Lisa Posch

Characterizing and Predicting Semantic Hashtag Categories via Usage Patterns on Twitter

Master's Thesis

Graz University of Technology

Knowledge Technologies Institute
Head: Univ.-Prof. Dipl.-Inf. Dr. Stefanie Lindstaedt

Supervisor: Prof. Dipl.-Ing. Dr.techn. Markus Strohmaier
Advisor: Dipl.-Ing. Claudia Wagner

Graz, September 2013

Lisa Posch

Charakterisierung und Vorhersage semantischer Hashtag-Kategorien mittels Nutzungsmustern auf Twitter

Masterarbeit

Technische Universität Graz

Institut für Wissenstechnologien
Vorstand: Univ.-Prof. Dipl.-Inf. Dr. Stefanie Lindstaedt

Begutachter: Prof. Dipl.-Ing. Dr.techn. Markus Strohmaier
Betreuerin: Dipl.-Ing. Claudia Wagner

Graz, September 2013

This document is set in Palatino, compiled with pdfLATEX2e.

The LATEX template from Karl Voit is based on KOMA script and can be found online: https://github.com/novoid/LaTeX-KOMA-template

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.

Graz, ______________________ ______________________

Date Signature

Eidesstattliche Erklärung[1]

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 entnommenen Stellen als solche kenntlich gemacht habe.

Graz, am ______________________ ______________________

Datum Unterschrift

[1]Beschluss der Curricula-Kommission für Bachelor-, Master- und Diplomstudien vom 10.11.2008; Genehmigung des Senates am 1.12.2008

Abstract

This thesis sets out to explore whether data about the usage of hashtags on Twitter contains information about their semantics. Towards that end, pragmatic measures were designed to capture the usage patterns of different hashtag streams. These measures describe different social and message-based structures of hashtag streams at specific points in time, and additionally capture how these structures change over time.

Initial statistical hypothesis tests were performed in order to quantify the association between usage patterns and semantic categories of hashtags. To assess the utility of the pragmatic measures for semantic analysis of hashtags, various hashtag stream classification experiments were conducted and their utility was compared with the utility of lexical features.

The results of these experiments indicate that semantic categories of hashtag streams do indeed differ with respect to their usage patterns, and that the pragmatic measures contain valuable information which can be used to distinguish between different semantic categories. Furthermore, the pragmatic measures can be successfully used for classifying hashtags into their semantic categories. Although pragmatic measures do not outperform lexical measures in the experiments presented in this work, pragmatic measures are important and relevant for settings in which textual information might be sparse or absent (e.g., in social video streams).

The results presented in this thesis 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.

Zusammenfassung

Das Ziel dieser Masterarbeit ist es, zu untersuchen, ob Daten über die Nutzung von Hashtags auf Twitter Informationen über ihre Semantik enthalten. Zu diesem Zweck wurden pragmatische Maße definiert, die die Nutzungsmuster von Hashtagstreams erfassen. Diese Maße beschreiben die verschiedenen sozialen und nachrichtenbasierten Strukturen zu bestimmten Zeitpunkten und erfassen außerdem die Veränderung dieser Strukturen über die Zeit.

Um den Zusammenhang zwischen Hashtagstreams und semantischen Kategorien zu quantifizieren, wurden statistische Hypothesentests durchgeführt. Zusätzlich wurden verschiedene Hashtagstream-Klassifikationsexperimente durchgeführt um die Nützlichkeit der pragmatischen Maße für die semantische Analyse von Hashtags zu beurteilen. In diesen Experimenten wurde die Nützlichkeit der pragmatischen Maße mit der Nützlichkeit von lexikalischen Maßen verglichen.

Die Ergebnisse der Experimente zeigen, dass die Nutzungsmuster semantischer Hashtag-Kategorien signifikant unterschiedlich sind, und dass pragmatische Maße wertvolle Informationen enthalten, mit denen zwischen semantische Kategorien werden kann. Weiters zeigen die Ergebnisse, dass pragmatische Maße erfolgreich dazu verwendet werden können, Hashtags in ihre semantischen Kategorien zu klassifizieren. Obwohl pragmatische Maße für die Klassifikation von Hashtagstreams nicht besser geeignet sind als lexikalische Maße, sind pragmatische Maße wichtig und relevant für Situationen, in denen wenig oder keine textuelle Information vorhanden ist (z.B. in Videostreams).

Diese Arbeit ist relevant für Forscher, die an der semantischen Analyse von Hashtags in sozialen Streams interessiert sind.

Acknowledgements

The research work presented in this thesis has been carried out at the Knowledge Technologies Institute at Graz University of Technology.

First of all I want to thank my supervisor Prof. Dr. Markus Strohmaier for his invaluable advice and inspiring discussions during the course of this thesis. This work would not have been possible without his essential guidance and support.

I want to thank my academic advisor Dipl.Ing. Claudia Wagner for providing frequent feedback, advice and motivation, and for encouraging and supporting me during my thesis.

I would like to express my gratitude to Dipl.Ing. Philipp Singer for the help, advice and feedback he has given me, and for his precious cooperation in the research papers which were published during the course of this work.

Also, I want to thank the Knowledge Technologies Institute and the Know-Center for providing facilities during the course of this thesis.

I would like to thank my colleagues at the Knowledge Technologies Institute for providing feedback and motivation, for challenging and enriching my ideas, and for hosting gaming nights.

I owe my deepest gratitude to my parents Hermine and Karl who provided constant support, encouragement and motivation, and who gave me the opportunity to carry out my studies.

Finally, I am very grateful for all the encouragement, feedback and moral support I received from my friends.

Thank you!

Contents

Abstract v

Zusammenfassung vi

Acknowledgements vii

1. Introduction 1
- 1.1. Motivation . . . 2
- 1.2. Research Questions and Contributions . . . 3
- 1.3. Thesis Outline . . . 5
- 1.4. Semantics and Pragmatics on the Web . . . 6
  - 1.4.1. Semantics . . . 6
  - 1.4.2. Pragmatics . . . 9
- 1.5. Twitter . . . 11
  - 1.5.1. Tweets . . . 13
  - 1.5.2. Social Relations . . . 14
  - 1.5.3. Trending Topics . . . 14
  - 1.5.4. Hashtags . . . 15
  - 1.5.5. Slashtags . . . 18

2. Related Work 19
- 2.1. Emergent Semantics . . . 19
  - 2.1.1. Tagging Vocabulary . . . 20
  - 2.1.2. Extracting Tag Networks . . . 21
  - 2.1.3. Semantic Grounding and Classification of Tags . . . 24
- 2.2. Research on Twitter . . . 27
  - 2.2.1. Characteristics of Twitter . . . 28
  - 2.2.2. Classification of Twitter Content . . . 33
  - 2.2.3. Further Research on Twitter . . . 34

