a rather active user.

- *User Account Age:* Measures the length of time (measured in days) that the user has been a member of the community;
- *Post Count:* Measures the number of posts that the user has made.
- *Post Rate:* Measures the number of posts made by the user per day.
- *In-degree:* For the author of each post, this feature measures the number of incoming communication connections to the user.
- *Out-degree:* This feature measures the number of outgoing communication connections from the user.

B. Focus Features

Focus features measure the topical concentration an author. Our intuition is that by gauging the topical focus of a user we will be able to capture his/her areas of interest or expertise. For the first two features, we use the frequency distribution of forums a user has published posts in to approximate his/her interests or expertise, while for the last three features we learn topics from a collection of posts and annotate users with topics by using LDA.

- *Forum Entropy:* Measures the forum focus of a user via the entropy of a user's forum distribution. Low forum entropy would indicate high focus.
- *Forum Likelihood:* Measures the likelihood that the user will publish a post within a forum given the past forum distribution of the user.
- *Topic Entropy:* Measures the topical focus of a user via the entropy of a user's topic distributions inferred via the posts he/she authored. Low topic entropy would indicate high focus.
- *Topic Likelihood:* Measures the likelihood that the user will publish a post about certain topics given the past topic distribution of the user's posts. Therefore, we measure how well the user's language model can explain a given post by using the likelihood measures:

$$likelihood(p) = \sum_{i=0}^{N_p} \ln P(w_i|\hat{\phi}, \hat{\theta}) \tag{1}$$

  N_p refers to the total number of words in the post, $\hat{\phi}$ refers to the word-topic matrix and $\hat{\theta}$ refers to the average topic distribution of a user's past posts. The higher the likelihood for a given post, the greater the post fits to the topics the user has previously written about.
- *Topic Distance:* Measures the distance between the topics of a post and the topics the user wrote about in the past. We use the Jensen-Shannon (JS) divergence to measure the distance between the user's past topic distribution and the post's topic distribution. The JS divergence is defined as follows:

$$D_{JS} = \frac{1}{2}D_{KL}(P||A) + \frac{1}{2}D_{KL}(A||P) \tag{2}$$

  where $D_{KL}(P||A)$ represents the Kullback Leibler divergence between a random variable P and A. The KL divergence is calculated as follows:

$$D_{KL}(P||A) = \sum_i P(i) \log \frac{P(i)}{A(i)} \tag{3}$$

  The lower the JS divergence, the greater the post fits the topics the user has previously written about.

C. Post Features

Post features describe the post itself and identify attributes that the content of a post should contain in order to start a discussion. For example, a post may only start a lengthy discussion if its content is informative or if it was published at a certain time in the day.

- *Post Length:* Number of words in the post.
- *Complexity:* Measures the cumulative entropy of terms within the post, using the word-frequency distribution, to gauge the concentration of language and its dispersion across different terms.
- *Readability:* Gunning fog index using average sentence length (ASL) [15] and the percentage of complex words (PCW): $0.4 * (ASL + PCW)$ This feature gauges how hard the post is to parse by humans.
- *Referral Count:* Count of the number of hyperlinks within the post.
- *Time in day:* The number of minutes through the day from midnight that the post was made. This feature is used to identify key points within the day that are associated with seed or non-seed posts.
- *Informativeness:* The novelty of the post's terms with respect to other posts. We derive this measure using the Term Frequency-Inverse Document Frequency (TF-IDF) measure.
- *Polarity:* Assesses the average polarity of the post using Sentiwordnet.[2] Let n denote the number of unique terms in post p, the function $pos(t_.)$ returns the positive weight of the term $t_.$ from the lexicon and $neg(t_.)$ returns the negative weight of the term. We therefore define the polarity of p as:

$$\frac{1}{n}\sum_{i=1}^{n} pos(t_i) - neg(t_i) \tag{4}$$

[2] http://sentiwordnet.isti.cnr.it/

D. Community Features

Community features describe relations between a post or its author and the community with which the post is shared. For example, members of a community might be more likely to reply to a post which fits their areas of interest or they might be likely to reply to someone who contributed a lot to discussions in the past.

- *Topical Community Fit:* Measures how well a post fits the topical interests of a community by estimating how well the post fits into the forum. We measure how well the community's language model can explain the post by using the likelihood measure which is defined in equation 1, where $\hat{\theta}$ refers to the average topic distribution of posts that were previously published in that forum. The higher the likelihood of the post, the better the post fits to the topics of this community forum.
- *Topical Community Distance:* Measures the distance between the topics of a post and the topics the community discussed in the past. We use the Jensen-Shannon (JS) divergence to measure the distance between a community's past topic distribution and a post's topic distribution. The JS divergence is defined in equation 2. The lower the JS divergence, the greater the post fits the topical interests of the community.
- *Evolution score:* Measures how many users of a given community have replied to a user in the past, differing from *in-degree* by being conditioned on the forum. Theories of evolution [16] suggest a positive tendency for user A replying to user B if A previously replied to B. Therefore, we define the evolution score of a given user u_j as follows:

$$evolution(u_j) = \sum_{i}^{U} \frac{U(u_{j,i}) + 1}{U} \tag{5}$$

  where U refers to the total number of users in a given forum and $U(u_{j,i})$ refers to the number of users who replied to user u_j in the past.
- *Inequity score:* Measures how many users of a given community a user has replied to in the past, differing from *out-degree* by being conditioned on the forum. Equity Theory [17] suggests a positive tendency for user A replying to user B if B previously replied more often to A than A to B. Therefore, we define the inequity score of a user u_j as follows:

$$inequity(u_j) = \sum_{i}^{U} \frac{|P(u_{i,j})_{reply}|}{|P(u_{j,i})_{reply} + 1|} \tag{6}$$

  where U refers to the total number of users in a given forum, $P(u_{i,j})_{reply}$ refers to the probability that user u_i replies to user u_j and $P(u_{j,i})_{reply}$ refers to the probability that user u_j replies to user u_i

E. Title Features

Title features describe the title of a post itself and identify attributes that the title should contain in order to start a discussion. We decided to separate title features from post features in order to be able to capture potential affects of the user interface since the current Boards.ie user interface encourages users to decide which post to read based on the title. Therefore, our intuition is that in some community forums, title features may have a greater influence on the start of discussions as well as on the development of lengthy discussions.

- *Title Length:* Number of words in the title of the post.
- *Title Question-mark:* Measures the absence or presence of a question-mark in the title.

V. Experiments

Understanding what drives attention in different forums and their implicit communities enables us to reveal key differences between those forums. To detect such deltas we apply our two-stage prediction approach to (i) detect seed posts within each forum and (ii) predict the level of activity that such seed posts will generate. We begin by explaining our experimental setup before going on to discussing our findings and observing how the communities differ from one another in their discussion dynamics.

A. Experimental Setup

For our experiments we took all the thread starter posts - i.e. that were both seeds and non-seeds - published in each of the 20 forums throughout the year 2006. For each thread starter we constructed the features as described in the previous section. We performed two experiments using our generated datasets, each intended to explore the research questions: (i) *Which factors may impact the attention level a post gets in certain community forums?* and (ii) *How do these factors differ between individual community forums?*

1) Seed Post Identification: The first experiment sought to identify the factors that help differentiating between posts that initiate discussions and posts that do not get any attention in different communities. To this end, we performed *seed post identification* through a binary classification task using a logistic regression model. For each forum, we divided the forum's dataset into a training/testing split using an 80/20% split, trained the logistic regression model using the former split and applied it to the latter. We tested each of the five feature sets in isolation - i.e. user, focus, post, community and title - such that the model was trained using only those features, and then tested all the features combined together. To assess how well each model performed, we measured the F1 score, which is the harmonic mean of precision and recall, and the Matthews correlation coefficient (MCC), which is a balanced measure of the quality of binary classification and can be used even if the classes are of very different sizes. The MCC measure returns a value between -1 and $+1$: a coefficient of $+1$ represents a perfect prediction, 0 is no better than random prediction and -1 indicates total disagreement between prediction and observation. The F1 score is frequently used by the Information Retrieval community, while the MCC is widely used by the Machine Learning community and in statistics where it is known as phi (ϕ) coefficient.

The best performing model was then chosen based on the F1 score and MCC value and the coefficients of the logistic regression model were inspected to detect how the features were associated with seed posts, thereby identifying the factors which impact reply behaviour. To gain further insights into which features contribute most to the classification model, we also ranked the features of the best performing model by using the Information Gain Ratio (IGR) as a ranking criterion.

2) Activity Level Prediction: For our second experiment, we sought to identify the factors that were correlated with lengthy discussions and how they differed between communities. To do this we performed *seed post activity level prediction* through a linear regression model. We maintained the same splits as in our previous experiment and filtered through the seed posts in the 20% test split using the best performing model in each community. We then trained a linear regression model using the seed posts in the training split and predicted a ranking for the identified seed posts in the test split based on expected discussion volume. This allowed us to pick out the key factors that were associated with generating the most activity by concentrating our rank assessments on the top portion of the posts. We trained the linear regression model using each of the five feature sets in isolation and then used all the features combined together. We chose the best performing model based on its rank prediction accuracy and assessed the statistically significant coefficients of the regression model for the relation between increased attention and its features.

To evaluate our predicted rank, we used the Normalised Discounted Cumulative Gain (nDCG) at varying rank positions, looking at the performance of our predictions over the top-k documents where $k = \{1, 5, 10, 20, 50, 100\}$, and then averaging these values. nDCG is derived by dividing the Discounted Cumulative Gain (DCG) of the predicted ranking by the actual rank defined by (iDCG). DCG is well suited to our setting, given that we wish to predict the most popular posts and then expand that selection to assess growing ranks, as the measure penalises elements in the ranking that appear lower down when in fact they should be higher up. Let $rank_i$ be the actual position in the ranking that seed post i should appear and N be the number of items in the total set of seed posts that are to be predicted, we then define $rel_i = N - rank_i + 1$ and DCG based on the definition from [18] as:

$$DCG_k = \sum_{i=1}^{k} \frac{rel_i}{\log_2(1+i)} \tag{7}$$

B. Results: Seed Post Identification

Comparing the F1 score and MCC values of different forums in Table II reveals interesting differences between communities and corroborates our hypothesis that the reply behaviour of users in different communities is impacted by different factors. Table II shows the 9 forums for which a classifier trained with our features outperformed the baseline classifier. We decided not to analyse the results from the other 11 forums, since our classifier did not outperform (but only matched) the performance of the baseline. We assume that this happens because most of these 11 forums are rather inactive forums such as forum 44 or 858 (i.e. only a few messages have been published in 2006 and therefore our classifier had not enough examples of seed and/or non-seed posts to learn general attention patterns). Another potential explanation is that the discussion behaviour of these communities is in part rather random and/or driven by other, external factors which we could not take into account in our study. For example the discussion behaviour of the communities around specific locations or regions (such as community 190 and 625) might for example be impacted by spatial properties of users while the discussion behaviour of the community around forum 227 (Television) seems to be mainly driven by external events (e.g. start of a new series).

Our results from the seed post identification experiment show that for most of the 9 forums a classifier trained with a combination of all features achieves the highest performance boost. Only for the community around forum 267 (Astronomy and Space) a classifier trained with content features alone performs best. This example nicely shows that this community seems to be mainly content driven since its main purpose is to share information and content. Another exception is the community of practice around forum 221 (Spanish) for which a classifier trained with title features alone and a classifier trained with user features alone outperforms a classifier trained with all feature groups. This indicates that the features of those two groups best capture the characteristics of seed and non-seed posts in this community.

To gain further insights into the factors that impact attention in different communities we inspected the statistically significant coefficients of the best performing feature group learned by the logistic regression model. The coefficients can be interpreted as the log-odds for the features. Therefore, a positive coefficient denotes a higher probability of getting replies for posts having this feature. In addition to interpreting the statistically significant coefficients we also ranked the features of the best performing feature group by using the Information Gain Ratio (IGR) as a ranking criterion. The higher the information gain of a feature the higher the average purity of the subsets that it produces. A feature with a maximum information gain ratio of 1 would enable perfect separation between seed and non seed posts. Due to space constraints we only discuss features with an $IGR >= 0.1$.

Our results suggest that in the community around forum 10 (Work & Jobs) which has a support and marketplace function, longer posts (content length's *coef* $= 0.063$ and $p < 0.001$) which do not really contain new information (informativeness *coef* $= -0.028$ and $p < 0.001$) and/or links (*coef* $= -0.592$ and $p < 0.01$) are far more likely to get replies. Further, posts which contain question marks (*coef* $= 0.454$ and $p < 0.01$) in their title are more likely to attract the attention of this support-oriented community. Finally, since the topic of this community is quite general, posts are not required to be topically similar to other posts in the forum (community fit's *coef* $= -221.844$

TABLE II
F1 SCORE AND MATTHEWS CORRELATION COEFFICIENT (MCC) FOR DIFFERENT FORUMS WHEN PERFORMING SEED POST IDENTIFICATION. THE BEST PERFORMING MODEL FOR EACH FORUM IS MARKED IN BOLD.

forumid	User		Focus		Content		Community		Title		All	
	MCC	F1	MCC	F1	MCC	F1	MCC	F1	MCC	F1	MCC	F1
10	0.0	0.75	0.0	0.75	0.071	0.76	0.0	0.75	0.0	0.75	**0.1**	**0.766**
607	0.332	0.839	0.0	0.802	0.0	0.802	0.0	0.802	0.0	0.802	**0.359**	**0.857**
343	0.0	0.769	0.0	0.769	0.093	0.782	0.0	0.769	0.0	0.769	**0.148**	**0.789**
267	0.078	0.609	-0.132	0.531	**0.242**	**0.673**	0.078	0.609	0.0	0.549	0.181	0.643
865	0.0	0.533	0.0	0.533	0.0	0.533	0.0	0.533	0.0	0.533	**0.632**	**0.815**
544	0.0	0.818	0.0	0.818	-0.052	0.809	0.0	0.818	0.0	0.818	**0.109**	**0.828**
55	0.0	0.913	0.0	0.913	0.0	0.913	0.0	0.913	0.0	0.913	**0.144**	**0.918**
221	0.447	0.625	-0.447	0.25	0.0	0.486	0.0	0.333	**0.707**	**0.829**	0.0	0.333
630	0.0	0.678	0.0	0.678	-0.044	0.675	0.0	0.678	0.0	0.678	**0.109**	**0.686**

and $p < 0.01$) in order to attract attention.

Another support and advise oriented community is the community around forum 343 (Golf). The topic of this community is a more specific than the topic of the previous community. In this community the content of a post needs to be rather complex (*coef* $= 2.261$ and $p < 0.01$) and should also not contain links (*coef* $= -0.586$ and $p < 0.05$) in order to attract attention. Further posts which are topically distinct from what the Golf community usually talks about (community distance *coef* $= -4.528$ and $p < 0.05$) are less likely to get replies. This indicates that within the community specialist terminology is used and the divergence away from such vocabularies reduces the likelihood of generating attention to a new post.

The community around forum 865 (HE Video Players & Recorders) has an advice seeking and experience sharing purpose but only for one specific group of products. For this community forum all features' coefficients are not significant. However, a classification model trained with all features outperformed a random baseline classification model with a MCC value of 0.632. By looking at the feature list ranked by the *IGR*, we note that only one feature contributed to this performance boost, namely the inequity score (*IGR* $= 0.7$). The coefficient of the inequity score in the regression model is negative (*coef* $= -5.025$) which indicates that a post is less likely to get replies if it is authored by a user who replied to many posts in this forum in the past but hasn't got many replies himself in this forum. One possible explanation is that in support oriented communities users who reply to many posts are more likely to be experts. It is not surprising that posts of such expert users are less likely to get replies since less users have enough expertise to answer or comment on the post of an expert.

The main purpose of the community around forum 544 (Banking & Insurance & Pensions) is also for seeking advice and sharing experiences and information. In this community shorter posts (content length *coef* $= -0.017$ and $p < 0.05$) authored by users who are new to the topic - or have not published anything about the topic before (topic distance *coef*= 2.890 and $p < 0.01$) - are more likely to get replies. When inspecting the *IGR* based feature ranking of the content group, we find that only the complexity of content is a useful feature for informing a classifier which has to differentiate between seed and non seeds (*IGR* $= 0.354$). This indicates that short, but complex posts which have been authored by newbies are most likely to catch the attention of this community.

The main purpose of the community around forum 267 (Astronomy & Space) is to share information and content and to engage in discussions. Long posts (*coef* $= 0.083$ and $p < 0.05$) which do not contain many novel terms (informativeness *coef* $= -0.029$ and $p < 0.05$) but are positive in their sentiment (polarity's *coef* $= 4.556$ and $p < 0.05$) are very likely to attract the attention of this community. The content feature with the highest *IGR* is the number of links per post (*IGR* $= 0.1$). Since the coefficient of the number of links is positive in our regression model we can conclude that a higher number of links indicates that the post is more likely to get replies (*coef* $= 0.157$) in this forum. This suggests that in this forum posts which are long, informative and re-use the vocabulary of the community are more likely to attract attention.

Also for the topical community around forum 55 (Satellite) the main purpose is to share information and content and to engage in discussions. In this community posts authored by users who have a high forum likelihood are less likely to get replies (*coef* $= -5.891$ and $p < 0.01$). This suggests that users who stimulate discussions in this community have to focus their activity away from this forum. Further posts which are topically distant from the topics the community usually talks about are again less likely to get replies (*coef* $= -2.944$ and $p < 0.01$). This pattern indicates that users who focus their activity away from this community and then post a new thread that is about topics which seem to be in the topical interest area of the community are more likely to get replies.

The community around forum 221 (Spanish) is a community of practice which means that the community members have a common interest in a particular domain or area, and learn from each other. This community is mainly impacted by user and title factors, however all features' coefficients are not significant. Ranking the features by their *IGR* shows that the most important feature for discriminating between posts getting replies and posts not getting replies is the title length (*IGR* $= 0.558$). Interestingly in this forum, posts with short titles are more likely to get replies. The longer the title the less likely a post gets replies (title length's *coef* $= -0.326$). The second most important feature is the user account age (*IGR* $= 0.381$). Users who have owned an account for longer are more likely to get replies in this forum than users who recently created their account. This suggests that in communities where the members share a common long-term goal and/or have a shared interest which is rather stable over time, the duration of users' community membership is a good feature to predict if a post will become a seed post or not.

The community around forum 630 is a rather open and diverse community of users who are interested in all kind of events and/or want to promote events. For forum 630 (Real-World Tournaments & Events) a classifier trained with all features performed best. The only significant feature for this forum is the community distance (*coef* $= -1.185$ and $p < 0.05$). This indicates that posts which do not fit the topical interests of this community are less likely to get replies.

C. Results: Activity Level Prediction

To explore which factors may affect the number of replies a post gets, we first identified the feature groups which lead to the best model for each community forum (see Table III) and then analysed the statistically significant coefficients of the best performing model from each community.

Interestingly, our results suggest that the factors that impact whether a discussion starts around a post tend to differ from the factors that impact the length of a discussion. For example for the support and advise oriented community around forum 343 (Golf) content and community features contribute most to the identification of seed posts, but focus features are most important for predicting the activity level of discussions around seed posts. This indicates that it is important that a post's content has certain characteristics (e.g. contains only few links) and fits the topical interests of the community in order to start a discussion, but afterwards it is important that the author of a post has certain topical and/or forum focus in order to stimulate a lengthy discussion in this forum. In forum 865 (HE Video Players & Recorders) the seed post identification works best when using features from all feature groups, but for predicting the activity level a post will produce a linear regression model trained with content features works best. This indicates that posts which manage to stimulate lengthy discussions in this forum share some content characteristics. Also for the community around forum 544 (Banking & Insurance & Pensions) which also has an advice seeking purpose a model using all feature groups performs best in the seed post identification task. However, for predicting the length of discussions which a seed post will generate a model trained with community features only ranked the posts most accurately according to their discussion length. This suggest that in this forum it makes a difference who authored a post and how this person relates to the community when predicting the discussion length around a post. For the topical community around forum 55 (Satellite) the main purpose is to share information and content and to discuss satellite television. Also in this community a model trained with all feature groups performs best in the seed post identification task. However for predicting the discussion length of seed posts a regression model trained with title features only works best. This indicates that in this community title features impact if a post will stimulate a long discussion. Our results show that seed posts with longer titles (*coef*=0.03003 and $p < 0.05$) are more likely to stimulate lengthy discussions.

For certain communities, such as the community around forum 267 (Astronomy & Space) whose main purpose is to share information and content, the same group of features, namely content features, works best for identifying posts around which a discussion will start and for predicting the length of a discussion. This indicates that in this community users' discussion behaviour is mainly impacted by characteristics of posts' content and therefore content features alone are sufficient to predict users' reply behaviour. Other factors play a minor role in this community.

For the community around forum 630 (Real-World Tournaments & Events) and the community around forum 10 (Work & Jobs) a classification model using all features performs best in both tasks, the seed post identification and the activity level prediction tasks. For the community around forum 10 (Work & Jobs) our results show that posts authored by users who replied to many other users in the past (*coef* of users' out-degree is 0.005 and $p < 0.01$) and have longer titles (*coef*=0.034 and $p < 0.01$) are more likely to stimulate lengthy discussions than other posts. One potential explanation is that posts with longer titles are more likely to attract the attention of this community and that users in this community are more likely to be involved in lengthy discussions with users who have replied to them before. For the community around forum 630 our results suggest that posts authored by users with a high inequity score are more likely to lead to lengthy discussions (*coef*=0.0015 and $p < 0.05$). This suggests that in this community rather active users who frequently reply to other community members' posts but do not get many replies themselves are most likely to stimulate lengthy discussions. It seems that users in this community are more likely to reply to posts of other users who replied to their own posts in the past. Also in this community posts with longer titles are slightly more likely to stimulate lengthy discussions (*coef*=0.04145 and $p < 0.001$). One potential explanation for that is that posts with longer titles tend to catch the attention of more users who then read the post and reply to it. However, one needs to note that although the effect is statistically significant the effect size is very small which indicates that the dependent variable (discussion length) is expected to only increase slightly when that independent variable (title length) increases by one, holding all the other independent variables constant.

Finally, in the community of practice around forum 221 (Spanish) no lengthy discussions happened within the selected time period and therefore we could not analyse factors that impact lengthy discussions.

VI. Discussion of Results

Our findings from the seed post identification experiment demonstrate that different community forums exhibit interesting differences in terms of how attention is generated and that

TABLE III
AVERAGED NORMALISED DISCOUNTED CUMULATIVE GAIN $nDCG@k$ VALUES USING A LINEAR REGRESSION MODEL WITH DIFFERENT FEATURE SETS. A $nDCG@k$ OF 1 INDICATES THAT THE PREDICTED RANKING OF POSTS PERFECTLY MATCHES THEIR REAL RANKING. POSTS ARE RANKED BY THE NUMBER OF REPLIES THEY GOT.

Forum	User	Focus	Content	Commun'	Title	All
10	0.599	0.561	0.452	0.516	0.418	**0.616**
221	0.887	0.954	0.863	0.954	0.88	**0.985**
267	0.63	0.703	**0.773**	0.6	0.75	0.685
343	0.558	**0.727**	0.612	0.634	0.572	0.636
544	0.5	0.514	0.607	**0.684**	0.461	0.574
55	0.574	0.42	0.655	0.671	**0.73**	0.692
607	0.77	0.632	0.814	0.48	0.686	**0.842**
630	0.707	0.459	0.635	0.547	0.485	**0.762**
865	0.673	0.612	**0.85**	0.643	0.771	0.796

the same features which have a positive impact on the start of discussions in one community can have a negative impact in another community. For example, our results from the seed post identification experiment suggest that a high number of links in a post has a negative impact on the post getting replies especially in communities having a supportive purpose (such as community 343 and 10). However, in the community around forum 267, which mainly has an information and content sharing purpose, the contrary is the case. Posts which tend to have many links are more likely to get replies in this community forum. This example nicely shows that the purpose of a community may influence how individual factors impact the start of discussions in a community forum.

It is also interesting to note that for support oriented forums (such as forum 865 and 544) users which seem to be rather new to a topic (i.e. have not published posts before which are topically similar to the content produced by this community) are more likely to get replies. Further, we notice that the importance of whether a post fits the topical focus of a community or not is largely dependent on the subject specificity of the community. In other words communities around very specific topics (such as the community around the sport Golf) require posts to match the topical focus of the community in order to attract attention, while communities around more general topics (such as the community around topic Work and Jobs) do not have this requirement.

In our previous work [4] we learnt a general pattern for generating attention on Boards.ie by performing seed post identification using all data from 2006, not just a selection of forums. The best performing model contained all features (user, content and focus), and indicated that the inclusion of hyperlinks was correlated with non-seed posts, while seed posts were those that had a high forum likelihood - i.e. the user had posted in the forum before and was therefore familiar with the forum. The results from our current work have identified the key differences between this general attention pattern and the patterns that each community exhibits. For instance for the 9 analysed forums, 7 perform best when using all features - similar to our previous work - while for the 2 remaining forums, one forum performs best when using content features and another when using title features. Additionally we find differences in the patterns: for forum 55 we find that the lower the forum likelihood the greater the likelihood that the user will generate attention, this being the converse of the general pattern learnt previously [4]. For forums 10 and 343 we find that an increased number of hyperlinks reduces the likelihood of the post generating attention, agreeing with the general attention pattern, while for forum 267 a greater number of hyperlinks increases the likelihood of generating attention.

Our results from the activity level prediction experiment show that the factors that impact whether a discussion starts around a post tend to differ from the factors that impact the length of this discussion. For example, in the community around forum 10 (Work & Jobs) a posts which has question marks in the title is more likely to get a reply but in order to stimulate lengthy discussions it is more important that the title of a post has a certain length rather than that it contains question marks.

It is also interesting to note that the title length is the only feature which has a significant positive impact across several communities on the number of replies a post gets. This suggests that in some communities posts with longer titles are more likely to stimulate lengthy discussions. We assume that this happens because long titles may on the one hand attract more users to read the posts and on the other hand long titles may be correlated with high quality or substantivity of posts's content. It is also likely to be an effect caused by the platform's interface, as users are presented with a list of threads in a given community each of which is listed by its title. The first piece of information, along with the username of the author, that community members see is the title of the post.

We also found a shared attention pattern between the Golf and Real-World Tournaments and Events communities, since in these communities posts which are topically distant from what these communities usually talk about are less likely to stimulate lengthy discussions. Therefore we can conclude that although most attention patterns which we identified in our work are local and community-specific, cross-community patterns also exist and can be identified with our approach.

Comparing these findings to our previously work [4] once again reveals interesting differences between the general pattern learnt across the entirety of Boards.ie for activity level prediction and the per-forum patterns that we have found in this paper. For instance in [4] the general pattern indicated that lower forum entropy and informativeness together with increased forum likelihood lead to lengthier discussions, while for forum 343 we found an increase in forum entropy to be associated with an increase in activity. For the other features none were found to be significant.

VII. CONCLUSIONS, LIMITATIONS AND FUTURE WORK

In this paper, we have presented work that identifies attention patterns in community forums and shows how such patterns differ between communities. Our exploration was facilitated through a two-stage approach that provided novel features able to capture the community and focus information pertaining to the creators of community content.

Our results show that the attention patterns of different communities are impacted by different factors and therefore suggest that these patterns may only be valid in a certain context and that the existence of global, context-free attention patterns is highly questionable. In our previous work [4] we focussed on identifying global attention patterns and found amongst others that posts including links are less like stimulate discussions. In this work we show by analysing attention patterns of individual communities that this global attention pattern is only valid for certain forums. The global attention patterns one learns heavily depend on the mixture and constitution of the sample of communities which one analyses. Therefore, we can conclude that *ignorance isn't a bliss* since understanding the idiosyncrasies of individual communities seem to be crucial for predicting which post will catch the attention of a community and manages to stimulate lengthy discussions in a forum.

We found for example that in support-oriented or advice seeking communities posts which contain many links in their content are less likely to get replies, while in information and content sharing oriented communities a high number of links may even have a positive impact and make posts more likely to attract the attention of such a community. Further we observed that in support-oriented communities especially posts authored by newbies tend to be more likely to get replies. This suggests that the purpose of a community impacts which factors drive the reply behaviour of this community. Beside the purpose of a community we also found that the specificity of the subject of a community may impact which factors explain the discussion behaviour of a community. Communities around very specific topics require posts to fit to the topical focus of the community in order to attract attention while communities around more general topics do not have this requirement. Finally we also found that the factors which impact the start of discussions in communities often differ from the factors which impact the length of discussions.

Although our work is limited to a small number of communities on one message board platform, Boards.ie, it uncovers an interesting problem: the problem of identifying the context in which attention patterns may occur. In our work we use the number of replies a post gets to assess how much attention it attracts. However, we want to point out that the number of replies is just a proxy metric and other metrics such as the number of views could be used as well. Since these metrics tend to be correlated we believe that using other proxy metrics would lead to similar results.

Community managers and hosts invest time, effort and money into providing a community which is useful and attractive to its users. By understanding what factors influence community attention patterns, we can provide actionable information to community managers who are in desperate need for systematic support in decision making and community development. We hope that our research is a first step towards analysing the context in which certain types of behavioural patterns hold. Our future work will further investigate the context of attention patterns in different communities by clustering communities according to the factors which are best for predicting which post will get the attention of a community.

Acknowledgment

Claudia Wagner is a recipient of a DOC-fForte fellowship of the Austrian Academy of Science. The work of Matthew Rowe and Harith Alani was supported by the EU-FP7 project Robust (grant no. 257859).

References

[1] J. Cheng, D. Romero, B. Meeder, and J. Kleinberg, "Predicting reciprocity in social networks," in *he Third IEEE International Conference on Social Computing (SocialCom2011)*, 2011.

[2] D. Sousa, L. Sarmento, and E. Mendes Rodrigues, "Characterization of the twitter @replies network: are user ties social or topical?" in *Proceedings of the 2nd international workshop on Search and mining user-generated contents*, ser. SMUC '10. New York, NY, USA: ACM, 2010, pp. 63–70. [Online]. Available: http://doi.acm.org/10.1145/1871985.1871996

[3] M. Rowe, S. Angeletou, and H. Alani, "Predicting discussions on the social semantic web," in *Extended Semantic Web Conference*, Heraklion, Crete, 2011.

[4] ——, "Anticipating discussion activity on community forums," in *The Third IEEE International Conference on Social Computing*, 2011.

[5] H. Rangwala and S. Jamali, "Defining a Coparticipation Network Using Comments on Digg," *IEEE Intelligent Systems*, vol. 25, no. 4, pp. 36–45, 2010. [Online]. Available: http://dx.doi.org/http://dx.doi.org/10.1109/MIS.2010.98

[6] L. Hong, O. Dan, and B. D. Davison, "Predicting popular messages in twitter," in *Proceedings of the 20th international conference companion on World wide web*, ser. WWW '11. New York, NY, USA: ACM, 2011, pp. 57–58.

[7] N. Naveed, T. Gottron, J. Kunegis, and A. C. Alhadi, "Bad news travel fast: A content-based analysis of interestingness on twitter," in *WebSci '11: Proceedings of the 3rd International Conference on Web Science*, 2011.

[8] S. A. Macskassy and M. Michelson, "Why do People Retweet? Anti-Homophily Wins the Day!" in *Proceedings of the Fifth International Conference on Weblogs and Social Media*. Menlo Park, CA, USA: AAAI, 2011. [Online]. Available: http://www.aaai.org/ocs/index.php/ICWSM/ICWSM11/paper/view/2790

[9] G. Szabo and B. A. Huberman, "Predicting the popularity of online content," *Commun. ACM*, vol. 53, no. 8, pp. 80–88, 2010.

[10] T.-A. Hoang and E.-P. Lim, "Virality and susceptibility in information diffusions," in *ICWSM*, 2012.