2.3. Semantics and Pragmatics of Hashtags 38

3. Dataset **47**

3.1. Concepts 47

3.1.1. Hashtag Streams 47

3.1.2. Social Connections 48

3.1.3. User Ranking and Audience 49

3.2. Structure of the Dataset 50

3.3. Data Collection 51

3.3.1. Hashtag Selection 51

3.3.2. Overview of the Process 52

3.3.3. Implementation 54

4. Methodology **57**

4.1. Static Pragmatic Measures 57

4.1.1. Entropy Measures 58

4.1.2. Overlap Measures 60

4.1.3. Coverage Measures 61

4.2. Dynamic Pragmatic Measures 63

4.2.1. Divergence Measures 63

4.3. Lexical Measures 65

4.3.1. Term Frequency 66

4.4. Time Frames 66

5. Experiment 1: Usage Patterns of Hashtag Categories **68**

5.1. Experimental Setup 68

5.2. Results 72

5.2.1. Comparison of Categories 73

5.2.2. Comparison of Measures 75

6. Experiment 2: Classification of Hashtag Streams **79**

6.1. Experimental Setup 79

6.1.1. Classification Models 81

6.1.2. Evaluation 83

6.1.3. Feature Ranking 84

6.2. Results 84

6.2.1. Classification Results 84

6.2.2. Results by Category 85
6.2.3. Feature Ranking 86

7. Discussion of Results 89

8. Conclusions and Implications 91
8.1. Threats to Validity 91
8.2. Limitations and Future Work 93

A. Hashtag Category Descriptions 96

B. Feature Distributions 98

C. Precision and Recall by Category 103

Bibliography 105

List of Figures

1.1. Example of a tweet containing a hashtag 2
1.2. Initial Web 2.0 concept 8
1.3. First sketch of Twitter 12
1.4. Sample of worldwide trending topics 14
1.5. Messina's tweet proposing the use of hashtags 15
1.6. First use-case for hashtags 16
1.7. Hashtag search query result for the hashtag #nevertrust. . . 17

3.1. Followers, followees and friends 48
3.2. Audience Ranking 50
3.3. Timeline of the data collection process 53
3.4. Babel Library 56

4.1. Illustration of the time frames 67

5.1. Feature distribution for $t_{0\to1}$ 77
5.2. Feature distribution for $t_{1\to2}$ 78

6.1. Classification results 85
6.2. Classification results by category 87

8.1. Multimedia tweet example 94

B.1. Author Entropy 98
B.2. Follower Entropy 98
B.3. Followee Entropy 99
B.4. Friend Entropy 99
B.5. Author-Follower Overlap 99
B.6. Author-Followee Overlap 99
B.7. Author-Friend Overlap 100

B.8. Informational Coverage 100
B.9. Conversational Coverage 100
B.10. Retweet Coverage 100
B.11. Retweet Coverage 101
B.12. Author Dynamics 101
B.13. Follower Dynamics 101
B.14. Followee Dynamics 101
B.15. Friend Dynamics 102

C.1. Precision by category 103
C.2. Recall by category 104

List of Tables

3.1. Top 10 authors, followers, followees and friends of the hashtag #ebay . . . 49
3.2. Description of the complete dataset . . . 51
3.3. Randomly selected hashtags per category (ordered alphabetically) . . . 52
3.4. Trending hashtags streams retrieved at t_1 and t_2 . . . 53

5.1. Statistically significant differences in both time frames . . . 72
5.2. Statistically significant differences in $t_{0 \to 1}$ . . . 73
5.3. Statistically significant differences in $t_{1 \to 2}$ . . . 74

6.1. Classification Models . . . 82
6.2. Top features for two different datasets ranked via Information Gain . . . 86

A.1. Hashtag category descriptions . . . 96

List of Abbreviations

Pragmatic Measures	
norm_entropy_author	Author Entropy
norm_entropy_follower	Follower Entropy
norm_entropy_followee	Followee Entropy
norm_entropy_friend	Friend Entropy
overlap_authorfollower	Author-Follower Overlap
overlap_authorfollowee	Author-Followee Overlap
overlap_authorfriend	Author-Friend Overlap
informational	Informational Coverage
conversational	Conversational Coverage
retweet	Retweet Coverage
hashtag	Hashtag Coverage
kl_authors	Kullback-Leibler divergence of authors
kl_followers	Kullback-Leibler divergence of followers
kl_followees	Kullback-Leibler divergence of follwees
kl_friends	Kullback-Leibler divergence of friends

1. Introduction

Micro-blogging and online social networks have become extremely popular over the recent years. The wealth of information that is created collaboratively by users in social networks is often publicly accessible and has therefore attracted researchers from a wide range of disciplines. Online platforms such as Twitter have given researchers unprecedented access to large datasets which contain aggregated information created by millions of people worldwide.

Drawing on the abundance of information contained in social networks and social tagging systems, researchers have attempted to capture the semantics which emerge from this collaboratively generated data. Social tagging systems have been investigated extensively, and recently, also Twitter has been studied in terms of its emergent semantics. The micro-blogging platform is interesting for researchers, as a large part of its message stream is accessible via public APIs [Twi13a]. Few of these studies, however, have focused on the link between usage patterns and semantics. The work of this thesis presents a first step in exploring whether pragmatic features of social streams may be useful for revealing information about their semantics.

This chapter is structured as follows. Section 1.1 describes the motivation for the work presented in this thesis, Section 1.2 introduces the research questions which were addressed in this work, and Section 1.3 gives an overview of the structure of this thesis. In Section 1.4, the concepts of semantics and pragmatics are explained. Finally, Section 1.5 gives a brief overview of the Twitter platform and its functionality.

1.1. Motivation

In the past years, social media applications such as Twitter, Facebook, Google+ or Flickr have emerged rapidly and gained a tremendous amount of participants. Social media, constituting natural platforms for the spread of ideas and thoughts [TR12] have demonstrated exponential growth, making them a very popular activity on the Internet [WNHT09]. The large amounts of data which are created by users of such social media applications have inspired researchers to investigate different aspects of these large-scale social systems. One of the aspects of social tagging sites and micro-blogging systems which has been investigated is the semantics which emerge from collaboratively generated tagging structures.

This thesis investigates if and to what extent pragmatic features of social awareness streams, such as hashtag streams on Twitter, reveal information about the semantics of the streams' topics. Hashtags are strings of characters preceded by the hash (#) character and they are used on platforms like Twitter as descriptive labels or to build communities around particular topics [TR12]. 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. Figure 1.1 shows an example of a tweet containing a hashtag.

Figure 1.1.: Example of a tweet containing a hashtag

A simplistic view on Wittgenstein's work [Wit53] suggests that *meaning is use* (also see Section 1.4.2). Wittgenstein postulates 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. The work presented in this thesis builds on this observation, but focuses on the pragmatics of a word (i.e., how a word, or in this case a hashtag, is used by a large group of users) – rather than its narrow, lexical context.

Pragmatic features have the advantage of being language independent and they 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. The pragmatic aspects that were examined in this work do not focus on single individuals nor on identifying socio-cultural norms, but rather investigate the structural patterns of the social connections which a collection of individuals (the set of users employing a certain hashtag) exhibits, as well as the structural context in which a hashtag occurs.

1.2. Research Questions and Contributions

The aim of this work is not only to discover the idiosyncrasies of hashtag usage within one semantic category but also in the deltas between different semantic categories. The experiments are designed to explore to what extent pragmatic properties of hashtag streams, which capture how a large group of users uses a hashtag, can be used to gauge the semantic category of a hashtag.

Specifically, the following research questions are addressed:

1. Do different semantic categories of hashtags reveal substantially different usage patterns?
2. To what extent do pragmatic and lexical properties of hashtags help to predict the semantic category of a hashtag?