[11] J. Chan, C. Hayes, and E. Daly, "Decomposing Discussion Forums using Common User Roles," in *Proceedings of the WebSci10: Extending the Frontiers of Society On-Line*, Apr. 2010.

[12] D. M. Blei, A. Ng, and M. Jordan, "Latent dirichlet allocation," *JMLR*, vol. 3, pp. 993–1022, 2003.

[13] H. M. Wallach, "Structured topic models for language," Ph.D. dissertation, University of Cambridge, 2008.

[14] H. M. Wallach, D. Mimno, and A. McCallum, "Rethinking LDA: Why priors matter," in *Proceedings of NIPS*, 2009. [Online]. Available: http://books.nips.cc/papers/files/nips22/NIPS2009_0929.pdf

[15] R. Gunning, *The Technique of Clear Writing*. McGraw-Hill, 1952.

[16] B. McKelvey, "Quasi-natural organization science," *Organization Science*, vol. 8(4), 1997.

[17] J. Adams, "Inequity in social exchange," *Adv. Exp. Soc. Psychol.*, vol. 62, pp. 335–343, 1965.

[18] C.-F. Hsu, E. Khabiri, and J. Caverlee, "Ranking Comments on the Social Web," in *Computational Science and Engineering, 2009. CSE '09. International Conference*, vol. 4, August 2009.

What Catches Your Attention? An Empirical Study of Attention Patterns in Community Forums

Claudia Wagner
Institute of Information and
Communication Technologies
JOANNEUM RESEARCH
Graz, Austria
claudia.wagner@joanneum.at

Matthew Rowe
Knowledge Media Institute
The Open University
Milton Keynes, United Kingdom
m.c.rowe@open.ac.uk

Markus Strohmaier
Knowledge Management Institute
Graz University of Technology
Graz, Austria
markus.strohmaier@tugraz.at

Harith Alani
Knowledge Media Institute
The Open University
Milton Keynes, United Kingdom
h.alani@open.ac.uk

Abstract

Online community managers work towards building and managing communities around a given brand or topic. A risk imposed on such managers is that their community may die out and its utility diminish to users. Understanding what drives attention to content and the dynamics of discussions in a given community informs the community manager and/or host with the factors that are associated with attention. In this paper we gain insights into the idiosyncrasies that individual community forums exhibit in their attention patterns and how the factors that impact activity differ. We glean such insights by using logistic regression models for identifying seed posts and explore the effectiveness of a range of features. Our findings show that the discussion behaviour of different communities is clearly impacted by different factors.

Introduction

Social media applications such as blogs, video sharing sites or message boards allow users to share various types of content with a community of users. The different nature and intentions of online communities means that what drives attention to content in one community may differ from another. For example, what catches the attention of users in a question-answering or a support-oriented community may not have the same effect in conversation-driven or event-driven communities. In this paper we use the number of replies that a given post on a community message board yields as a measure of its attention and explore factors that impact the attention level a post gets in certain community forums.

Through an empirical study of attention patterns in 10 different forums on the Irish community message board Boards.ie[1], we analysed how attention is generated in different community forums. Our study was facilitated through a classification experiment which aims to identify seed posts - i.e. thread starter posts on a community message board that got at least one reply - and the use of five distinct feature sets - *user*, *focus*, *content*, *community* and *post title* features. We find interesting differences between these communities in terms of what drives users to reply to thread starters initially. Our work is relevant for researchers interested in behavioural analysis of communities and analysts and community managers who aim to understand the factors that are associated with attention within a community.

Dataset: Boards.ie

In this work, we analysed data from an Irish community message board, Boards.ie, which consists of 725 community forums ranging from communities around specific computer games or spiritual groups to communities around general topics such as films or music. For our analysis we used all data published in the year 2006. Table 1 describes the properties of the dataset.

Table 1: Description of the Boards.ie dataset

Posts	Seeds	Non-Seeds	Replies	Users
1,942,030	90,765	21,800	1,829,465	29,908

Since our goal was to uncover the idiosyncrasies of individual community forums and the deltas between them, we selected 10 distinct forums for analysis of their attention patterns. These forums were selected by computing 5 statistics using data from 2005 (*average post count, average number of users, average number of replies, average number of seeds* and *average number of non-seeds per forum*), plotting each community in a PCA space and then selecting forums that appeared away from one another in the space. Table 2 provides a brief description of the selected community forums.

Feature Engineering

Understanding what factors drive reply behaviour in online communities involves defining a collection of features and then assessing which are important for identifying seed posts. We defined the following five feature groups: *User features* describe the author of a post via his/her past behaviour, while *focus features* measure the topical concentration of posts by an author. *Post features* capture characteristics of a post, while *title features* focus on the title of a post itself and identify attributes that the title should contain in order to start a discussion. *Community features* describe relations between a post or its author and the community with which the post is shared. Table 3 provides a brief descrip-

[1] http://www.boards.ie

Table 2: Overview of selected community forums

ID	Name	Description
7	After hours	General discussion forum with the highest level of activity on the platform.
9	Computers and Technology	Computer support-oriented forum containing posts enquiring about issue resolution.
552	Wanted General	Forum where users state items and products that they would like which other users could provide.
483	Cuckoo's Nest	Conversation forum for liberally minded individuals.
47	Motors	Contains posts related to motoring spanning topics such as new cars, purchasing advice and general motoring discussion.
11	Flight Simulator General	Community for discussions about the video game Flight Simulator.
556	Wanted Tickets	Forum for users to state their needs for event tickets, ranging from sports through to music concerts.
468	TCD	Forum for discussions related to Trinity College Dublin (TCD), one of the largest universities in Ireland.
411	Mobile Phones and PDAs	Contains discussions related to mobile phone issues and portable devices that are emerging on the market. Often contains support requests and allows users to resolve problems they are having.
453	Flight Simulator Discs	Forum for the exchange and sale of computer discs for the video game Flight Simulator.

tion of the features we used and relates each feature with a feature group.

For each thread starter post we computed the features by taking a 6-month window prior to when the post was made. That means, we used all the author's past posts within that window to construct the necessary features - i.e. constructing a social network for the user features, assessing the forums in which the posts were made for the focus features and inferring topic distributions per user and month based on the content of posts he/she authored within the previous 6 month. For the features that relied on topic models, we first trained a Latent Dirichlet Allocation (Blei, Ng, and Jordan 2003) model which we use later for inferring users' topic distributions. For training the LDA model we aggregated all posts authored by one user in 2005 into an artificial user document and chose the default hyperparameters ($\alpha = 50/T$, $\beta = 0.01$ and $T = 50$) which we optimised during training. We used this model to infer the monthly average topic distributions (averaged over 10 independent runs of a Markov chain) of users who authored at least one post in 2006 based on all posts they authored within the last 6 months. We use monthly-increments for scalability.

Experimental Setup

Our experiment sought to identify the factors that were associated with discussions in different communities. To that end, we conducted binary classification experiment using a logistic regression model and the features as described in the previous section. For each forum, we divided the forum's dataset into a training/testing split using an 80/20% split, trained the logistic regression model using the former split and applied it to the latter. We tested each of the five feature sets in isolation - i.e. user, focus, post, community and title - such that the model was trained using only those features, and then tested all the features combined together. The best performing model was then chosen and the coefficients of the logistic regression model were inspected to detect how the features were associated with seed posts, thereby identifying the factors that impact reply behaviour of users in different community forums.

To assess how well each model performed, we measured the Area Under the ROC Curve (AUC). A curve that maximises the AUC, and therefore achieves $AUC = 1$, is optimal.

Results: Seed Post Identification

Comparing the AUC values of different forums in Table 4 reveals interesting differences between communities and corroborates our hypothesis that the reply behaviour of users in different communities is impacted by different factors. While content features are most important for community forum 411 (Mobile phones and PDAs), user features are most important for the communities around forum 453 (Flight Simulator Discs) and 483 (Cuckoo's nest). That means that in forum 411 it mainly depends on a post and its characteristics whether the post gets replies or not, while in forum 453 and 483 posts are far more likely to get replies if they were authored by certain types of users. For the communities around forum 556 (Tickets wanted), 552 (Wanted) and 11 (Flight Simulator General), which all have relatively low discussion levels (i.e. many posts get no replies), community features were most important for predicting which post will get replies. It suggests that in those communities only posts and/or users which fit into the community and/or contribute to the community will get replies. Finally, for the communities 7 (After Hours), 9 (Computers and Technology), 468 (TCD) and 47 (Motors), a classifier based on all features performed best in differentiating between posts which get replies and posts which do not stimulate any discussion.

Table 4: Area under the ROC curve (AUC) for different forums when performing seed post identification

Forum	User	Focus	Content	Commun'	Title	All
7	0.612	0.660	0.661	0.536	0.522	**0.711**
9	0.556	0.590	0.559	0.463	0.568	**0.631**
552	0.434	0.469	0.510	**0.532**	0.518	0.502
483	**0.918**	0.890	0.415	0.765	0.530	0.700
47	0.573	0.542	0.631	0.490	0.548	**0.687**
11	0.596	0.539	0.578	**0.604**	0.410	0.603
556	0.434	0.545	0.624	**0.683**	0.465	0.552
468	0.597	0.582	0.473	0.442	0.570	**0.601**
411	0.469	0.468	**0.526**	0.396	0.497	0.489
453	**0.678**	0.602	0.509	0.574	0.585	0.612

To gain deeper insights into the factors which impact users' reply behaviour, we further analysed the coefficients of the logistic regression model which indicate the features' influence on the probability of a post getting replies. In the following we only discuss statistical significant coefficients. For example, when further analysing the Mobile phones and PDAs community, for which content factors seem to play a crucial role, we noted that in this community posts which have a higher polarity ($c = 3.14$) and are therefore more positive are far more likely to get replies. This community seems to be mainly driven by content factors, while characteristics of authors or relations between authors and the rest of the community play a minor role. Community 9 (Computers and Technology) seems to have a supportive purpose. Posts are far more likely to get replies if titles contain question marks ($c = 0.528$), articles ($c = 0.0211$) and negated words ($c = 0.0581$) and if the post's content has high complexity ($c = 0.988$), and therefore uses more expressive language. Outsiders, i.e. users which seem to be rather new to the topic they are writing about (high topic distance $c = 0.970$) and which are not really focused on this particular forum (high forum entropy $c = 0.163$), are more likely to get replies. Interestingly, long titles (title length $c = -0.0109$) and long posts ($c = -0.0103$) have a negative impact on posts getting replies in such support oriented forums. Users who replied to many others (higher outdegree $c = -0.0216$) in the past are also less likely to get replies. Similarly the community around forum 47 (Motors) also seems to have a supportive purpose where content is an important factor for anticipating the start of discussions. Posts which fit into the community (high topical community fit $c = 0.0758$), whose title contains question marks ($c = 0.0554$) and whose content contains a wider vocabulary of terms (high complexity $c = 0.719$) are more likely to catch the attention of this community.

Communities oriented around a very specific subject such as the community in forum 468 (Trinity College Dublin) are more likely to reply to users who are new to the platform (lower user account age $c = -1.58E^{-5}$) and the topic of community's interest (high topic distance $c = -3.53$). The more engaged a user is in a forum (high forum likelihood $c = 0.192$) and the more positive his/her post is (high polarity $c = 3.968$) the more likely he/she will catch the attention of this community. This suggests that naivety of the user plays a role, where a new or prospective student could be asking the community for information about the university. Communities which are oriented around a more general subject, such as the one around forum 7 (After Hours) also require users to engage in a forum (high forum likelihood $c = 6.94$) but do not require them to only focus on one community (high forum entropy $c = 0.379$) in order to get replies. New users (high topic distance $c = 2.00$) which have a topical focus (low topical entropy $c = -0.515$) are likely to get replies. Further, short posts ($c = -0.0117$) which have high complexity ($c = 0.797$) are as well more likely to attract the attention of this community.

Conclusions and Future Work

In this paper, we have presented work that identifies attention patterns in community forums and shows how such patterns differ between communities. Our findings demonstrated that different community forums exhibit interesting differences in terms of how attention is generated. Our results suggest understanding the purpose and nature of a community, including the specificity of its subject, seems to be crucial for identifying the right features to anticipate community behaviour. Communities that seem to have a partly supportive purpose (such as community 9 and 47) tend to be content driven and such communities are more likely to reply to users who are new to the area, not greatly involved in the community and who are seeking help by publishing a post which is about a topic which fits in the community. Communities around very specific subjects (such as the community 468) tend to reply to users who are new to the community and focussed, while communities around more general subjects such as the After Hours community (7) do not have this requirement. In communities that lack specificity everyone can participate, but posts are required to be rather short in order to minimise effort while still containing distinct terms in order to attract attention. We also note that for support-oriented communities there are common patterns in the inclusion of a question-mark and complexity of the language used - requiring an wider vocabulary of terms.

Although our work is limited to a small number of communities on one message board platform, Boards.ie, it uncovers an interesting problem: the problem of identifying the context in which attention patterns may occur. Our results show that the attention patterns of different communities are impacted by different factors and therefore suggest that these patterns may only be valid in a certain context and that the existence of global, context-free attention patterns is highly questionable. Our previous work in (Rowe, Angeletou, and Alani 2011) focussed on identifying global attention patterns and suggested that the initial reply behaviour of communities on Boards.ie tends to be driven by content-factors while our findings show that this is only true for certain types of communities. Our future work will explore this avenue by comparing similar communities for the existance of similar attention patterns.

Acknowledgment

Claudia Wagner is a recipient of a DOC-fForte fellowship of the Austrian Academy of Science. The work of Matthew Rowe and Harith Alani was supported by the EU-FP7 project Robust (grant no. 257859).

References

Adams, J. 1965. Inequity in social exchange. *Adv. Exp. Soc. Psychol.* 62:335–343.

Blei, D. M.; Ng, A.; and Jordan, M. 2003. Latent dirichlet allocation. *JMLR* 3:993–1022.

Gunning, R. 1952. *The Technique of Clear Writing*. McGraw-Hill.

McKelvey, B. 1997. Quasi-natural organization science. *Organization Science* 8(4).

Rowe, M.; Angeletou, S.; and Alani, H. 2011. Anticipating discussion activity on community forums. In *The Third IEEE International Conference on Social Computing*.

Tausczik, Y. R., and Pennebaker, J. W. 2010. The Psychological Meaning of Words: LIWC and Computerized Text Analysis Methods. *Journal of Language and Social Psychology* 29(1):24–54.

Table 3: Overview of the features and their group memberships.

Group	Name	Description
User	User Account Age	Measures the length of time that the user has been a member of the community.
User	Post Count	Measures the number of posts that the user has made.
User	Post Rate	Measures the number of posts made by the user per day.
User	In-degree	Measures the number of incoming connections to the user.
User	Out-degree	Measures the number of outgoing connections from the user.
Focus	Forum Entropy	Measures the forum focus of a user via the entropy of a user's forum distribution. Low forum entropy would indicate high focus.
Focus	Forum Likelihood	Measures the likelihood that the user will publish a post within a forum given the past forum distribution of the user.
Focus	Topic Entropy	Measures the topical focus of a user via the entropy of a user's topic distributions inferred via the posts he/she authored. Low topic entropy would indicate high focus.
Focus	Topic Likelihood	Measures the likelihood that the user will publish a post about certain topics given his/her past topic distribution. Therefore, we measure how well the user's language model can explain a given post by using the likelihood measures: $likelihood(p) = \sum_{i=0}^{N_p} \ln P(w_i\|\hat{\phi}, \hat{\theta}) \qquad (1)$ N_p refers to the total number of words in the post, $\hat{\phi}$ refers to the word-topic matrix and $\hat{\theta}$ refers to the average topic distribution of a user's past posts. The higher the likelihood for a given post, the greater the post fits to the topics the user has previously written about.
Focus	Topic Distance	Measures the distance between the topics of a post and the topics the user wrote about in the past. We use the Jensen-Shannon (JS) divergence to measure the distance between the user's past topic distribution and the post's topic distribution. The lower the JS divergence, the greater the post fits the topics the user has previously written about.
Post	Post Length	Measures the number of words in the post.
Post	Complexity	Measures the cumulative entropy of terms within the post, using the word-frequency distribution, to gauge the concentration of language and its dispersion across different terms.
Post	Readability	This feature gauges how hard the post is to parse by humans by using Gunning fog index (Gunning 1952) which uses average sentence length (ASL) and the percentage of complex words (PCW): $0.4 * (ASL + PCW)$.
Post	Referral Count	Measures the number of hyperlinks within the post.
Post	Time in day	The number of minutes through the day from midnight that the post was made. This feature is used to identify key points within the day that are associated with seed or non-seed posts.
Post	Informativeness	Measures the novelty of the post's terms with respect to other posts. We derive this measure using the Term Frequency-Inverse Document Frequency (TF-IDF) measure.
Post	Polarity	Assesses the average polarity of the post using Sentiwordnet.[2]
Community	Topical Community Fit	Measures how well a post fits the topical interests of a community by estimating how well the post fits into the forum. We measure how well the community's language model can explain the post by using the likelihood measure which is defined in equation 1, where $\hat{\theta}$ refers to the average topic distribution of posts that were previously published in that forum. The higher the likelihood of the post, the better the post fits to the topics of this community forum.
Community	Topical Community Distance	Measures the distance between the topics of a post and the topics the community discussed in the past. We use the Jensen-Shannon (JS) divergence to measure the distance between a community's past topic distribution and a post's topic distribution. The lower the JS divergence, the greater the post fits the topical interests of the community.
Community	Evolution score	Measures how many users of a given community have replied to a user in the past, differing from *in-degree* by being conditioned on the forum. Theories of evolution (McKelvey 1997) suggests a positive tendency for user A replying to user B if A previously replied to B.
Community	Inequity score	Measures how many users of a given community a user has replied to in the past, differing from *out-degree* by being conditioned on the forum. Equity Theory (Adams 1965) suggests a positive tendency for user A replying to user B if B previously replied more often to A than A to B.
Title	Length	Number of words in the title of the post.
Title	Questionmark	Measures the absence or presence of a question-mark in the title.
Title	Linguistic Dimension	Measures the proportion of words per linguistic dimension using LIWC (Linguistic Inquiry and Word Count) (Tausczik and Pennebaker 2010) which categorises 2300 words or word stems into over 70 linguistic dimensions. Rather than using all 70 dimensions we chose five evocative dimensions for our analysis and derived a feature for each one: human terms (e.g. adult, baby), anger (e.g. hate, loathe), sexual (e.g. horny, love), article (e.g. a, an) and negate (e.g. no, not).

The Utility of Social and Topical Factors in Anticipating Repliers in Twitter Conversations

Johannes Schantl
JOANNEUM RESEARCH
Graz, Austria
johannes.schantl@student.tugraz.at

Claudia Wagner
JOANNEUM RESEARCH
Graz, Austria
claudia.wagner@joanneum.at

Rene Kaiser
JOANNEUM RESEARCH
Graz, Austria
rene.kaiser@joanneum.at

Markus Strohmaier
University of Technology
Graz, Austria
markus.strohmaier@tugraz.at

ABSTRACT

Anticipating repliers in online conversations is a fundamental challenge for computer mediated communication systems which aim to make textual, audio and/or video communication as natural as face to face communication. The massive amounts of data that social media generates has facilitated the study of online conversations on a scale unimaginable a few years ago. In this work we use data from Twitter to explore the predictability of repliers, and investigate the factors which influence who will reply to a message. Our results suggest that social factors, which describe the strength of relations between users, are more useful than topical factors. This indicates that Twitter users' reply behavior is more impacted by social relations than by topics. Finally, we show that a binary classification model, which differentiates between users who will and users who will not reply to a certain message, may achieve an F1-score of 0.74 when using social features.

Author Keywords

Twitter, social media communication, reply behavior, reply prediction

ACM Classification Keywords

J.4 Computer Applications: Social and Behavioral Sciences

INTRODUCTION

Social media platforms like Twitter or Facebook are used for interacting and communicating with other users. Many different kinds of conversations, ranging from informal chats to formal discussions, can emerge on these platforms. The massive amounts of data that social media generates has facilitated the study of online conversations on a scale unimaginable a few years ago.

Identifying patterns in online conversations is important for at least two reasons: First, such patterns can be incorporated into the design of online conversation tools (e.g. *orchestrated* video communication systems as described in [9]) and social media services. Second, such patterns can provide an empirical test of social theoretical models that have been proposed in the literature (see e.g. [12]). Therefore, this work sets out to explore patterns in online conversations and investigates the predictability of repliers in Twitter.

When it comes to the theoretical study of online conversations, a natural assumption would be that the closer the friendship between two users A and B, the more likely user A replies to a message of user B and vice versa. A competing hypothesis would be that conversations are driven by topical factors rather than social factors, and that therefore the probability of user A replying to user B increases with their topical similarity – i.e., with the extent to which they talk about the same topics.

In this work, we aim to explore these two competing hypothesis and investigate the following research questions:

- RQ1: To what extent is communication of Twitter users influenced by social and topical factors?
- RQ2: To what extent are repliers on Twitter predictable?

To this end, we measure the predictability of users' reply behavior in Twitter conversations. We propose a comprehensive set of features to quantify the major social and topical factors which may impact users' communication behavior. In addition to topical and social factors we also add activity features (e.g. number of tweets, number of replies or number of followers) as covariates which describe how active, how communicative and how popular a user is on Twitter. We decided to add activity features since we are interested in exploring to what extent social and topical features help predicting repliers above and beyond the effects of activity features.

To address our research questions, we constructed a dataset consisting of user pairs $\langle a, c\rangle$ where either a user c saw a message m authored by user a and replied to it (positive samples), or where a user c saw a message m authored by user a and

WebSci'13, May 2–4, 2013, Paris, France.

did not reply to it (negative samples). In this work we use the variable a to refer to the user who authored the start message of a conversation and the variable c to refer to a potential reply candidate.

Gathering the aforementioned negative samples is obviously difficult since no factual data is available on which tweet has been read by which users. Finding out who has seen a certain message would require approximating unobservable variables such as the time a user spends reading messages which are shown on his/her Twitter timeline, the number of messages which are published on his/her timeline every minute and the extent to which users consume tweets which are not shown on their timeline (e.g. via using Twitter search). In this work we use a simplification and assume that the followers of a user are those users who are likely to see a message authored by this user.

Given this dataset, we first examine which features may have the potential to differentiate between users who see a certain message and reply to it, and users who see the same message but do not reply to it, by conducting statistical hypothesis tests. The null hypothesis states that the users who see the message and reply, and the users who see the message and do not reply, do not differ significantly, i.e., the feature distributions of both user groups are similar. Further, to assess the predictive power of individual features, we conduct a logistic regression analysis using positive and negative user-message pairs as samples. In addition to analyzing the statistically significant coefficients which reveal information about the impact of individual features, we also test the predictive power of the logistic regression model using a 10-fold cross validation.

Our results are in line with results from previous research [18] and suggest that on Twitter social features, which describe the strength of the relation between users, are more useful than topical features for predicting if a user will reply to another user or not. This suggests that conversations on Twitter might be more driven by social relations than by topics. Further, our results show that a binary classification model which aims to differentiate between users who will and users who will not reply to a certain message of another user may achieve an F1-score of 0.76.

This paper is structured as follows: In the next section we introduce some basic terminology used within our work and provide some background information about Twitter. In the *Related Work* Section we discuss research about the nature and the predictability of online conversations in social media applications. In the *Experimental Setup* Section we present our dataset, features and methodology. Our results are described in the *Results* Section. We conclude this work by drawing final conclusions in *Conclusion and Further Work.*

BACKGROUND AND TERMINOLOGY

Twitter was launched in 2006 and is one of the most popular microblogging services in the world. Users may write short messages, called *tweets*, which are limited to 140 characters. Information consumption on Twitter is mainly driven by explicitly defined social networks. That means, a user sees the messages authored by the users he/she follows on their Twitter timeline in reverse chronological order. We call a user u_1 a *follower* of user u_2 if u_1 has established a follow relation with u_2. In the same example, user u_2 is a *followee* of user u_1. We call a user u_3 a *friend* of user u_1 if u_1 has established a follow relation with u_3 and vice versa.

In this work we define a conversation as an interaction between at least two users, consisting of at least two messages, the original start message and the reply message. The Twitter API provides information which allows reconstructing conversation threads since for each message which is a reply, the ID of the message to which it is replying can be retrieved. Therefore, one can recursively find for any reply message the original start message. However, it is not possible to find the end of a conversation without using a temporally restricted definition of a conversation. In our work we therefore decided to predict only the first user who replied to a message rather than all users who replied to it since it is impossible to know if more users will reply to the message in the future. Further, 89% of conversations in our dataset consist of only two users and therefore predicting the first user who replies is most times equal to predicting all users who will join a conversation.

RELATED WORK

Previous research has focused on exploring how users use Twitter in general, and to what extent this platform is used for communication purposes. For example, in one of the first papers about Twitter usage intention, Java et al. [8] found that Twitter is often used for discussing events of daily life, sharing information or URLs, reporting news and for conversations, which we focus on in this study. Java et al. show that 21% of Twitter users participate in conversations, and 1/8 of all Twitter messages are part of conversations. They use the *@mention* sign as indicator for a conversation.

Macskassy et al. [10] show that 92% of dialogues are between two people and that the average number of messages in dialogues is less than 5 tweets. Honeycutt and Herring [7] evaluate conversations in Twitter and give insight about the nature of the *@mention* usage. They found that *@mention* is used in 90.96% for addressivity reasons, and that the median/mean number of users participating in conversations is 2/2.5. Naaman et al. [13] developed a content based categorization system for Twitter messages and found that most users focus on themselves (so-called “meformers”) while less users are “informers”.

Understanding the nature and dynamics of conversations on social media applications like Twitter was also subject of previous studies. For example, in [1] the authors explore the problem of predicting directed communication intention between users who did not communicate with each other before. The authors use various network and content features and conduct a link prediction experiment to assess the predictive power of those features. Their work focuses on predicting only new communication links between users, while our work aims to predict who will reply to a certain message of a certain author no matter if the user has communicated with the author before or not.

Most similar to our work is the work of [18] which explores if the reply behavior of users is mainly driven by topical or social factors. Similar to our findings their findings suggest that social factors are on average more important. For users with larger and denser ego-centric networks, they observed a slight tendency for separating their connections depending on the topics discussed. Unlike our work, their work focuses on three broad topics (sport, religion and politics) and therefore they only analyze the replies of messages which belong to one of these topics. Further, their work focuses on Portuguese tweets while we focus on English tweets. Finally, their work uses a different approach for addressing the same research question as we do. For each pair of topics, they analyze and compare the ego-centric networks of users who have replied to messages from both topics, while we use topical and social features to fit a regression model using user-pairs as observations and the reply-status of a user as dependent binary variable.

Wang and Huberman [20] study the predictability of online interactions both at the group and individual level. They measure the predictability of online user behavior by using information-theoretic methods applied to real time data of online user activities from Epinions, a who-trust-whom consumer review site and Whrrl, a location based online social network game. Their work shows that the users' interaction sequences have strong deterministic components. In addition, they show that individual interactions are more predictable when users act on their own rather than when attending group activities. The work presented in [2] describes an approach for recommending interesting conversations to Twitter users. They are using topic and tie strength between users and preferred thread length as factors to recommend conversations. Their approach gives interesting insights about which conversations different types of users prefer but they don't take into account if the users are also willing to join a conversation.

Research about predicting social links in online social networks is also related to our research about predicting communication links. For example, Rowe et al. [15] study the follow behavior of Sina Weibo users and found that the users' follow behavior is more driven by topical than by social factors. In [16] the authors present an approach that allows inferring social links between users by considering patterns in friendship formation, the content of people's messages and user location. Unlike the aforementioned work, our work solely focuses on communication links rather than on social links (i.e. follower relations). In addition to predicting the existence of social links, researchers also started being interested in predicting the strength of a link. Gilbert et al. [4] try to classify social relations in Facebook into strong and weak ties, referring to user with strong social relation and users with weak social relation. In [3] the authors apply the same approach to Twitter, and found that their Facebook tie strength model largely generalizes to Twitter.

Related to users' reply behavior is also users' retweet behavior and users' question answering behavior. The work of [11] explores the retweet behavior of Twitter users. They present four retweeting models (general model, content model, homophily model, and recency model) and found that content based propagation models were better at explaining the majority of retweet behaviors in their data. That means in contrast to our work they found that content and topics drive the retweet behavior of Twitter users, while we found that the reply behavior is more driven by social factors. Paul et al. [14] conducted a study of question asking and answering behavior on Twitter. They examined what characteristics of the asker might improve his/her chances of receiving a response. They found that the askers' number of followers and their Twitter account age are good predictors of whether their questions will get answered. However, the number of tweets the asker had posted or his/her frequency of use of Twitter do not predict whether his/her question will get answered. Finally, they examined the relationship between asker and replier and found that 36% of relationships are reciprocal and 55% are one-way. Surprisingly, 9% of answerers are not following the askers. Paul et al. focus on one specific type of message, namely questions, while our work is not limited to any message type. Further, they explore characteristics of the questions and the askers in order to predict the number of answers a question will receive, while we are interested in exploring characteristics of user pairs in order to predict if they will communicate with each other or not.

EMPIRICAL STUDY

The aim of our empirical study is to explore how predictable repliers are on Twitter and to what extent users' reply behavior is driven by topical and social factors. In the following Section we describe our experimental setup – i.e., we describe our dataset, features and methodology.

Dataset and Sample Generation

To obtain a random sample of Twitter conversations we firstly crawled Twitter's public timeline[1] by using its publicly available API, and filtered English tweets[2] containing a *reply_to_status_id* – i.e., tweets which were published in reply to another message. Since those tweets are part of a conversation, we reconstructed the conversation thread by recursively crawling all past messages which belong to this conversation. The conversations were crawled on November 20th, 2012 and we obtained 3,850 random conversations in total.

For each conversation we have exactly one positive author-candidate pair which consists of the author of the start message of the conversation and the first user who replied to this message. Further, we randomly selected for each of the remaining conversations one negative sample by selecting one follower of the author of the start message who has not replied to it. We decided to only keep positive author-candidate pairs where the candidate is a follower of the author of the start message, because we wanted to make sure that positive and negative samples are constructed in a consistent way. Surprisingly we had to remove 19.22% sample conversations since users who were not following the author of the message replied to it. This finding confirms the finding of [14] who found that 9% of answerers are not following the askers.

[1] https://dev.twitter.com/docs/api/1/get/statuses/public_timeline

[2] For language detection the *guess_language* python library was used, see: http://pypi.python.org/pypi/guess-language

	median	mean	std
Conversation length	3.0	5.3	12.2
Tweets per user	1,991.9	1,702.2	1,047.7
List memberships per user	0.0	33.2	456.2
Created lists per user	0.0	0.1	0.7
Character length of bio information per user	73.4	68.7	52.4
Followers	266.0	1,524.1	13,819.7
Followees	295.7	1,205.2	8,237.7

Table 1. Characteristics of the dataset consisting of 3,850 conversations from 12,701 different users.

We ended up having 3,215 positive and 3,215 negative samples. For all users who are part of the positive or negative samples (containing 9122 users) we further crawled their most recently published messages (up to 3,200 tweets), their user list memberships, the user lists they created, their user profile information and their followers and followees. We checked that there are no duplicate author/candidate pairs in the positive and negative samples. We want to point out that this information was crawled one day after the conversations were crawled, on the 21th of November 2012. This implies that the information about user's social network, their users lists and their biography may have changed during that day. Therefore features which are based on this information may contain future information which was not available when the conversation happened.

Table 1 shows the basic characteristics of our dataset. The zero median value for the number of participating membership lists and the created membership lists per user indicates that many user do not use or create membership lists. Further one can see from the table that the number of followers per user have a high standard deviation coming from outliers having multiple millions of followers.