To address these research questions, an empirical study was conducted 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 [RMK11] and have shown to be useful for grouping hashtags. From each of the eight categories, ten sample hashtags were selected at random and temporal snapshots of messages containing at least one of these hashtags at three different points in time were collected. To quantify how hashtags are used over time, the set of pragmatic stream measures which were introduced in [WS10] were extended and applied to the hashtag streams in the 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.

To answer the first research question, statistical standard tests were used, which allow to quantify the association between pragmatic characteristics of hashtag streams and their semantic categories. To tackle the second research question, lexical features were computed using a standard bag-of-words model with term frequency (TF). Then, several classification models were trained with lexical features only, pragmatic features only and a combination of both. The performance of different classification models was compared 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 the classification models, a control dataset was constructed by randomly shuffling the category labels of the hashtag streams, using the average performance of 100 repetitions. That means that the original relationship between the pragmatic properties and the semantic categories of hashtags was destroyed.

The main contributions of this thesis are the following:

1. The compilation of a dataset consisting of three time frames (encompassing a total time span of three months). The data was collected from Twitter and contains hashtag streams, the social connections of the hashtag streams' authors as well as tweets posted by the streams' audiences.
2. The creation of a set of pragmatic measures which describe the usage patterns of hashtag streams.
3. The demonstration that different semantic categories of hashtags do indeed show different usage patterns, that certain measures are better suitable for distinction than other measures and that some categories show more distinct usage patterns than others.

4. The demonstration that pragmatic measures can be used to gauge the semantic category of a hashtag and that some categories are better predictable than others.

The results of this work 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).

1.3. Thesis Outline

This thesis is organized in the following way: Chapter 2 gives an overview of related research on analyzing the semantics of tags in social bookmarking systems, research on Twitter in general, and research on the semantics and pragmatics of hashtags. In Chapter 3, the dataset which was used for the experiments and comprised during the course of this work is described. Chapter 4 presents the different metrics which were developed to characterize usage patterns of hashtag streams. Chapter 5 describes the experimental setup, evaluation approach and results obtained from the first experiment. The first experiment targets the first research question, exploring to which extent usage patterns of hashtag streams in different semantic categories are indeed significantly different. The second experiment is described in Chapter 6. It addresses the second research question and explores whether social and structural properties of hashtag streams may be used to identify the semantic category of hashtags. The results of the experiments are further discussed in Chapter 7 and finally, Chapter 8 concludes this thesis.

Parts of this thesis have already been published in [PWSS13] and in [WSPS13]:

- Lisa Posch, Claudia Wagner, Philipp Singer, and Markus Strohmaier. Meaning as collective use: predicting semantic hashtag categories on twitter. In *Proceedings of the 22nd International Conference on World Wide Web Companion*, pages 621–628. International World Wide Web Conferences Steering Committee, 2013.

- Claudia Wagner, Philipp Singer, Lisa Posch, and Markus Strohmaier. The wisdom of the audience: An empirical study of social semantics in twitter streams. In *The Semantic Web: Semantics and Big Data*, pages 502–516. Springer, 2013.

My main responsibilities in [PWSS13] were the data collection and creation of the dataset (see Chapter 3), the calculation of the pragmatic measures (see Chapter 4), as well as the design and conduction of *Experiment 1: Usage Patterns of Hashtag Categories* (see Chapter 5). The design of the pragmatic features and *Experiment 2: Classification of Hashtag Streams* (see Chapter 6) were conducted in cooperation with Claudia Wagner and Philipp Singer. In *Experiment 2*, I was involved in the design of the experimental setup, the conduction of the classification, and responsible for the measure ranking via information gain. In [WSPS13], I was mainly responsible for the data collection and creation of the dataset (including the audience ranking) and the calculation of the pragmatic measures.

1.4. Semantics and Pragmatics on the Web

This section introduces the general concepts underlying this work. It gives an overview of how the concepts of social and emergent semantics were developed, introduces the linguistic field of pragmatics, and briefly describes Wittgenstein's use-theory of meaning, which constitutes a philosophical basis for the research questions of this work.

1.4.1. Semantics

In its early days, the Internet resembled an extensive library and was used mainly for accessing information. This role started to change at the beginning of this century, and the change was amplified by the rise of social networks and micro-blogging platforms. These services allowed members to post short pieces of digital content, to be accessed by the public or by certain groups of people which the authors of the content defined. People started sharing information about themselves and social relations began to play a bigger role. On different social networking sites, people became connected with each other, and large amounts of social data were being published. With the burst of the dotcom bubble in 2001, the concept of Web 2.0. emerged. Abandoning the clear distinction between information providers and information consumers which was present in the Web 1.0, the Web 2.0 had a focus on user-driven design and social participation. In an initial brainstorming session for the concept of Web 2.0, O'Reilly and MediaLive International formulated their concept of "Web 2.0" by example, shown in figure 1.2. The concept was intended as one with a gravitational core, rather than having a hard boundary [Ore07]. At the same time, the concept of the Semantic Web emerged, presenting a vision of a Web of machine-processable content that could be maintained efficiently, enabling logical reasoning and giving information a well-defined meaning [Mik07b].

In the early Web, most URIs identified a web page by allowing direct access to it. With the changing role of the Web, the question what URIs identify became less trivial. This new question of how URIs get their meaning on the Web produced opposing views: On the one side advocators, like Berners-Lee, argued that URIs always identified one thing, while proponents of the other side, like Hayes [HH08], stated that URIs were always ambiguous due to the fact that different models interpreted the formal semantics the same URI in different ways. A third position, which arose with the Web 2.0, was that the meaning of names was defined by the way in which they were used. This position is sometimes called *social semantics*, having its philosophical roots in the later works of Wittgenstein [HT09]. This position states that *"meaning in language exists due to a form-of-life, and so names have a sense as a mechanism for the co-ordination of actions among multiple agents"* [Hal13].

Besides providing an ease of publishing user-generated content, a wide range of online services also started to allow users to tag the URIs with freely choosable keywords, giving rise to a new phenomenon: social tagging. The first service to adopt this idea was del.icio.us, a social bookmarking site [Mik07a], but since then, a large variety of resource types have become taggable, including photos on Flickr, tweets on Twitter, audio tracks on last.fm and videos on YouTube. Social tagging was defined as the practice of "publicly labeling or categorizing resources in a shared,

Web 1.0		Web 2.0
DoubleClick	-->	Google AdSense
Ofoto	-->	Flickr
Akamai	-->	BitTorrent
mp3.com	-->	Napster
Britannica Online	-->	Wikipedia
personal websites	-->	blogging
evite	-->	upcoming.org and EVDB
domain name speculation	-->	search engine optimization
page views	-->	cost per click
screen scraping	-->	web services
publishing	-->	participation
content management systems	-->	wikis
directories (taxonomy)	-->	tagging ("folksonomy")
stickiness	-->	syndication

Figure 1.2.: Initial Web 2.0 concept [Ore07]

on-line environment" [Tra09], "allowing anyone – especially consumers – to freely attach keywords or tags to content" [GH06]. In social tagging systems, users create resources and annotate them with freely chosen natural language words, called tags. These tags are often public and shared with other users [SKK10] [KBS+10] [CKJ11]. The success of tagging has been attributed to the fact that it acts as a balance between the individual and the community, having a low cost of participation and benefiting both the individual and the community [NM08a]. Aberer et al. [ACMO+04] stated that information sharing will be IT's primary goal for the 21st century (in contrast to information processing).