Feature Engineering

We introduce three different groups of features. *Topical features* capture the topical similarity between the author of a message and a reply candidate. *Social features* describe the social relationship between the author of a message and a reply candidate. Finally, *Activity features* describe how active and popular a user is on Twitter. We added activity features since we are interested in exploring to what extent social and topical features help predicting repliers beyond the effects of activity features which may function as confounding variables. If we would not take into consideration the users' activity level, we might observe that some social or topical features are highly correlated with a user's reply probability, although they are only correlated with the user's activity level.

Topical Features

Topical features capture the topical similarity between the author of a message and a reply candidate. To identify topics we evaluated three different topic-annotation methods: First we used the concept and keyword extraction service from Alchemy[3], a third party information extraction service, and secondly we used a Twitter-specific Part-of-Speech Tagger (POS)[4]. The tagger reaches an overall tagging accuracy of 90% on Tweets [5] and performs better than the commonly used Stanford POS Tagger for text including abbreviations, interjections, and text which is not grammatically correct written. We decided to keep only proper nouns and hashtags since they often reveal information about the topic of a tweet. In [17] Saif et al. evaluate several open APIs for extraction semantic concepts and entities from tweets. They found that the AlchemyAPI, which we use in our work, extracted the highest number of concepts, and has also the highest entity-concept mapping accuracy. The concept extraction method takes a raw text as input and returns DBpedia[5] concepts and relevance scores as output, while the keyword extraction method extracts relevant unigrams and bigrams from a given input text. We experimented with using Dbpedia concepts, Alchemy generated keywords and POS tagger generated keywords. In this paper we only report the results which we obtained when using topical features produced by the Twitter POS tagger because we obtained the best model fit using this type of topical feature. That means we picked the best performing topical feature. Further in this work we will use the term concept to refer to our topical features.

	median	mean	std
Tweet concepts per user	10.3	8.5	5.6
List concepts per user	0.0	5.4	11.7
Bio concepts per user	1.3	1.9	2.0

Table 2. Number of concepts per user extracted from three types of information provided by a user. First, the aggregation of all tweets written by the user. Second, the aggregation of all membership list names and descriptions the user participates and finally the user's profile description.

We use the following three methods for representing users as documents:

- First, we represent each user as an aggregation of messages which he/she recently published (up to 3,200).
- Second, we represent each user as an aggregation of the names and descriptions of the user lists he/she is a member of.
- Third, we represent each user by his/her personal description obtained from his/her user profile page.

Each topic annotation method combined with each document representation method provides us with a different concept-vector for a user and allows computing the topical similarity between the author of a message and the potential reply candidate based on their concept-vectors. Table 2 shows the mean number of concepts which can be obtained for a user using the different types of user information. Not surprisingly, tweets allow to obtain the highest number of concepts per user, followed by lists and bio information.

We calculate the similarity of the concept-vector of user a and the concept vector of user c using the cosine similarity which

[3] http://www.alchemyapi.com

[4] http://www.ark.cs.cmu.edu/TweetNLP/

[5] http://dbpedia.org

is defined as follows:

$$sim(a,c) = \frac{\langle concepts(a), concepts(c)\rangle}{||concepts(a)|| \cdot ||concepts(c)||} \quad (1)$$

Using the three aforementioned methods for representing users via text and using cosine similarity as similarity measure, for each pair of users $\langle a, c\rangle$ we compute the following features: The *TweetConceptSimilarity* describes how similar two users are, given the concepts they are tweeting about. The *ListConceptSimilarity* describes how similar users' list memberships are, given the concepts the lists are about. Finally, the *BioConceptSimilarity* reveals how topically similar two users are, given the concepts extracted from their personal descriptions on Twitter.

Social Features

Social features capture the strength of the social relation between the author a and a reply candidate c. We introduce the following six social features: The *NumReplyRelation* feature describes how often the reply candidate has communicated with the author in the past. The *ReplyPartnerOverlap* feature reveals if the author and the reply candidate tend to have similar communication partners. The *FriendsOverlap* feature describes how many similar *friends* the author and the reply candidate have in their follower/followee network. The *isFriend* feature is a boolean value describing if the author and candidate have a bidirectional *follower/followee* relation or not. The *CommonListMembership* feature measures the overlap between the list memberships of the author and the candidate – i.e. in how many common lists they are both members. Finally, the *CandInAuthorsList* feature measures the overlap between the lists the author has created and the lists the candidate is member of.

For computing the overlap between the set of users or lists related with the author a ($users(a)$ or $lists(a)$) and the set of users or lists related with the potential reply candidate c ($users(c)$ or $lists(c)$) we use Jaccard similarity coefficient which is defined as follows:

$$Jaccard(a,c) = \frac{|users(a) \cap users(c)|}{|users(a) \cup users(c)|} \quad (2)$$

Activity Features

The third category of features are the activity features. These features capture how active or communicative, and also how popular a reply candidate is. Activity features do not measure any association between the reply candidate and the author but rely solely on characteristics of the candidate. Activity features represent common confounding variables since they might be correlated with some topical and social features. Activity features represent of course not the only confounding factor. For example, external events or happenings or users' current locations might be other confounding variables. However, those factors can unfortunately not be obtained from our observational dataset. However, since we constructed our positive and negative samples randomly (with a slight bias towards active users in the case of positive samples) we can assume that other confounding factors are equally distributed across positive and negative samples.

We compute the following six activity features as follows: The *TweetActivity* feature measures the general activity level of a user on Twitter based on the number of tweets he/she has written in the past. The *AvgTweetActivityLastWeek* feature measures the user's average tweet activity per day within the last week. The *ReplyActivity* feature shows how communicative a user is given the number of reply messages the user has written in the past. The *Openness* feature reveals how open a user is giving the number of users he/she is communicating with. The *Followers* feature captures the popularity of a user given his/her *number of followers*. The *Followees* feature indicates the number of users a user is interested in given his/her *number of followees*.

All feature values are normalized by firstly subtracting the mean in each feature and secondly dividing the values of each feature by its standard deviation. Consequently, values of each feature have zero-mean and unit-variance

Methodology

In this section we describe the methodology which we use to answer our research questions.

Feature Analysis

To answer the first research question (*To what extent is communication of Twitter users driven by social and topical factors?*) we assess the association between each feature and the users' probability of replying. Therefore, we use statistical hypothesis tests and measure the potential of each feature to differentiate between the positive and negative class (i.e., user replies or does not reply). The null hypothesis states that the users who see the message and reply and the users who see the message and do not reply do not differ significantly – i.e., the feature distributions of both user groups are similar. We use the Wilcoxon rank sum test for ordinal features and the Chi-Squared test for categorical features. Unlike the t-test which works best for normally distributed ordinal data, the Wilcoxon rank sum test does not have any requirements in the distribution of the data.

Since the statistical tests compute the significance for each individual feature without taking the combination of features into account, we further use a logistic regression model. The dependent variable in our model is binary and indicates for each author-candidate pair $\langle a, c\rangle$ if the candidate has replied to the author or not. We add the previously described social, topical and activity features as independent variables. A logistic regression model reveals if the discriminative power of a feature persists, given all other variables are held constant.

When multicollinearity appears in a regression model, the standard error of the coefficients tend to be very large, and the coefficients are unreliable. Two commonly used ways for dissolving collinearity are combining the correlated features or neglecting one of them. As Figure 1 shows, the *ReplyPartnerOverlap* and *FriendsOverlap* (Pearson correlation coefficient 0.78) and the *ReplyActivity* and *TweetActivity* (Pearson correlation coefficient 0.76) were highly correlated (i.e. correlation coefficient > 0.75).

For the *ReplyPartnerOverlap* and *FriendsOverlap* we decided to neglect the *FriendsOverlap* because it is based on the *Fol-*

Feature	Description	Mathematical Description
Topical Features		
TweetConceptSimilarity	Cosine similarity between *tweet_concepts* of the candidate c and author a.	$\frac{\langle tweet\_concepts(a), tweet\_concepts(c)\rangle}{\|\|tweet\_concepts(a)\|\| \cdot \|\|tweet\_concepts(c)\|\|}$
BioConceptSimilarity	Cosine similarity between *profile_concepts* candidate c and author a.	$\frac{\langle profile\_concepts(a), profile\_concepts(c)\rangle}{\|\|profile\_concepts(a)\|\| \cdot \|\|profile\_concepts(c)\|\|}$
ListConceptSimilarity	Cosine similarity between *list_concepts* of candidate c and author a.	$\frac{\langle list\_concepts(a), list\_concepts(c)\rangle}{\|\|list\_concepts(a)\|\| \cdot \|\|list\_concepts(c)\|\|}$
Social Features		
CommonListMembership	Jaccard similarity between list memberships of candidate c and author a.	$\frac{\|lists(a) \cap lists(c)\|}{\|lists(a) \cup lists(c)\|}$
CandInAuthorsList	In how many list candidate c appears of author a.	$\frac{\|created\_lists(a) \cap created\_lists(c)\|}{\|created\_lists(a) \cup created\_lists(c)\|}$
NumRepliesRelation	Number of replies of candidate c to Author a in the past.	$replies(a, c).$
ReplyPartnerOverlap	Jaccard similarity between between *reply_partners* of candidate c and author a.	$\frac{\|reply\_partner(a) \cap reply\_partner(c)\|}{\|reply\_partner(a) \cup reply\_partner(c)\|}$
isFriend	Is the candidate c a follower of author a and vice versa.	$isFollowing(a, c) \cap isFollowedby(a, c)$
FriendsOverlap	Jaccard similarity between candidate c and author a given their *friends*.	$\frac{\|friends(a) \cap friends(c)\|}{\|friends(a) \cup friends(c)\|}$
Activity Features		
TweetActivity	Number of tweets posted by the candidate c.	$num\_tweets(c)$
ReplyActivity	Number of replies the candidate c was participating.	$num\_replies(c)$
AvgTweetActivityLastWeek	Average tweets per day the candidate c writing within the last week.	$avg\_tweets\_week(c)$
Openness	Number of users the candidate c was replying to.	$num\_replyingto(c)$
Followers	Number of *followers* of the candidate c.	$num\_followers(c)$
Followees	Number of *followees* of the candidate c.	$num\_friends(c)$

Table 3. Overview of all features used in our empirical study.

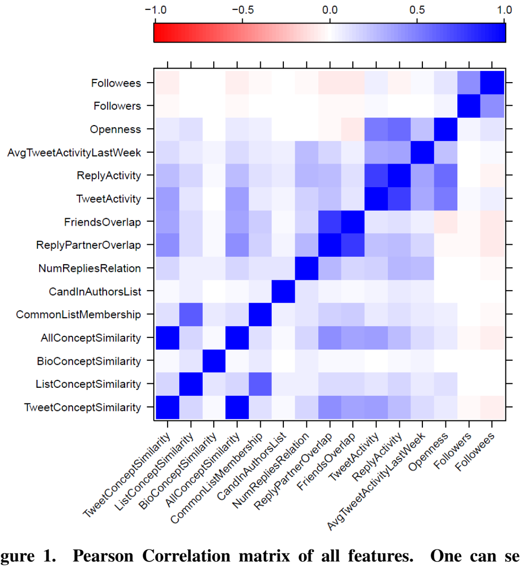


Figure 1. Pearson Correlation matrix of all features. One can see from this figure that the *ReplyPartnerOverlap* and *FriendsOverlap* and the *ReplyActivity* and *TweetActivity* are strongly correlated. When multicollinearity appears in a regression model, the standard error of the coefficients tend to be very large, and the coefficients are unreliable. We solved this issue by neglecting one of the highly correlated features.

lowers and *Followees* information which we crawled one day after the conversation took place. In theory, the social network as well as the list memberships may have changed within this day and therefore the features which rely on this information may contain future information. Finally, for the *ReplyActivitiy* and *TweetActivity* we decided to keep the *ReplyActivity* because we assume that this feature has more power to predict repliers than the more general *TweetActivity*.

After the removal of collinear features we fit the logistic regression model to our dataset. We use *Nagelkerkes pseudo* R^2 measure to assess how well the model fits our data. This value ranges from 0 to 1, where 1 denotes a perfect fit to the observed data and 0 the model doesn't fit at all.

Nagelkerkes pseudo R^2 is defined as follows:

$$Nagelkerkes\_pseudo\_R^2 = \frac{1 - \frac{L(M_{intercept})}{L(M_{full})}^{2/N}}{1 - L(M_{intercept})^{2/N}} \quad (3)$$

where N denotes the number of samples, $L(M_{full})$ refers to the likelihood to obtain the training data when using all features and $L(M_{intercept})$ without using any feature in the logistic regression model.

To gain further insights into the usefulness of individual features, we interpret the statistical significant coefficients of the model. The coefficients returned from a logistic regression model are log-odds ratios and can tell us how the log-odds of a "success" (in our case a reply) changes with a one-unit change in the independent variable.

Prediction Experiment

In addition to looking into the utility of individual features, we are also interested in assessing the predictive power of the whole model in order to answer the second research questions

(*To what extent are repliers on Twitter predictable?*). Therefore we conduct a 10-fold cross validation and train and test the logistic regression model on our dataset. Since our dataset is balanced, i.e. it contains an equal number of positive and negative samples, a random guesser baseline would lead to a performance of 50%. We use *Precision*, *Recall* and the *F1-score* which is the harmonic mean of Precision and Recall as evaluation measures.

RESULTS

In this Section, we present the results from our empirical data analysis which aims to gain insights into patterns of conversations on Twitter and the factors which may potentially drive them.

Feature Analysis

Answering our first research question *RQ1* requires gaining insights into the utility of individual features. Towards that end, we conducted statistical significance tests and fitted a logistic regression model using all features as independent variables and the binary variable (replies or not) as dependent variable.

Statistical Hypothesis Tests

The results from the Wilcoxon rank sum test and the Chi-Squared test show that all features except *Followers* and *BioConceptSimilarity* are statistically significant (see Table 4). This indicates that almost all features are significantly associated with our binary variable (replies or not).

One potential explanation why the *BioConceptSimilarity* seems to be irrelevant is that the bio information of users tends to be short with a mean length of 75 characters per user and that around 14% of the users do not provide any bio information. In our previous work [19] we found that the users' bio information is almost as useful as tweets for predicting users' expertise. However, one needs to note that the dataset we used in [19] was biased towards active expert users who had a high Wefollow[6] rank, while our dataset in this work consist of average users who use Twitter for a conversational purpose. The number of followers seems to be unrelated with users' reply behavior which indicates that users' popularity does not impact their probability of replying.

Regression Analysis

Since the statistical tests compute the significance for each individual feature without taking the combination of features into account, we further fitted a logistic regression model. The dependent variable of our logistic regression model is binary and indicates for each author-candidate pair $\langle a, c \rangle$ if the candidate has replied to the author or not. The previously described social, topical and activity features are added as independent variables.

Table 5 shows the regression coefficients of each feature and their significance level. All features are normalized, so we can rank their influence using their coefficients. Figure 2 shows the distribution of the most significant features for each class. The more the class-specific feature distributions differ, the higher the ability of these features to discriminate the two classes.

Feature	p-Values	Significance
Wilcoxon Rank Sum Test (numerical features)		
TweetConceptSimilarity	4.873e-112	***
ListConceptSimilarity	3.882e-10	***
BioConceptSimilarity	0.4008	
CommonListMembership	1.904e-17	***
CandInAuthorsList	2.740e-08	***
NumRepliesRelation	0.00e+00	***
ReplyPartnerOverlap	3.644e-261	***
FriendOverlap	3.641e-120	***
TweetActivity	3.418e-98	***
ReplyActivity	2.948e-206	***
AvgTweetActivityLastWeek	8.255e-238	***
Openness	2.198731e-93	***
Followees	1.640e-36	***
Followers	0.151	
Chi-Squared Test (categorical features)		
isFriend	2.2e-16	***

Table 4. Results from the statistical hypothesis tests.

One can see from Table 5 that the activity features *AvgTweetActivityLastWeek* and *ReplyActivity* are significant and have a positive coefficient. This demonstrates that the activity level of a user is indeed a significant factor, which influences if a user will reply to a message or not. Not surprisingly, active users are more likely to reply than non active users. The features which are related with the popularity and social status of a user (*Openess* and *Followers*) are not significant which means that the users' reply behavior is not influenced by how open they are or by how many users they follow.

In addition to the activity features, the following social features have a significant positive coefficient – i.e., they help predicting repliers beyond the effects of activity features: *NumRepliesRelation*, *isFriend* and *ReplyPartnerOverlap*. This shows that previous communication relations as well as bidirectional friendship relations are very important for predicting who will reply to a message of a certain user. Friends of the author of the message who have communicated with each other before are more likely to reply than others. The only significantly negative feature is the *Followees* feature. This indicates that the more users a user is following the less likely he/she replies to their messages, as also shown in Figure 2. Intuitively this makes sense as we assume that every user has a maximum number of tweets to which he/she will reply e.g. per hour. The more people a user is following, the more new tweets will show up in his/her timeline. That means the users' reply probability is spread across more tweets and is therefore lower for each individual tweet.

Finally, the logistic regression model shows that topical features like the *TweetConceptSimilarity* and the *BioConceptSimilarity* are also significantly positively correlated with users' reply probability. This indicates that there is a slight tendency that users who are interested into similar topics are more likely to reply to each other. However, one needs to

[6] http://wefollow.com/

	Coefficient	Significance
(Intercept)	-0.0151	
TweetConceptSimilarity	0.1472	***
BioConceptSimilarity	0.0710	*
ListConceptSimilarity	-0.0575	
NumRepliesRelation	2.6073	***
ReplyPartnerOverlap	0.2638	***
CommonListMembership	0.0281	
CandInAuthorsList	0.0727	
isFriend	0.3962	***
ReplyActivity	0.3418	***
AvgTweetActivityLastWeek	0.3505	***
Openness	0.0726	
Followers	0.6063	
Followees	-1.9698	***

Table 5. Results from the logistic regression model using topical, social and activity features as independent variables and reply or not as binary dependent variable.

note that the coefficients of the significant topical features are much smaller than the coefficients of the significant social features. This indicates that users' reply behavior on Twitter is more influenced by social factors than by topical factors.

Prediction Experiment

To answer our second research question *RQ2* we conducted a prediction experiment using the same features as in the aforementioned logistic regression experiment. We trained our logistic regression model and tested the predictive power of the model using a 10 fold cross-validation.

Our results in Table 7 show that when using all three types of features we achieve an average F1-score of 0.76 while a naive baseline (random guesser) would achieve 0.5 since our dataset is balanced. The confusion matrix in Table 6 shows that the model classified more users who replied as non-repliers than users who did not reply as repliers. Interestingly, using social features alone was almost as good as using a combination of all features (F1=0.74). This indicates that social features contribute most to the performance of the classification model. Also, activity features alone performed very well (F1=0.70) as shown in Table 7. This confirms our hypothesis that the activity level of a user is a common confounding variable when analyzing the factors that influence users' reply behavior.

Finally, Table 7 shows that the performance is worst when using topical features alone (F1=0.63). Also Table 8 indicates that a logistic regression model using only topical features as independent variables is worst in explaining the variability in the training dataset, while a combination of all features is best, followed by using social features alone.

	predicted non replier	predicted replier
non replier	2582	633
replier	924	2291

Table 6. Confusion matrix of the logistic regression classification results using all features. The columns of the confusion matrix show the predicted values and the rows show the reference values.

	Precision	Recall	F-Score
All features			
non replier class	0.74	0.80	0.77
replier class	0.79	0.71	0.75
average	0.76	0.76	0.76
Topical features			
non replier class	0.61	0.73	0.67
replier class	0.67	0.54	0.60
Average	0.64	0.64	0.63
Social features			
non replier class	0.70	0.84	0.76
replier class	0.80	0.64	0.71
Average	0.75	0.74	0.74
Activity features			
non replier class	0.67	0.77	0.72
replier class	0.73	0.62	0.67
Average	0.77	0.70	0.70

Table 7. Classification accuracy of our logistic regression model using all features, topical features, social features and activity features.

Our results clearly demonstrate that conversations on Twitter are not driven by topics but by social relations. Further our work shows that in addition to social relations users' activity level plays an important role since more active users are also more likely to reply (i.e., have a higher prior probability of replying). Researchers need to consider activity information since they may function as confounding variables when neglected. Including activity features into our models allows us to conclude that social features help predicting repliers above and beyond the effects of activity features.

CONCLUSIONS, LIMITATIONS AND FUTURE WORK

In this work we conducted an empirical study about the nature and predictability of conversations on Twitter.

Concretely, our work answers the following research questions:

- *RQ1:To what extent is communication of Twitter users influenced by social and topical factors?* Our results show that social features, which describe the strength of the relation between users, help predicting repliers above and beyond the effects of activity features and are more useful than topical features for predicting if a user will reply to another user or not. This suggests that conversations on Twitter are more driven by friendships and social relations rather than topics. The best social features were the *NumRepliesRelation*, the *isFriend* and the *FriendsOverlap* features. This suggests that users are far more likely to reply to a message authored by a user who is a friend of them, to whom they have talked in the recent past frequently and with whom they share common friends.
- *RQ2: To what extent are repliers on Twitter predictable?* Our work shows that a binary classification model that dif-

	all	topical	social	activity
R^2	0.402	0.105	0.337	0.246

Table 8. Goodness of fit of the logistic regression model measured using the Nagelkerke pseudo R^2.

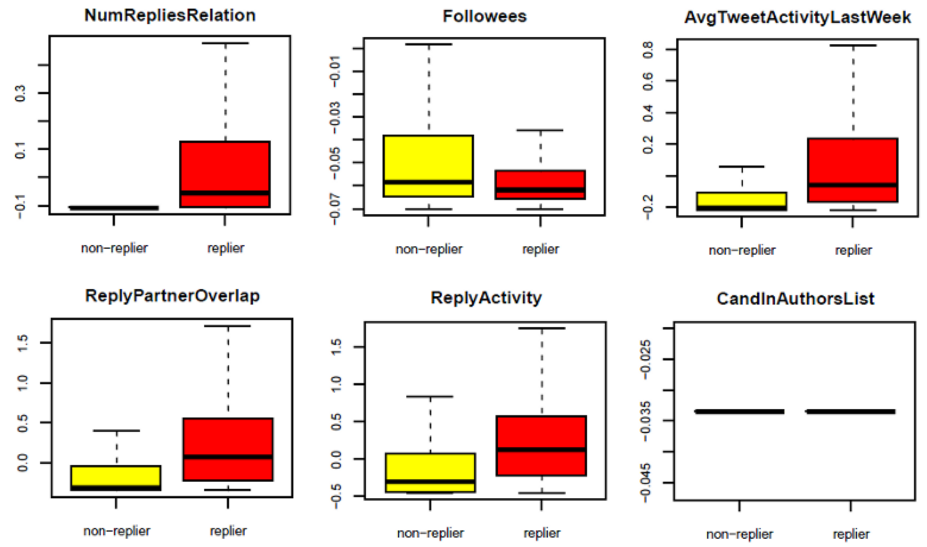


Figure 2. The six most discriminative numerical features from the logistic regression analysis. One can see that users are more likely to reply to a message if they have a high conversation partner overlap with the author of the message (*ReplyPartnerOverlap*) or if they communicated with the author of the message before (*NumRepliesRelation*). Further, users who reply tend to be more active – i.e. they have a higher *AvgTweetActivityLastWeek* and a higher *ReplyActivity*. One can also see that the users who have many *Followees* are less likely to reply.

ferentiates between users who will and will not reply to each other may achieve an F1-score of 0.75 using social, topical and activity features. Using topical features as independent variables leads to the worst statistical model, while using a combination of all features works best, followed by using social features alone. We were able to increase the average F1 score of a random baseline classifier by 24% when using social features alone.

Our work has certain limitations since our assumption that all users who follow a user are similar likely to see messages authored by this user is a simplification which may not reflect the reality. By adding activity features as covariates we addressed this limitation to some extent. Further, this work focuses on the first replier on a single branch of the conversation, and does not take the long-term dynamics of social media conversations into account. We also want to point out that any crawling strategy might introduce a certain bias, as comprehensively studied and described in [6].

In this work we focused on features which can be computed between pairs of users rather than triples (consisting of the two users and the current message) since we are interested in integrating this work into a real-time video communication tool [9] which exploits users' social media stream as background knowledge for orchestrating the video communication. Therefore, it is necessary to compute the features at the beginning of each communication session rather than re-computing them after each message or sentence. For future work we plan to analyze the influence of the current message on users' reply behavior and update the initial communication prediction model during the course of a conversation.

ACKNOWLEDGMENTS

This work was performed within Vconect - Video Communications for Networked Communities (grant agreement no 287760) and received funding from the European Commission under the Seventh Framework Programme (http://www.vconect-project.eu). This work was supported in part by a DOC-fForte fellowship of the Austrian Academy of Science to Claudia Wagner.

REFERENCES

1. Chelmis, C., and Prasanna, V. K. Predicting communication intention in social networks. In *Proceedings of the 2012 ASE/IEEE International Conference on Social Computing and 2012 ASE/IEEE International Conference on Privacy, Security, Risk and Trust*, SOCIALCOM-PASSAT '12, IEEE Computer Society (Washington, DC, USA, 2012), 184–194.
2. Chen, J., Nairn, R., and Chi, E. Speak little and well: recommending conversations in online social streams. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, CHI '11, ACM (New York, NY, USA, 2011), 217–226.
3. Gilbert, E. Predicting tie strength in a new medium. In *CSCW*, S. E. Poltrock, C. Simone, J. Grudin, G. Mark, and J. Riedl, Eds., ACM (2012), 1047–1056.
4. Gilbert, E., and Karahalios, K. Predicting tie strength with social media. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, CHI '09, ACM (New York, NY, USA, 2009), 211–220.
5. Gimpel, K., Schneider, N., O'Connor, B., Das, D., Mills, D., Eisenstein, J., Heilman, M., Yogatama, D., Flanigan,

J., and Smith, N. A. Part-of-speech tagging for twitter: annotation, features, and experiments. In *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies: short papers - Volume 2*, HLT '11, Association for Computational Linguistics (Stroudsburg, PA, USA, 2011), 42–47.

6. Gonzalez-Bailon, S., Wang, N., Rivero, A., Borge-Holthoefer, J., and Moreno, Y. Assessing the Bias in Communication Networks Sampled from Twitter. *Social Science Research Network Working Paper Series* (2012).
7. Honeycutt, C., and Herring, S. C. Beyond microblogging: Conversation and collaboration via twitter. In *Proceedings of the Forty-Second Hawai'i International Conference on System Sciences (HICSS-42). Los Alamitos, CA.*, IEEE Computer Society (Los Alamitos, CA, USA, 2009), 1–10.
8. Java, A., Song, X., Finin, T., and Tseng, B. Why we twitter: understanding microblogging usage and communities. In *Proceedings of the 9th WebKDD and 1st SNA-KDD 2007 workshop on Web mining and social network analysis*, WebKDD/SNA-KDD '07, ACM (New York, NY, USA, 2007), 56–65.
9. Kaiser, R., Weiss, W., Falelakis, M., Michalakopoulos, S., and Ursu, M. F. A rule-based Virtual Director enhancing group communication. In *ICME Workshops*, IEEE (2012), 187–192.
10. Macskassy, S. A. On the study of social interactions in twitter. In *The International AAAI Conference on Weblogs and Social Media (ICWSM)*, The AAAI Press (2012).
11. Macskassy, S. A., and Michelson, M. Why do people retweet? Anti-homophily wins the day! In *ICWSM*, L. A. Adamic, R. A. Baeza-Yates, and S. Counts, Eds., The AAAI Press (2011).
12. Monge, P., and Contractor, N. *Theories of Communication Networks*. Oxford university Press, 2003.
13. Naaman, M., Boase, J., and Lai, C.-H. Is it really about me?: message content in social awareness streams. In *Proceedings of the 2010 ACM conference on Computer supported cooperative work*, CSCW '10, ACM (New York, NY, USA, 2010), 189–192.
14. Paul, S. A., Hong, L., and Chi, E. H. Is twitter a good place for asking questions? A characterization study. In *ICWSM*, L. A. Adamic, R. A. Baeza-Yates, and S. Counts, Eds., The AAAI Press (2011).
15. Rowe, M., Stankovic, M., and Alani, H. Who will follow whom? exploiting semantics for link prediction in attention-information networks. In *International Semantic Web Conference (ISWC)*, vol. 7649 of *Lecture Notes in Computer Science*, Springer (2012), 476–491.
16. Sadilek, A., Kautz, H., and Bigham, J. P. Finding your friends and following them to where you are. In *Proceedings of the fifth ACM international conference on Web search and data mining*, WSDM '12, ACM (New York, NY, USA, 2012), 723–732.
17. Saif, H., He, Y., and Alani, H. Semantic sentiment analysis of twitter. In *The 11th International Semantic Web Conference (ISWC)* (2012), 508–524.
18. Sousa, D., Sarmento, L., and Mendes Rodrigues, E. Characterization of the twitter @replies network: are user ties social or topical? In *Proceedings of the 2nd international workshop on Search and mining user-generated contents*, SMUC '10, ACM (New York, NY, USA, 2010), 63–70.
19. Wagner, C., Liao, V., Pirolli, P., Nelson, L., and Strohmaier, M. It's not in their tweets: Modeling topical expertise of twitter users. In *Proceedings ASE/IEEE International Conference on Social Computing (SocialCom2012)* (2012).
20. Wang, C., and Huberman, B. How random are online social interactions? *Scientific Reports 2* (2012).

4 Conclusions

Previous research analyzed semantics of social streams as well as the structural evolution of social streams by predicting user's future activities. However, previous research mainly focused on analyzing social streams as just another textual document and neglected the fact that social streams emerge through user activities. The publications contained in this thesis explore the relation between the semantics of social streams, their structural evolution and the user activities which guide this evolution. Within several empirical studies this thesis explores if such a relation exists and to what extent this relation can be exploited for (1) the creation of semantic annotations of social streams and users and (2) the prediction of users' future activities in social streams. The findings of this work show that structural stream properties and activity patterns can indeed be exploited for learning semantic annotations of user and hashtag streams. Further, this thesis shows that communities around different topics differ in their activity patterns, which again suggests that there exists a relation between users' activities and the semantic context in which the activities take place.

4.1 Research Results

To conclude and to summarize this work, the answers to the research questions defined in Section 1.3 are presented.

4.1.1 Emergent Structure: What is the emergent structure of social streams and how can we formally describe it?

In this thesis a new model [Wagner and Strohmaier, 2010] which allows to formally describe the complex emerging structure of social streams is introduced. The model can be seen as an extended and extensible version of a folksonomy model. The folksonomy model allows to describe the structure emerging from collaborative tagging systems, while the model presented in this work allows to formally describe a great variety of social streams and is not limited to collaborative tagging systems. Further, various structural stream measures are presented which allow to characterize social streams via structural properties. Finally, this thesis proposes two novel measures for assessing if and to what extent the structure emerging from social streams becomes stable over time and presents empirical investigations on the stabilization process of tag streams [Wagner et al., 2013c].

4.1.2 Emergent Semantics: To what extent may structural metadata and usage metadata contribute to acquiring emerging semantics from streams and annotating streams with semantics?

In several empirical studies this thesis explores to what extent structural and usage metadata can be exploited for semantically annotating social streams. Concretely, this work focuses on user streams and hashtag streams. The results presented in this thesis show amongst others, that a semantic categorization system of hashtags can be constructed by only exploiting structural properties of hashtag streams since different semantic categories of hashtag streams reveal significantly different usage patterns and structural stream properties [Posch et al., 2013]. Further, this work reveals that structural metadata which emerge from users' communication activities in social streams may indeed be useful for creating semantic annotations of user streams [Wagner et al., 2011] and that the audience of a social stream possesses knowledge which may help to interpret the meaning of stream's messages [Wagner et al., 2013b]. Finally, our results show that linguistic-style and social-semantics play a crucial role on social media, since they may reveal information about latent user characteristics such as their professional area and personality-related attributes [Wagner et al., 2013a]. Social-semantics have also shown to be useful for assessing users' expertise areas [Wagner et al., 2012a].