While tags provide metadata to the content, they are a flat and unstructured collection of keywords [NM08a], describing various aspects of the content of resources, such as locations, dates, people [OSvZ09], contextual information about the resources, subjective qualities and opinions about them, or organizational aspects [CKJ11]. In contrast to Semantic Web ontologies, which are an "explicit specification of a conceptualization" [G+95], these collaboratively generated collections of tags do not exhibit a predefined and hierarchical structure [Hal13]. However, such a structure can be constructed from them [SHB+12]. Due to ontologies being structured "a-priori" agreements on concepts, they are not sufficient in ad-hoc and dynamic situations as they occur in social tagging systems

[ACMO+04]. As a community evolves, the knowledge coded in an ontology will be invalidated with time – a problem called "ontology drift" [Mik07a].

In an attempt to address these problems that occurred with ontologies, the concept of *emergent semantics* was created. In the work of Aberer et al. [ACMO+04], principles for emergent semantics were defined in a collaborative effort. According to them, *emergent semantics* refers to a set of principles and techniques for analyzing the semantic interoperability of decentralized semantic structures which are constructed incrementally in widely distributed information systems. Another effort to address the problem of ontology drift included incorporating the social dimension into the traditional bipartite model of ontologies (which consists of concepts and instances). Mika [Mik07a] demonstrated in his work that a tripartite model of actors, concepts and instances could show how community-based semantics emerged, through a process of graph transformation.

The emergent information structures of collaborative tagging systems were named *folksonomies*, a neologism resulting from the words "folk" and "taxonomy" (also see Section 2.1.2) [VW07] [MCM+09]. Folksonomies have also been described as the "core data structure of collaborative tagging systems" [KBS+10], the "collective assemblage of tags assigned by many users" [Tra09], or the "collaboratively generated annotations of web pages" [WS10].

Since the rise of social tagging, a large amount of studies have been conducted on investigating how and to what extent semantics emerge from folksonomies and whether the data produced in social tagging systems can be used to engineer light-weight ontologies (e.g., [Mik07a], [BKH+11]). Some researchers see the usage of folksonomy-derived information, in combination with the Semantic Web, as the transition towards the Web 3.0 [MCM+09].

1.4.2. Pragmatics

Pragmatics is considered to be one of the five structural components of language (the other four being phonology, semantics, syntax and morphology) [BSGHP06]. The field of pragmatics is commonly defined as "the study of how context and situation affect meaning," "the general study of how context affects linguistic interpretation" [FRH11], or "the study of the use of context to make inferences about meaning" [Fas90]. It studies language's relation to contextual background features [Cut02], which may be linguistic (such as what was previously said or written) or knowledge of the world (such as the speech situation) [FRH11]. Pragmatics is primarily concerned with the interplay between context, text and function by analyzing the parts of meaning which can be explained by knowledge of the physical and social world and other contextual factors [Cut02].

Bach [Bac99] differentiated between semantics and pragmatics simply by saying: "Narrow context is semantic, wide context pragmatic." The wide context of pragmatics includes facts about the speaker, the utterance and the people she is speaking with, the speaker's beliefs, the beliefs the speaker shares with her audience, and the speakers' intentions [oP12]. Pragmatics views language as related to the user. This view of language in pragmatics is very similar to the view of language Wittgenstein advocated in his later works, most notably in *Philosophical Investigations* [Wit10].

Meaning is Use

In Wittgenstein's earlier works, the most famous being the *Tractatus Logico-Philosophicus* [Wit21], he proposed a picture-theory of meaning. The central doctrine of this theory was that meaning is a mirroring between language and the facts which constitute the world. The emphasis of the picture-theory of meaning in the Tractatus, which was strongly logicist, was the construction of a symbolism that would allow to generate accurate meaning. In his later works, however, Wittgenstein replaced the picture-theory with a *"pragmatic description of intersubjective communicative practice"* [Liv04]. *Philosophical Investigations*, published posthumously in 1953, constitutes a repudiation of his earlier position in the Tractatus. It deals with pragmatic aspects of language and focuses on the use of words [Reh10]. This "use-theory" of meaning is what Wittgenstein became famous for [Con98] [Liv04].

The core statement of the "use-theory" of meaning is that the meaning of sentences is defined by how they are used, and not by the objects which the words refer to. Wittgenstein opines that language acquires its meaning by its use, that it is part of a "form of life". His later works state that language is a public activity – that a sentence only has its meaning in the circumstances in which it is actually used [Reh10]. In *Philosophical Investigations*, Wittgenstein wrote: *"For a large class of cases – though not for all – in which we employ the word "meaning" it can be defined thus: the meaning of a word is its use in the language"* [Wit10]. What he meant by this is that the question "What do the words mean?" cannot be posed properly without considering the contexts of use. Wittgenstein called these contexts of significant use *"language-games"* [Con98].

In the context of semantics on the Web, it has been argued that tagging can be seen as adding words to the language game, and that tagging, as well as the use of search engines, can be viewed as a form of life – the form of life of the Web [Hal13] [HT09].

1.5. Twitter

Launched in 2006, Twitter[1] soon became one of the most popular micro-blogging platforms. Twitter enables real-time discussion of current topics and describes itself as a "real-time information network that connects you to the latest stories, ideas, opinions and news about what you find interesting" [Twi13b], and "a public forum where anyone can read, write and share messages" [Twi13c]. The platform is available in over 33 languages [Twi13d].

Originally, the service was designed as a "mobile status update service" [Twi09]. Figure 1.3 shows the first sketch of the service, drawn by Twitter co-founder Jack Dorsey in March 2006 [Twi12]. During the first years of Twitter, users were encouraged to post short answers to the question "What are you doing?", in a maximum of 140 characters [BB11]. In November 2009, this question was changed to "What's happening?" as the Twitter

[1] http://twitter.com

developers noticed that the original question was completely ignored. Instead, Twitter users were seemingly using the service for both asking and answering this more immediate question [Twi09]. Now, Twitter simply prompts users to "Compose new Tweet," indicating an even more diverse use for the platform. A large-scale study of the service conducted by Kwak et al. [KLPM10] showed that Twitter is used as a hybrid between communication media and an online social network.