4.1.3 Emergent Usage: To what extent do structural, usage and semantic metadata help predicting future usage activities in social streams?

In several empirical studies this thesis explores to what extent structural, usage and semantic metadata may help anticipating users' future activities in a social stream and therefore allow annotating it with usage metadata. Concretely, this work focuses on usage metadata which captures if anyone will reply to a message or not (i.e., is the message of interest to anyone), how many users will reply to the message (i.e., how interesting is the message from a global perspective) and who will reply to the message (i.e., how interesting is the message from the local perspective of an individual user).

The findings presented in this thesis show that streams around different topics exhibit interesting differences in terms of how attention is generated [Wagner et al., 2012c] [Wagner et al., 2012b] and show that on certain platforms like Twitter users' communication behavior is more driven by structural social factors than it is by topical factors [Schantl et al., 2013]. This indicates that users' behavior in social stream depends on the context (e.g., platform context or topical context) and for predicting users' future activities context-specific prediction models are required.

4.2 Contributions

By answering the previously described research questions the contributions of this dissertation can be outlined as follows:

- First, a network theoretic model of social streams was developed which allows to formally describe the structure that is emerging from social streams. Further, various structural stream measures which enable a comparison of structural properties of social streams and two novel methods for assessing if and to what extent social streams become stable have been introduced.

- Second, this thesis presents empirical evidence that structural metadata combined with usage metadata can be exploited for semantically annotating social streams.
- Finally, the empirical results of this work show that semantic metadata and information about users' activities on social streams (and the structure emerging from those activities) can be exploited for predicting future activities of users in social streams. However, users' activities may differ depending on the context (e.g., platform context or topical context) and therefore context-specific user model may increase the accuracy of predictions about users' future activities.

4.3 Implications of this Work

The network-theoretic model which is introduced in this work, is relevant for researchers interested in social streams and social stream mining since it allows to formally describe social streams and the structure which emerges from them. The model is very general and extensible and allows not only to formally describe existing social streams but also future manifestations of them. The introduced structural stream measures are capable of identifying interesting differences and properties of social streams and are useful for comparing different types of social streams. The proposed measures for assessing the stability of structures emerging from social streams are applicable to a great variety of social streams. Assessing the stability of social stream structures is an important issue since a high stability indicates that although the stream keeps changing the emergent patterns which can be observed remain stable.

In addition to the theoretical model of social streams, the structural stream measures and the stability measures, this work provides empirical evidence for the fact that structural metadata and usage metadata can be exploited for semantically annotating social streams. This has implications for researcher and engineers working on social stream mining techniques since those techniques can benefit from going beyond existing text mining methods by exploiting structural metadata and usage metadata.

While it seems to be intuitive to assume that incorporating any kind of usage metadata related with content leads to better learning algorithms, our results show that not all types of metadata contribute in the same way. Previous research [Rosen-Zvi et al., 2004] as well as our own research [Wagner et al., 2011] show that for example topic models which incorporate usage metadata such as the author topic model tend to overfit data. That means incorporating metadata into a topic model can lead to model assumptions which are too strict and which yield the model to perform worse. This example nicely shows that not all usage metadata contribute equally and that possessing the knowledge about the utility of different usage metadata for different tasks is crucial for researchers and engineers working on new social media mining methods for solving real world problems. This thesis provides knowledge about the utility of different usage metadata for selected tasks such as expertise mining or hashtag categorization.

Further, this work shows that behavioral patterns in different topical forums reveal interesting differences. This suggests that these patterns may only be valid in a certain (semantic) context and that the existence of global, context-free attention patterns is highly questionable. This work also shows that the communication behavior on certain platforms like Twitter is more driven by social than topical factors. These findings have implications for researchers interested in studying users' online behavior since it suggests that the global behavioral patterns which one learns may heavily depend on the constitution of the data sample as well as the platform from which the sample was obtained. Therefore, one can conclude that understanding the idiosyncrasies of individual social streams is crucial for predicting users' future activities on social streams.

4.4 Limitations and Future Work

While the structural framework introduced in this thesis is general and can be used to formally describe and characterize the structure of social streams, the scope of the empirical studies presented in this thesis is limited by the data sets which have been analyzed. Some research directions that can be envisioned from the current status of this work are sketched in the following.

Causal relation between user activities and semantics: This thesis shows within several empirical studies that there exists a relation between user activities which generate a stream (and therefore impact the emergent structure of the stream) and the semantics of the stream. However, this thesis does not explore causal relations between semantics and user activities, but investigates if relational patterns exist. If such relational patterns exist, this work further explores to what extent these patterns can be exploited for developing methods which allow annotating social streams with semantic and usage metadata. Exploring causal relations between semantics and user behavior would require to semantically manipulate a social stream and explore if users' behavior changes in response to the manipulation. Randomized experiments [Aral and Walker, 2011] where users are randomly assigned to certain streams to which they may contribute, provide one method for studying to what extent different (semantic) contexts may impact users' behavior.

Define the (semantic) context in which a theory holds: In science a theory is an attempt to explain and predict behavior in particular contexts. Social theories which try to explain human behavior in particular contexts have been developed over the past few decades and within the last couple of years researchers also started investigating to what extent theories (e.g., homophily and proximity theory) can explain user behavior on the web and in networks (e.g., phone call networks). However, little is known about the context in which those theories hold. The empirical results presented in this thesis suggest that user behavior varies in different (semantic) contexts. However, models which describe properties of a context in which a certain theory holds are missing and have not been subject of the investigations in this thesis.

Theoretical model which explains the stable patterns emerging from social streams: Previous research mainly focused on exploring the imitation behavior of users as a potential cause for the stabilization of tag streams. However, the empirical results presented in this thesis, as well the experimental study of Bollen and Halpin [Bollen et al., 2009] suggest that tagging systems become stable over time regardless of whether tag suggestions are provided to the user or not. Therefore, investigating other factors which may explain the stabilization process (e.g., shared background knowledge or the regularities and stability of natural language [Zipf, 1949] [Cohen et al., 1997] [Ferrer-i Cancho and Elvevåg, 2010]) is a promising avenue for future research.

List of Figures

Bibliography

[Abel et al., 2011] Abel, F., Gao, Q., Houben, G.-J., and Tao, K. (2011). Analyzing User Modeling on Twitter for Personalized News Recommendations. In *International Conference on User Modeling, Adaptation and Personalization (UMAP), Girona, Spain.* Springer.

[Aberer et al., 2004] Aberer, K., Cudré-Mauroux, P., Ouksel, A. M., Catarci, T., Hacid, M.-S., Illarramendi, A., Kashyap, V., Mecella, M., Mena, E., Neuhold, E. J., Troyer, O. D., Risse, T., Scannapieco, M., Saltor, F., Santis, L. D., Spaccapietra, S., Staab, S., and Studer, R. (2004). Emergent semantics principles and issues. In Lee, Y.-J., Li, J., Whang, K.-Y., and Lee, D., editors, *Proceedings of the 9th International Conference on Database Systems for Advanced Applications (DASFAA'04)*, volume 2973 of *Lecture Notes in Computer Science*, pages 25–38. Springer.

[Aral and Walker, 2011] Aral, S. and Walker, D. (2011). Identifying Social Influence in Networks Using Randomized Experiments. *IEEE Expert / IEEE Intelligent Systems*, 26:91–96.

[Asur et al., 2011] Asur, S., Huberman, B. A., Szabó, G., and Wang, C. (2011). Trends in social media : Persistence and decay. *CoRR*, abs/1102.1402.

[Augustson and Minker, 1970] Augustson, J. G. and Minker, J. (1970). An analysis of some graph theoretical cluster techniques. *J. ACM*, 17(4):571–588.

[Begelman, 2006] Begelman, G. (2006). Automated tag clustering: Improving search and exploration in the tag space. In *In Proc. of the Collaborative Web Tagging Workshop at WWW'06.*

[Blei et al., 2003] Blei, D., Ng, A., and Jordan, M. (2003). Latent dirichlet allocation. *J. Mach. Learn. Res.*, 3:993–1022.

[Bollen et al., 2009] Bollen, J., Pepe, A., and Mao, H. (2009). Modeling public mood and emotion: Twitter sentiment and socio-economic phenomena. *CoRR*, abs/0911.1583.

[Bryden et al., 2013] Bryden, J., Funk, S., and Jansen, V. A. A. (2013). Word usage mirrors community structure in the online social network twitter. *EPJ Data Science*, 2(1):3+.

[Burger et al., 2011] Burger, J. D., Henderson, J., Kim, G., and Zarrella, G. (2011). Discriminating gender on twitter. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing*, EMNLP '11, pages 1301–1309, Stroudsburg, PA, USA. Association for Computational Linguistics.

[Canini et al., 2010] Canini, K. R., Suh, B., and Pirolli, P. (2010). Finding relevant sources in twitter based on content and social structure. In *NIPS Workshop*.

[Cattuto et al., 2007] Cattuto, C., Loreto, V., and Pietronero, L. (2007). Semiotic dynamics and collaborative tagging. *Proceedings of the National Academy of Sciences*, 104(5):1461–1464.

[Chelmis and Prasanna, 2012] Chelmis, C. and Prasanna, V. K. (2012). Predicting communication intention in social networks. In *Proceedings of the 2012 ASE/IEEE International Conference on Social Computing and 2012 ASE/IEEE International Conference on Privacy, Security, Risk and Trust*, SOCIALCOM-PASSAT '12, pages 184–194, Washington, DC, USA. IEEE Computer Society.

[Chen et al., 2011] Chen, J., Nairn, R., and Chi, E. (2011). Speak little and well: recommending conversations in online social streams. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, CHI '11, pages 217–226, New York, NY, USA. ACM.

[Chen et al., 2010] Chen, J., Nairn, R., Nelson, L., Bernstein, M. S., and Chi, E. H. (2010). Short and tweet: experiments on recommending content from information streams. In Mynatt, E. D., Schoner, D., Fitzpatrick, G., Hudson, S. E., Edwards, W. K., and Rodden, T., editors, *CHI*, pages 1185–1194. ACM.

[Cheng et al., 2011] Cheng, J., Romero, D., Meeder, B., and Kleinberg, J. (2011). Predicting reciprocity in social networks. In *he Third IEEE International Conference on Social Computing (SocialCom2011).*

[Choudhury et al., 2012] Choudhury, M. D., Counts, S., and Gamon, M. (2012). Not all moods are created equal! exploring human emotional states in social media. In Breslin, J. G., Ellison, N. B., Shanahan, J. G., and Tufekci, Z., editors, *ICWSM*. The AAAI Press.

[Cimiano et al., 2005] Cimiano, P., Hotho, A., and Staab, S. (2005). Learning concept hierarchies from text corpora using formal concept analysis. *Journal of Artificial Intelligence Research (JAIR)*, 24:305–339.

[Cohen et al., 1997] Cohen, A., Mantegna, R. N., and Havlin, S. (1997). Numerical analysis of word frequencies in artificial and natural language texts. *Fractals*.

[Corning, 2002] Corning, P. A. (2002). The re-emergence of "emergence": A venerable concept in search of a theory. *COMPLEXITY*, 7(6):2002.

[Deerwester et al., 1990] Deerwester, S., Dumais, S. T., Furnas, G. W., Landauer, T. K., and Harshman, R. (1990). Indexing by latent semantic analysis. *JOURNAL OF THE AMERICAN SOCIETY FOR INFORMATION SCIENCE*, 41(6):391–407.

[Dellschaft and Staab, 2008] Dellschaft, K. and Staab, S. (2008). An epistemic dynamic model for tagging systems. In *HT '08: Proceedings of the nineteenth ACM conference on Hypertext and hypermedia*, pages 71–80, New York, NY, USA. ACM.

[Dhillon et al., 2001] Dhillon, I. S., Fan, J., and Guan, Y. (2001). Efficient clustering of very large document collections.

[Ferrer-i Cancho and Elvevåg, 2010] Ferrer-i Cancho, R. and Elvevåg, B. (2010). Random Texts Do Not Exhibit the Real Zipf's Law-Like Rank Distribution. *PLoS ONE*, 5(3):e9411+.

[Frey and Dueck, 2007] Frey, B. J. J. and Dueck, D. (2007). Clustering by passing messages between data points. *Science*.

[Gabrilovich and Markovitch, 2007] Gabrilovich, E. and Markovitch, S. (2007). Computing semantic relatedness using wikipedia-based explicit semantic analysis. In *Proceedings of the 20th international joint conference on Artifical intelligence*, IJCAI'07, pages 1606–1611, San Francisco, CA, USA. Morgan Kaufmann Publishers Inc.

[Gemmell et al., 2008] Gemmell, J., Shepitsen, A., Mobasher, M., and Burke, R. (2008). Personalization in folksonomies based on tag clustering. In *Proceedings of the 6th Workshop on Intelligent Techniques for Web Personalization and Recommender Systems.*

[Golbeck et al., 2011] Golbeck, J., Robles, C., Edmondson, M., and Turner, K. (2011). Predicting personality from twitter. In *SocialCom*, pages 149–156. IEEE.

[Golder and Huberman, 2006] Golder, S. A. and Huberman, B. A. (2006). Usage patterns of collaborative tagging systems. *J. Inf. Sci.*, 32(2):198–208.

[Golder and Macy, 2011] Golder, S. A. and Macy, M. W. (2011). Diurnal and seasonal mood vary with work, sleep, and daylength across diverse cultures. *Science*, 333(6051):1878–1881.

[Gower and Ross, 1969] Gower, J. C. and Ross, G. J. S. (1969). Minimum spanning trees and single linkage cluster analysis. *Journal of the Royal Statistical Society. Series C (Applied Statistics)*, 18(1).

[Guy et al., 2013] Guy, I., Avraham, U., Carmel, D., Ur, S., Jacovi, M., and Ronen, I. (2013). Mining expertise and interests from social media. In *Proceedings of the 22nd international conference on World Wide Web*, WWW '13, pages 515–526, Republic and Canton of Geneva, Switzerland. International World Wide Web Conferences Steering Committee.

[Halpin et al., 2007] Halpin, H., Robu, V., and Shepherd, H. (2007). The complex dynamics of collaborative tagging. In *Proceedings of the 16th* *international conference on World Wide Web*, WWW '07, pages 211–220, New York, NY, USA. ACM.

[Hearst, 1992] Hearst, M. A. (1992). Automatic acquisition of hyponyms from large text corpora. In *Proceedings of the 14th conference on Computational linguistics*, pages 539–545, Morristown, NJ, USA. Association for Computational Linguistics.

[Helic et al., 2011] Helic, D., Strohmaier, M., Trattner, C., Muhr, M., and Lerman, K. (2011). Pragmatic evaluation of folksonomies. In *Proceedings of the 20th international conference on World wide web*, WWW '11, pages 417–426, New York, NY, USA. ACM.

[Heymann and Garcia-Molina, 2006] Heymann, P. and Garcia-Molina, H. (2006). Collaborative creation of communal hierarchical taxonomies in social tagging systems. Technical Report 2006-10, Computer Science Department.

[Heymann et al., 2008] Heymann, P., Koutrika, G., and Garcia-Molina, H. (2008). Can social bookmarking improve web search? In *WSDM '08: Proceedings of the international conference on Web search and web data mining*, pages 195–206, New York, NY, USA.

[Hofmann, 1999] Hofmann, T. (1999). Probabilistic Latent Semantic Indexing. pages 50–57.

[Holland, 2006] Holland, J. (2006). Studying Complex Adaptive Systems. *Journal of Systems Science and Complexity*, 19(1):1–8.

[Honeycutt and Herring, 2009] Honeycutt, C. and Herring, S. C. (2009). Beyond microblogging: Conversation and collaboration via twitter. In *Proceedings of the Forty-Second Hawai'i International Conference on System Sciences (HICSS-42). Los Alamitos, CA.*, pages 1–10, Los Alamitos, CA, USA. IEEE Computer Society.