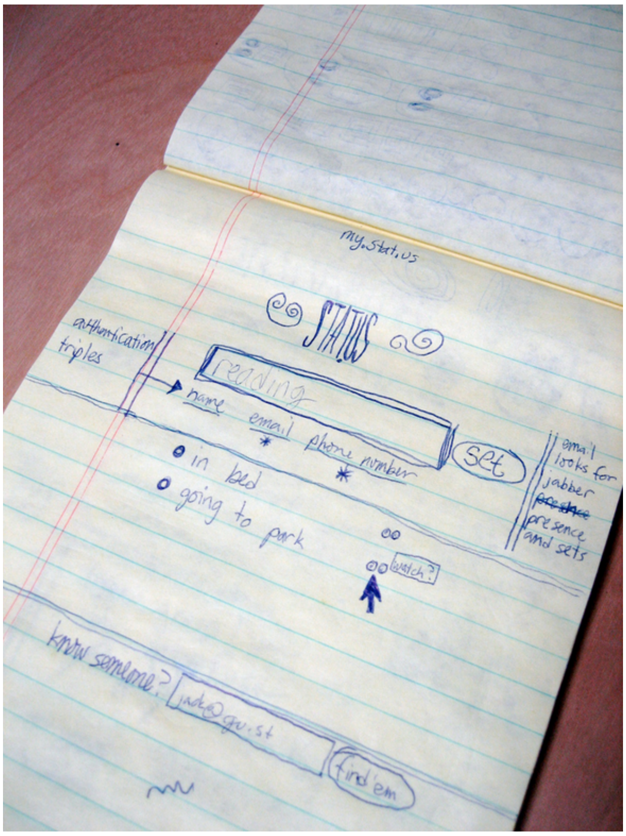


Figure 1.3.: First sketch of Twitter, drawn by Twitter cofounder Jack Dorsey in March 2006 [Twi12].

Twitter experienced an enormous growth over the years, reaching half a billion accounts in June 2012 – more than 140 million accounts in the United States alone [Sem12]. The platform grew 1,400% in a single year (from 2009 to 2010) [Twi10]. In 2007, 500,000 messages were posted per quarter. This number grew to four billion messages per quarter in 2010 [AK10], and in 2012, 340 million tweets were created each day [Twi12]. At the moment, Twitter reports 200 million active users and an average of 400 million tweets posted every day [Twi13d].

1.5.1. Tweets

Messages on Twitter are called *tweets*. These posts have a maximum length of 140 characters, but may include videos and photos. Tweets often exhibit an informal language, and due to the length restriction they often contain abbreviations. Special characters, such as # or @, may be used in tweets to indicate meta information. Meta information in tweets includes hashtags, slashtags, mentions, replies and whether a tweet is a retweet.

- **Retweets:** Retweets are copies of tweets posted by another user. They start with "RT @<username>," where <username> is the user name of the original author of the tweet. Retweets can be exact copies without modifications or copies with added comments. It is also possible to create a retweet via the functionality of the Twitter web interface, without manually adding "RT @username".
- **Mentions:** To mention a user, @<username> (where <username> is the username of the user to be mentioned) can be included anywhere in a tweet. Twitter supports mentions in their web interface: All mentions of a user are displayed in the "Mentions" tab.
- **Replies:** It is possible to post a reply to a certain user by starting a tweet with @<username>, where <username> represents the username of the user who is replied to. Replies are also considered mentions, and Twitter's web interface supports this functionality through the "Reply" button.
- **Hashtags:** Hashtags are strings of characters preceded by the hash (#) symbol, used for different purposes such as indicating the topic of a tweet, for building communities or for representing memes. The concept of hashtags is described in detail in Section 1.5.4.
- **Slashtags:** Slashtags are keywords preceded by the slash (/) symbol. There are three types of slashtags, which are described in Section 1.5.5.

1.5.2. Social Relations

In contrast to most social networking sites like Facebook or MySpace, the social network on Twitter is not bidirectional. The social network on Twitter is defined by *following* relations, which are unidirectional: A user on Twitter may follow other users without requiring any confirmation from these users. When a user A follows another user B, user A receives user B's tweets – unless user B's profile is protected, in which case user A needs user B's approval to see the tweets.

1.5.3. Trending Topics

Twitter displays the top ten currently trending topics (keywords and hashtags) on its web interface, a feature which was introduced in summer 2008 [Twi08]. The trending topics are tailored to the viewing user's Twitter account by default, based on her location and who she follows. However, these settings can be changed to display the trending topics of specific locations and regions, or to display worldwide trending topics. Figure 1.4 shows a sample of worldwide trending topics.

Figure 1.4.: Sample of worldwide trending topics, captured on June 11th, 2013.

1.5.4. Hashtags

A hashtag is a sequence of alphanumeric characters, preceded by the hash symbol (#). In 2007, Chris Messina [Mes07a] introduced the idea of creating a system of *"channel tags"* which make use of the hash symbol for *"improving contextualization, content filtering and exploratory serendipity within Twitter."* The idea of these channels was to enable Twitter users to follow and contribute to conversations on particular topics. Messina expressed the general idea as *"Every time someone uses a channel tag to mark a status, not only do we know something specific about that status, but others can eavesdrop on the context of it and then join in the channel and contribute as well"* [Mes07a]. Figure 1.5 shows Messina's tweet proposing the use of hashtags.

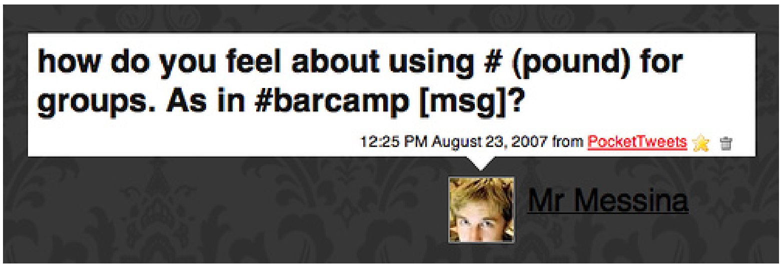


Figure 1.5.: Messina's tweet proposing the use of hashtags [Mes07a]

At first, the Twitter community was slow to pick up the concept of hashtags in their activities. Then, in October 2007, the bushfires in San Diego constituted a clear use case, and Messina urged people to use the hashtag #sandiegofire to mark tweets which referred to the bushfires [Mes07c]. In a blog post he states *"Hashtags become even more useful in a time of crisis or emergency as groups can rally around a common term to facilitate tracking"* [Mes07c]. Figure 1.6 shows the tweet which Messina used as an example to promote the usage of hashtags during the San Diego bushfires. Following this event, the practice of using hashtags became more and more widespread in the habits of the Twitter community [BB11]. A study by Laniado and Mika [LM10] found that in November 2009, 8.5% of the tweets contained at least one hashtag and 20% of the users were using hashtags.

Figure 1.6.: A tweet using the #sandiegofire hashtag, which represented a first use-case for the hashtag concept.

Hashtags became so popular that the word *hashtag* was voted as the *Word of the Year 2012* by the American Dialect Society [Soc13], in their 23rd annual words of the year vote. Ben Zimmer, chair of the New Words Committee of the American Dialect Society and executive producer of the Visual Thesaurus and Vocabulary.com, stated: *"This was the year when the hashtag became a ubiquitous phenomenon in online talk. In the Twittersphere and elsewhere, hashtags have created instant social trends, spreading bite-sized viral messages on topics ranging from politics to pop culture"* [Soc13].

Twitter users employ hashtags for explicitly marking the relevant topic of a tweet, but also for creating threads of conversation, for building communities and as a symbol of community membership [YSZM12] [LM10] [EMS+10] [SP11] [LSAA11]. Hashtags also frequently represent memes, which are short units of text that remain relatively intact while they travel through many online platforms [LBK09] [TR12]. Twitter supports this practice in its Web interface and API: By clicking on a hashtag in any tweet, a search query for this hashtag is conducted and the results are displayed. Figure 1.7 shows an exemplary query result of the hashtag #nevertrust – the query returns the most recent tweets which contain this hashtag. Terms which are not preceded by the # symbol but otherwise equal to the hashtag string, do not appear in the search results.

Hashtags underlie no regulation and are used to create "ad-hoc" channels [BB11]. They are created by users – to create or use a hashtag, all that is needed is inserting the # symbol in front of any string of alphanumeric characters. Hashtags show a high capacity for *cultural generativity* [Bur12], and their unique nature has attracted numerous researchers to investigate various aspects of them. While hashtags can be seen as lightweight social

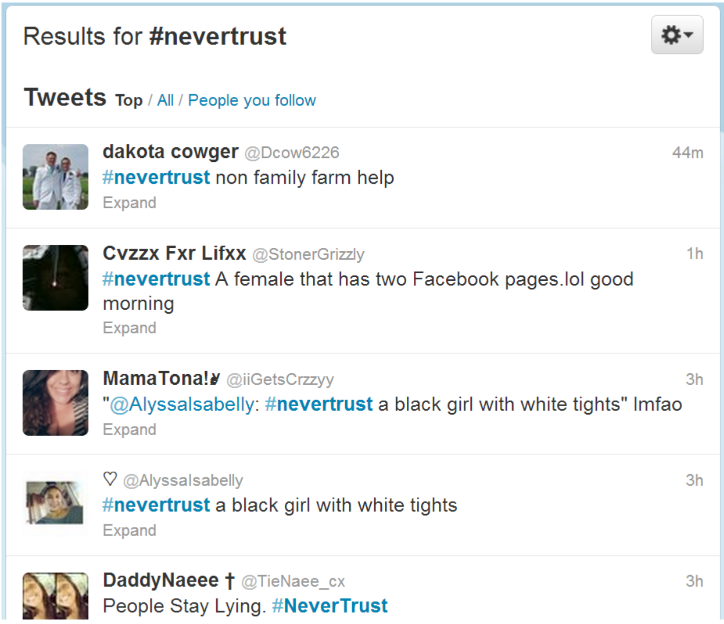


Figure 1.7.: Hashtag search query result for the hashtag #nevertrust.

annotations of the information streams that users consume [LGRC12] and are therefore practical for aggregating content, their meaning is often not clear at first sight – mostly due to the fact that many hashtags contain abbreviations. Furthermore, Laniado and Mika [LM10] found that not all hashtags aggregate around a topic and that some hashtags do not have a meaning. Relevant research on the pragmatics and semantics of hashtags is reviewed in detail in Section 2.3 of this thesis.

1.5.5. Slashtags

Slashtags are another form of microsyntax on Twitter. They use the slash symbol (/) as delimiter and were, like hashtags, also introduced by Messina [Mes07b]. Messina proposed three types of slashtags, in order to enable Twitter users to *"encode more meaning"* [Mes07b] in the length restriction of 140 characters. The three types of slashtags are:

- **/via:** This slashtag is used for giving credit to another user, without using a direct quote.
- **/cc:** This slashtag is an abbreviation of "carbon copy" and is used to indicate a user who the tweet is directed at.
- **/by:** This slashtag is used to *"attribute authorship for a longer-form piece"* [Mes07b], such as quoting a passage from a blog post.

2. Related Work

In the past decade, a considerable effort has been spent on studying the emergent semantics of tags (e.g., tags in social bookmarking systems). But also hashtags on Twitter, and the micro-blogging platform in general, have received attention from the research community. This chapter reviews the relevant literature of the past years regarding the semantics of tags, Twitter in general, and the semantics and pragmatics of hashtags on Twitter.

2.1. Emergent Semantics

Numerous studies have been conducted on the extraction of semantics from folksonomies. This section describes how researchers have leveraged the wealth of tags created in social tagging systems, capturing the emerging semantics with different approaches.

One of the aspects of social tagging systems which has been investigated is the structure and dynamics of such systems, and the vocabulary of tags which emerges over time. Section 2.1.1 briefly describes studies which have focused on this aspect. Researchers have explored to what extent semantics emerge from *folksonomies*, by investigating different algorithms for extracting tag networks and hierarchies from such systems (e.g., [BKH+11], [HGM06] or [Sch06]). Section 2.1.2 focuses on studies which have investigated this aspect of emergent semantics. A third aspect which has been investigated is to what extent tags (and the resources they annotate) can be semantically grounded and classified into predefined semantic categories. For example, Noll and Meinel studied the characteristics of tags [NM08a] and determined their usefulness for web page classification

[NM08b]. Section 2.1.3 focuses on semantic grounding and classification of tags.

2.1.1. Tagging Vocabulary

An early analysis of collaborative tagging systems was done by Golder and Huberman [GH06], who analyzed the structure and the dynamic aspects of the social bookmarking system del.icio.us. In their study, they investigated user activity, tag frequencies, kinds of tags used, bursts of popularity and relative proportions of tags within a given URL, finding patterns and regularities. The theory of social proof states that people act in ways they observe others acting because they come to believe that this is the correct behavior [CR93]. Therefore, according to this theory, it is likely that users adapt the tagging behavior of their community. While the results of Golder and Huberman [GH06] supported the social proof theory, they also concluded that imitation might not explain every aspect and that shared knowledge among taggers may also account for different taggers making the same choices. A study on the movie recommendation systems MovieLenses, conducted by Sen et al. [SLR+06], showed that the evolution of user's tagging vocabulary was influenced by the tagging vocabulary of her community influence and her personal tendency, also supporting the social proof theory.

The work of Halpin et al. [HRS07] showed that given sufficient active users, a stable distribution with a limited number of stable tags and a much larger long-tail of more idiosyncratic tags develops over time. A stable distribution indicates that there might be a user consensus around the categorization of information by tagging behaviors. They concluded that the agreement among users could overcome the problem of uncontrolled vocabulary of tags, including the three major problems of tagging pointed out by Golder and Huberman [GH06]: *polysomy*, *synonymy* and *basic level variation*.

2.1.2. Extracting Tag Networks

The neologism *folksonomy* was introduced by Vander Wal in 2004 [VW07], who initially used the word to describe the concept of a *"user-created bottom-up categorical structure development with an emergent thesaurus."* Folksonomies have since been defined in different ways, and the two following prevalent definitions are commonly used.

Plangprasopchok et al. [PLG10] defined folksonomies as *"common taxonom[ies]"* which are *"learned from social metadata created by many users."* This definition was adopted by, for example, Strohmaier et al. [SHB+12] who defined folksonomies as the output of algorithms which construct *"hierarchical structures from user-generated metadata"* [SHB+12], stating that the concept of folksonomies emerged to characterize the idea that *"latent hierarchical structures"* could be acquired from social tagging systems.

Vander Wal's definition of folksonomies [VW07] states that folksonomies are *"the result of personal free tagging of information and objects (anything with a URL) for one's own retrieval,"* and that they are *"created from the act of tagging by the person consuming the information"* [VW07]. Vander Wal stated that the three *"tenets"* of a folksonomy are the *tag*, the *object being tagged* and the *identity*. This definition coincides with the definition by Markines et al. [MCM+09] who defined folksonomies as the emergent information organizations of social tagging systems, with an *"inherent tripartite data structure"* consisting of users, tags and resources.