[Hong et al., 2011] Hong, L., Dan, O., and Davison, B. D. (2011). Predicting popular messages in twitter. In *Proceedings of the 20th international conference companion on World wide web*, WWW '11, pages 57–58, New York, NY, USA. ACM.

[Hong and Davison, 2010] Hong, L. and Davison, B. D. (2010). Empirical study of topic modeling in twitter. In *Proceedings of the IGKDD Workshop on Social Media Analytics (SOMA),*.

[Hotho et al., 2006] Hotho, A., Jäschke, R., Schmitz, C., and Stumme, G. (2006). Information retrieval in folksonomies: Search and ranking. In *Proceedings of the 3rd European Semantic Web Conference*, volume 4011 of *LNCS*, pages 411–426, Budva, Montenegro. Springer.

[Hotho et al., 2003] Hotho, A., Staab, S., and Stumme, G. (2003). Wordnet improves text document clustering. In *In Proc. of the SIGIR 2003 Semantic Web Workshop*, pages 541–544.

[Huang et al., 2010] Huang, J., Thornton, K. M., and Efthimiadis, E. N. (2010). Conversational tagging in twitter. In *Proceedings of the 21st ACM conference on Hypertext and hypermedia*, HT '10, pages 173–178, New York, NY, USA. ACM.

[Hughes et al., 2012] Hughes, D. J., Rowe, M., Batey, M., and Lee, A. (2012). A tale of two sites: Twitter vs. facebook and the personality predictors of social media usage. *Comput. Hum. Behav.*, 28(2):561–569.

[i Cancho and Solé, 2001] i Cancho, R. F. and Solé, R. V. (2001). The small world of human language. *Proceedings of The Royal Society of London. Series B, Biological Sciences*, 268:2261–2266.

[Java et al., 2007] Java, A., Song, X., Finin, T., and Tseng, B. (2007). Why we twitter: understanding microblogging usage and communities. In *WebKDD/SNA-KDD '07: Proceedings of the 9th WebKDD and 1st SNA-KDD 2007 workshop on Web mining and social network analysis*, pages 56–65, New York, NY, USA.

[Kang and Lerman, 2011] Kang, J.-H. and Lerman, K. (2011). Leveraging user diversity to harvest knowledge on the social web. In *Proceedings of the Third IEEE International Conference on Social Computing.*

[Kauffman, 1995] Kauffman, S. (1995). *At home in the universe: The search for laws of self-organization and complexity.* Oxford University Press, Oxford.

[Kim et al., 2010] Kim, D., Jo, Y., Moon, I.-C., and Oh, A. (2010). Analysis of twitter lists as a potential source for discovering latent characteristics of users. In *Workshop on Microblogging at the ACM Conference on Human Factors in Computer Systems. (CHI 2010).*

[Körner et al., 2010] Körner, C., Benz, D., Hotho, A., Strohmaier, M., and Stumme, G. (2010). Stop thinking, start tagging: Tag semantics emerge from collaborative verbosity. In *19th International World Wide Web Conference (WWW2010).* ACM.

[Kouloumpis et al., 2011] Kouloumpis, E., Wilson, T., and Moore, J. (2011). Twitter sentiment analysis: The good the bad and the OMG! In *Fifth International AAAI Conference on Weblogs and Social Media.*

[Lambiotte and Ausloos, 2006] Lambiotte, R. and Ausloos, M. (2006). Collaborative tagging as a tripartite network. In *Proceedings of the 6th international conference on Computational Science - Volume Part III*, ICCS'06, pages 1114–1117, Berlin, Heidelberg. Springer-Verlag.

[Laniado and Mika, 2010] Laniado, D. and Mika, P. (2010). Making sense of twitter. In Patel-Schneider, P. F., Pan, Y., Hitzler, P., Mika, P., Zhang, L., Pan, J. Z., Horrocks, I., and Glimm, B., editors, *International Semantic Web Conference (1)*, volume 6496 of *Lecture Notes in Computer Science*, pages 470–485. Springer.

[Lewes, 1879] Lewes, G. H. (1874-1879). *Problems of Life and Mind.* Tuebner, London.

[Macdonald and White, 2009] Macdonald, C. and White, R. W. (2009). Usefulness of click-through data in expert search. In *Proceedings of the 32nd international ACM SIGIR conference on Research and development in information retrieval*, SIGIR '09, pages 816–817, New York, NY, USA. ACM.

[MacQueen, 1967] MacQueen, J. (1967). Some methods for classification and analysis of multivariate observations. In *Proc. Fifth Berkeley Symp. on Math. Statist. and Prob.*, volume 1, pages 281–297. Univ. of Calif. Press.

[Macskassy, 2012] Macskassy, S. A. (2012). On the study of social interactions in twitter. In *The International AAAI Conference on Weblogs and Social Media (ICWSM).* The AAAI Press.

[Macskassy and Michelson, 2011] Macskassy, S. A. and Michelson, M. (2011). Why do people retweet? Anti-homophily wins the day! In Adamic, L. A., Baeza-Yates, R. A., and Counts, S., editors, *ICWSM*. The AAAI Press.

[Markines et al., 2009] Markines, B., Cattuto, C., Menczer, F., Benz, D., Hotho, A., and Stumme, G. (2009). Evaluating similarity measures for emergent semantics of social tagging. In *WWW '09: Proceedings of the 18th international conference on World wide web*, pages 641–650, New York, NY, USA.

[Mathioudakis and Koudas, 2010] Mathioudakis, M. and Koudas, N. (2010). Twittermonitor: trend detection over the twitter stream. In Elmagarmid, A. K. and Agrawal, D., editors, *SIGMOD Conference*, pages 1155–1158. ACM.

[Messina, 2007] Messina, C. (2007). Groups for twitter; or a proposal for twitter tag channels. http://factoryjoe.com/blog/2007/08/25/groups-for-twitter-or-a-proposal-for-twitter-tag-channels/.

[Mika, 2007] Mika, P. (2007). Ontologies are us: A unified model of social networks and semantics. *Web Semant.*, 5(1):5–15.

[Naaman et al., 2010] Naaman, M., Boase, J., and Lai, C.-H. (2010). Is it all about me? user content in social awareness streams. In *Proceedings of the ACM 2010 conference on Computer supported cooperative work*.

[Naveed et al., 2011] Naveed, N., Gottron, T., Kunegis, J., and Alhadi, A. C. (2011). Bad news travel fast: A content-based analysis of interestingness on twitter. In *WebSci '11: Proceedings of the 3rd International Conference on Web Science*.

[Overell et al., 2009] Overell, S., Sigurbjörnsson, B., and van Zwol, R. (2009). Classifying tags using open content resources. In *Proceedings of the Second ACM International Conference on Web Search and Data Mining*, WSDM '09, pages 64–73, New York, NY, USA. ACM.

[Pal and Counts, 2011] Pal, A. and Counts, S. (2011). Identifying topical authorities in microblogs. In *Proceedings of the fourth ACM interna-*

tional conference on Web search and data mining, WSDM '11, pages 45–54, New York, NY, USA. ACM.

[Paul et al., 2011] Paul, S. A., Hong, L., and Chi, E. H. (2011). Is twitter a good place for asking questions? A characterization study. In Adamic, L. A., Baeza-Yates, R. A., and Counts, S., editors, *ICWSM*. The AAAI Press.

[Pennacchiotti and Popescu, 2011] Pennacchiotti, M. and Popescu, A.-M. (2011). Democrats, republicans and starbucks afficionados: user classification in twitter. In *Proceedings of the 17th ACM SIGKDD international conference on Knowledge discovery and data mining*, KDD '11, pages 430–438, New York, NY, USA. ACM.

[Phan et al., 2008] Phan, X.-H., Nguyen, L.-M., and Horiguchi, S. (2008). Learning to classify short and sparse text & web with hidden topics from large-scale data collections. In *Proceedings of the 17th international conference on World Wide Web*, WWW '08, pages 91–100, New York, NY, USA. ACM.

[Posch et al., 2013] Posch, L., Wagner, C., Singer, P., and Strohmaier, M. (2013). Meaning as collective use: Predicting hashtag semantics on twitter. In *3rd Workshop on Making Sense of Microposts at WWW2013*.

[Quercia et al., 2011] Quercia, D., Kosinski, M., Stillwell, D., and Crowcroft, J. (2011). Our twitter profiles, our selves: predicting personality with twitter. In *SocialCom*.

[Rangwala and Jamali, 2010] Rangwala, H. and Jamali, S. (2010). Defining a Coparticipation Network Using Comments on Digg. *IEEE Intelligent Systems*, 25(4):36–45.

[Rao et al., 2010] Rao, D., Yarowsky, D., Shreevats, A., and Gupta, M. (2010). Classifying latent user attributes in twitter. In *Proceedings of the 2nd international workshop on Search and mining user-generated contents*, SMUC '10, pages 37–44, New York, NY, USA. ACM.

[Rapaport, 1988] Rapaport, W. (1988). *Syntactic Semantics: Founda-*

tions of Computational Natural-Language Understanding. Kluwer Academic Publishers.

[Romero et al., 2011] Romero, D. M., Meeder, B., and Kleinberg, J. (2011). Differences in the mechanics of information diffusion across topics: idioms, political hashtags, and complex contagion on twitter. In *Proceedings of the 20th international conference on World wide web*, WWW '11, pages 695–704, New York, NY, USA. ACM.

[Rosen-Zvi et al., 2004] Rosen-Zvi, M., Griffiths, T., Steyvers, M., and Smyth, P. (2004). The author-topic model for authors and documents. In *Proceedings of the 20th conference on Uncertainty in artificial intelligence*, UAI '04, pages 487–494, Arlington, Virginia, United States. AUAI Press.

[Rowe et al., 2011a] Rowe, M., Angeletou, S., and Alani, H. (2011a). Anticipating discussion activity on community forums. In *The Third IEEE International Conference on Social Computing.*

[Rowe et al., 2011b] Rowe, M., Angeletou, S., and Alani, H. (2011b). Predicting discussions on the social semantic web. In *Extended Semantic Web Conference*, Heraklion, Crete.

[Salton and McGill, 1986] Salton, G. and McGill, M. J. (1986). *Introduction to Modern Information Retrieval.* McGraw-Hill, Inc., New York, NY, USA.

[Sanderson and Croft, 1999] Sanderson, M. and Croft, B. (1999). Deriving concept hierarchies from text. In *SIGIR '99: Proceedings of the 22nd annual international ACM SIGIR conference on Research and development in information retrieval*, pages 206–213, New York, NY, USA. ACM.

[Schantl et al., 2013] Schantl, J., Wagner, C., Kaiser, R., and Strohmaier, M. (2013). The utility of social and topical factors in anticipating repliers in twitter conversations. In *ACM Web Science (WebSci2013).*

[Schmitz, 2006] Schmitz, P. (2006). Inducing ontology from flickr tags. In *Proceedings of the Workshop on Collaborative Tagging at WWW2006*, Edinburgh, Scotland.

[Sousa et al., 2010] Sousa, D., Sarmento, L., and Mendes Rodrigues, E. (2010). Characterization of the twitter @replies network: are user ties social or topical? In *Proceedings of the 2nd international workshop on Search and mining user-generated contents*, SMUC '10, pages 63–70, New York, NY, USA. ACM.

[Sriram et al., 2010] Sriram, B., Fuhry, D., Demir, E., Ferhatosmanoglu, H., and Demirbas, M. (2010). Short text classification in twitter to improve information filtering. In *Proceedings of the 33rd international ACM SIGIR conference on Research and development in information retrieval*, SIGIR '10, pages 841–842, New York, NY, USA. ACM.

[Strohmaier et al., 2012] Strohmaier, M., Helic, D., Benz, D., Körner, C., and Kern, R. (2012). Evaluation of folksonomy induction algorithms. *ACM Trans. Intell. Syst. Technol.*, 3(4):74:1–74:22.

[Szabo and Huberman, 2010] Szabo, G. and Huberman, B. A. (2010). Predicting the popularity of online content. *Commun. ACM*, 53(8):80–88.

[Tsur and Rappoport, 2012] Tsur, O. and Rappoport, A. (2012). What's in a hashtag? content based prediction of the spread of ideas in microblogging communities. In *Proceedings of the fifth ACM international conference on Web search and data mining*, WSDM '12, pages 643–652, New York, NY, USA. ACM.

[Vosecky et al., 2013] Vosecky, J., Jiang, D., and Ng, W. (2013). Limosa: a system for geographic user interest analysis in twitter. In *Proceedings of the 16th International Conference on Extending Database Technology*, EDBT '13, pages 709–712, New York, NY, USA. ACM.

[Wagner et al., 2013a] Wagner, C., Asur, S., and Hailpern, J. (2013a). Religious politicians and creative photographers: Automatic user categorization in twitter. In *we will see.*

[Wagner et al., 2012a] Wagner, C., Liao, V., Pirolli, P., Nelson, L., and Strohmaier, M. (2012a). It's not in their tweets: Modeling topical expertise of twitter users. In *Proceedings ASE/IEEE International Conference on Social Computing (SocialCom2012).*

[Wagner et al., 2012b] Wagner, C., Rowe, M., Strohmaier, M., and Alani, H. (2012b). Ignorance isn't bliss: An empirical analysis of attention patterns in online communities. In *Proceedings ASE/IEEE International Conference on Social Computing (SocialCom2012).*

[Wagner et al., 2012c] Wagner, C., Rowea, M., Strohmaier, M., and Alani, H. (2012c). What catches your attention? an empirical study of attention patterns in community forums. In *The International AAAI Conference on Weblogs and Social Media (ICWSM2012).*

[Wagner et al., 2013b] Wagner, C., Singer, P., Posch, L., and Strohmaier, M. (2013b). The wisdom of the audience: An empirical study of social semantics in twitter streams. In *Porceedings of the European Semantic Web Conference (ESWC2013).*

[Wagner et al., 2013c] Wagner, C., Singer, P., Strohmaier, M., and Huberman, B. (2013c). Semantic stability in people and content tagging streams. under review.

[Wagner and Strohmaier, 2010] Wagner, C. and Strohmaier, M. (2010). The wisdom in tweetonomies: Acquiring latent conceptual structures from social awareness streams. In *Proc. of the Semantic Search 2010 Workshop (SemSearch2010).*

[Wagner et al., 2011] Wagner, C., Strohmaier, M., and He, Y. (2011). Pragmatic metadata matters: How data about the usage of data effects semantic user models. In *Social Data on the Web Workshop co-located with 10th International Semantic Web Conference (ISWC2011).*

[Wang and Huberman, 2012] Wang, C. and Huberman, B. (2012). How random are online social interactions? *Scientific Reports*, 2.

[Weng et al., 2010a] Weng, J., Lim, E., Jiang, J., and He, Q. (2010a). Twitterrank: finding topic-sensitive influential twitterers. In *Proceedings of the third ACM international conference on Web search and data mining*, pages 261–270. ACM.

[Weng et al., 2010b] Weng, J., peng Lim, E., Jiang, J., and He, Q. (2010b). Twitterrank: Finding topic-sensitive influential twitterers. In

Third ACM International Conference on Web Search and Data Mining (WSDM 2010).

[White et al., 2009] White, R. W., Dumais, S. T., and Teevan, J. (2009). Characterizing the influence of domain expertise on web search behavior. In *Proceedings of the Second ACM International Conference on Web Search and Data Mining*, WSDM '09, pages 132–141, New York, NY, USA. ACM.

[Yang et al., 2012] Yang, L., Sun, T., Zhang, M., and Mei, Q. (2012). We know what @you #tag: does the dual role affect hashtag adoption? In *Proceedings of the 21st international conference on World Wide Web*, WWW '12, pages 261–270, New York, NY, USA. ACM.

[Yule, 1925] Yule, G. U. (1925). A Mathematical Theory of Evolution, Based on the Conclusions of Dr. J. C. Willis, F.R.S. 213(402-410):21–87.

[Zipf, 1949] Zipf, G. K. (1949). *Human behavior and the principle of least effort.* Addison-Wesley Press.