Several studies investigated the extraction of tag networks and hierarchies from tags of various social tagging systems. For example, Heymann and Garcia-Molina [HGM06] investigated the emergent semantics of a large amount of tags from the popular social bookmarking system del.icio.us and from CiteULike, an online service for organizing academic publications. They developed an algorithm which converts the corpus of tags into a navigable hierarchical taxonomy by calculating tag similarity and tag generality.

Similarly, Schmitz [Sch06] attempted to induce an ontology from tags on Flickr (an image and video hosting platform), with the aim of creating a system which combines social tagging and a taxonomy with fixed vocabulary. He proposed a probabilistic model which used seed ontologies, natural language processing techniques and domain knowledge for inducing a faceted ontology from the collected tags, intending to supplement the social tagging system. Such a combined system would have the advantage of preserving the flexibility of a social tagging system while also benefiting from the features of a faceted ontology, such as its suitability for searching and browsing. The model which resulted from his study was evaluated manually, resulting in 51% correct subsumption pairs.

The work of Benz et al. [BKH+11] measured tag abstractness in social tagging systems, in order to identify hierarchical relations between semantic concepts. Using a tagging dataset from del.icio.us and starting out with linguistic definitions of word abstractness, they applied folksonomy-based methods in order to measure the level of abstractness of given tags. The abstractness was measured in a number of different ways, including frequency-based measures, entropy-based measures, centrality measures and statistical subsumption. For evaluation, different large-scale ontologies and taxonomies were employed as grounding datasets for word generality: Yago, WordNet, DMOZ and the WikiTaxonomy.

The study found that the measures based on frequency, entropy and centrality correlated with the abstractness information of the grounding datasets, while the measure based on tag similarity graphs performed worst. Furthermore, the authors also found that popular tags are often more abstract.

Körner et al. [KBS+10] investigated how tagging usage patterns influence the quality of the emergent semantics. Using data from the social bookmarking system del.icio.us, they developed various measures which quantified a pragmatic differentiation of taggers, distinguishing between *describers* and *categorizers*.

The authors defined *describers* as users who employ a wide range of freely associated keywords, displaying a great verbosity in their tagging behavior, while *categorizers* were defined as users who employ a smaller variety of well-defined tags and use tagging as a replacement for hierarchical classification systems. The two types of taggers appeared to have different motivations for tagging: While *categorizers* tagged resources for later browsing, using subjective tags, *describers* tagged resources for later retrieval, using objective tags.

Several measures for tagging pragmatics were developed, classifying taggers into either *categorizers* or *describers*: vocabulary size, tag/resource ratio, average tags per post and orphan ratio (the percentage of tags that are assigned to very few resources). These measures were used to build folksonomy partitions through incrementally adding taggers of each class. The authors grounded the emerging semantics of each of the partitions by using semantically grounded tag relatedness measures.

The study found that 'verbose' taggers (*describers*) were more useful for the emergence of tag semantics than users who use a small set of tags (*categorizers*). Even a subset of only 40% of *describers* produced semantics that matched the semantic precision obtained from the whole dataset, in some cases even outperforming its precision. Furthermore, their results suggested a causal link between tagging pragmatics and the semantics which emerge from the folksonomy.

The study of Strohmaier et al. [SHB+12] evaluated three classes of state-of-the-art folksonomy induction algorithms in the context of five social tagging systems, using existing semantic evaluation techniques. Furthermore, their work used a pragmatic technique for folksonomy evaluation in terms of their usefulness for navigation.

In their work, the authors implemented three classes of folksonomy induction algorithms: affinity propagation, hierarchical k-means and generality in tag similarity networks. With tagging data from the social tagging systems BibSonomy, CiteULike, del.icio.us, Flickr and LastFm, which they used to create 20 folksonomies, they investigated the properties and characteristics of the different folksonomy induction algorithms, and identified limitations and challenges of folksonomy evaluation.

They evaluated the induction algorithms by reference-based semantic evaluation, choosing taxonomies derived from WordNet, Yago and the WikiTaxonomy as gold standards, as well as by pragmatic evaluation with greedy search (which evaluated the folksonomies from a navigation-oriented perspective). Both the reference-based evaluation and the pragmatic evaluation showed that those algorithms specifically developed to capture intuitions of social tagging systems outperformed traditional hierarchical clustering techniques.

2.1.3. Semantic Grounding and Classification of Tags

This section describes studies which investigated to what extent tags can be semantically grounded and classified into semantic categories. Similar to the work presented in this thesis, Overell et al. [OSvZ09] 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 showed that their ClassTag system increased the coverage of the vocabulary by 115% compared to a simple WordNet approach which classified Flickr tags by mapping them to WordNet via string-matching techniques. Unlike this work, they did not take into account how tags are used, but learned relations between tags and semantic categories by mapping them to Wikipedia articles.

Noll and Meinel [NM08b] investigated three types of metadata (social annotations, hyperlink anchor text and search queries) with respect to the characteristics of length, novelty, diversity and similarity. Their study found that tags are especially helpful for capturing the semantics of documents, and more useful for document classification than anchor words and search keywords. They also investigated social annotations on del.icio.us with respect to their usefulness for web page classification, attempting to gain insight into which documents get annotated more and in which way end users annotate the documents [NM08a].

Furthermore, they were interested in how the emerging folksonomy compared with an expert-maintained taxonomy, for the same documents. In contrast to a folksonomy, which emerges from social tagging, expert taxonomies have a well-defined, controlled vocabulary. In the taxonomy Open Directory, which they used for comparison, there are predefined category hierarchies; the categorization is done by experts and and goes through a process of peer-reviewing.

The study found that top-level documents are tagged more often than documents that are located deeper inside the content hierarchy of a web page. Furthermore, they found that the popularity of tags can help to identify the tags which provide more accurate classification information, but that also tags which are used infrequently provide helpful data for information retrieval and classification tasks. Other results of this study were that more popular documents had a higher consensus among taggers and that taggers preferred broad terms rather than specific terms.

Markines et al. [MCM+09] built an evaluation framework to compare folksonomy-based similarity measures, focusing on similarity among tags and resources on bibsomomy.org. They derived their measures from information-theoretic, statistical and practical measures and grounded them on WordNet and the Open Directory. Their framework, containing the tripartite folksonomy structure of users, tags and resources, compared how well similarity measures predict user-created tag relations. Additionally, they investigated the scalability of measures, studying whether they could be updated at a sufficiently fast pace to reflect new annotations.

In a two-step experiment, they first compared the ability of various tag similarity measures to predict user-created tag relations, and then provided an external semantic grounding. The results of their study showed that mutual information with distributional micro-aggregation across users, the most fine-grained approach, had the highest accuracy, extracting semantic similarity from a folksonomy best. However, this measure, which considers conditional probabilities between two objects, is computationally expensive and therefore not scalable. The best scalable approach was found to be per-user projection with collaborative aggregation, and its loss in accuracy was compensated by scalability. Their results were consistent across resource and tag similarity.

The work of Cantador et al. [CKJ11] set out to categorize social tags with the aim of improving folksonomy-based recommendations. In social tagging systems, content retrieval mechanisms face the limitation that tags can be freely chosen and that users have different intentions when tagging. For example, tags might describe the content of annotated resources, they might express contextual information about the resources, they might describe subjective qualities and opinions about them, or they might deal with organizational aspects of the tagger.

On the basis of the hypothesis that a large portion of tags are noisy for content retrieval, they presented a mechanism which automatically classified tags into purpose-oriented categories. Specifically, they examined whether it is generally possible to discover the underlying meanings of social tags, whether they can be automatically categorized based on the intention of the tagger, and whether this purpose-oriented categorization is useful for recommendation strategies.

For their experiments, the authors collected data from Flickr and developed a mechanism to automatically map social tags to semantic concepts, mapping the concepts of the tags to semantic entities on WordNet and Wikipedia. Then the concepts were transformed into semantic classes and context-based categories. Using these knowledge structures, they automatically inferred the semantic classes that can be used to determine the intention of a social tag. They were able to categorize 67.6% of their tags.

For an evaluation regarding whether the resulting tag categories really benefit a recommendation model, the authors used the Random Walk with Restarts method. Their results showed that some tag categories perform better than others – specifically, they found that content- and context-based tags performed better than subjective and organizational tags. Furthermore, they found that the recommendation system performed better when incorporating only content- and context-based tags. Subjective and organizational tags were not as useful for collaborative recommendation.

Cattuto et al. [CBHS08] introduced a systematic categorization and validation of tag similarity measures. For analyzing the different measures of tag similarity, they computed each measure on social tagging data obtained from the bookmarking system del.icio.us and then semantically grounded the pairs of similar tags by mapping them to pairs of synsets on WordNet. In addition to that, they used the measures of semantic distance, which had been validated in existing lexical databases, to characterize the semantic relations of the folksonomy. Five measures of tag similarity were analyzed: *co-occurrence count* of tags, *FolkRank* (an adaptation of PageRank to folksonomies), and the distributional measures of *tag context similarity*, *resource context similarity* and *user context similarity*.

Their study revealed several characteristics of the investigated similarity measures, showing that with the appropriate similarity measures, globally meaningful tag relations could be gathered from folksonomy vocabulary. Furthermore, they discovered which measures were more useful for different semantic applications. With respect to accuracy, they showed that distributional measures establish paradigmatic relations between tags. The distributional measure of *tag context similarity* is a computationally light measure that matched the accuracy of the most accurate and well-established similarity measure. Measures based on tag and resource context were found to be able to identify tags belonging to a common semantic concept.

Regarding the usefulness of the different measures regarding different semantic applications, the authors found that *tag* and *resource context similarity* were best for synonym discovery. For concept hierarchy, *FolkRank* and *co-occurrence* relatedness performed best, for tag recommendations, *FolkRank*, and finally, for query expansion, the best measures were *resource* and *tag context similarity*.

2.2. Research on Twitter

The popular micro-blogging platform Twitter has been extensively researched along many dimensions. This section reviews the related work dealing with Twitter. It firstly describes exploratory studies which focused on describing the structure of Twitter in general (Section 2.2.1), and then introduces related work regarding classification on Twitter (Section 2.2.2). Finally, a brief review of work regarding other aspects of Twitter is given 2.2.3. Research on hashtags is not included in this section; a review of the related work regarding hashtag semantics and pragmatics is given in the next section (Section 2.3).

2.2.1. Characteristics of Twitter

A study which was conducted by Huberman et al. in 2008 [HRW09] investigated social interactions on Twitter. They argued that a network constructed by following relationships was not actually the network that "mattered" to users, but that the actual network would have to be created via the patterns of interactions which users have with other users. Based on this hypothesis, they investigated with how many other users a Twitter user actually communicates directly through Twitter and defined a "friend" as a user to whom at least two posts were directed by the other user.

Comparing this number to the number of followees and followers that this user had, they found that users only interact directly with a few other users, even if they had a lot of followers and followees. The network that emerged from these "friend" relations was much sparser than the network of followers, which was very dense.

Concerning user activity, they found that users with many actual friends were more active than those with few actual friends, and that users with many followers or followees post updates less frequently than those with few followers or followees. They concluded that the driver of usage was a sparse and hidden network of connections underlying the network defined by the following relationships and that the number of "friends", as they defined it, is a more accurate predictor of a user's activity than his number of followers or followees.

Another early study on the topological and geographical aspects of the social network on Twitter was conducted by Java et al. [JSFT07] in 2007. They found that users tweeted about daily activities and used the platform to search for and share information. The main user intentions they found on Twitter were daily chatter, conversations, sharing information and reporting news. Additionally, there appeared to be three main categories of users on Twitter: information sources, information seekers, and friends. The study also found that users with similar interests or intentions connected with each other, a phenomenon called homophily which is a tendency for contacts between similar people occurring at a higher rate than between dissimilar people [MSLC01]. Homophily was also found in later studies on Twitter (e.g. [WpLJH10], [KLPM10]). Concerning the geographical distribution, the study found that Twitter was most popular in the United States, Europe and Asia, and that Tokyo, New York and San Francisco were the cities with the highest Twitter adoption.

Kwak et al. [KLPM10] conducted an extensive survey of Twitter, examining its uses, social aspects such as information diffusion, and the topological characteristics of its network. They crawled the entire Twitter site, collecting 41.7 million user profiles, 1.47 billion social relations and 106 million tweets. The authors were particularly interested in the way that the directed structure of relationships on Twitter impacted the topological characteristics of the network resulting from follower relationships. Comparing the characteristics of this network to the known characteristics of human social networks, they found considerable deviations: The distribution for followers did not follow a power law, the network exhibited a short effective diameter, and there was a low reciprocity in the following relations. Only 22.1% of the relations were reciprocal, while studies on other social networking services reported much higher numbers (for example, [CMG09] and [KNT10]). The authors stated that the short average path size of 4.12, which would be expected to be longer for a directed graph of this size, might suggest that twitter has a role other than social networking. This conclusion was recently confirmed by Kevin Thau, Twitter's vice president of business and corporate development, who stated that Twitter is not a social network, but rather a platform for news, content and information [Per10].

Ranking the users by different methods, they found that there was a gap in influence between the number of a user's followers and the popularity of his or her tweets. While the ranking of users by number of followers and PageRank was similar, the ranking by retweets was different. They also analyzed the tweets of the top trending topics regarding their temporal behavior and user participation. The majority of trending topics, over 85%, was headline news or persistent news. A large number of users participated in trending topics, and 15% of the participating users participated in more than 10 topics during a month. While the longest active trending topic, "big brother", lasted for 76 days, 31% of the life-spans were only one day. Only 7% of the trending topics lasted longer than 10 days.

Concerning information diffusion via retweets, they found that after the first retweet, information traveled fast and that any retweeted tweet reached an average of 1000 users, regardless of the number of followers of the author of the original tweet. They constructed retweet trees for trending topic and examined factors that impacted the spread of information. Up to about 1000 followers, the average number of additional recipients was not affected by the number of followers of the original author. Temporal analysis of retweeting showed that half of the retweeting occurred within less than an hour of the posting of the original tweet and 75% in less than one day. Only 10% of the retweets occurred after a month or later which indicated a fast diffusion of information.

For examining whether there was a link between geographic location and popularity, they used time zone as indicator and calculated the time differences between users. The median difference slowly increased as the number of reciprocal relations increased – users with less than 2000 reciprocal friends tended to be geographically close. Concerning popularity, they found a positive correlation between the number of followers of two reciprocally linked users, which also indicated the presence of homophily in reciprocal relations.

The study of Weng et al. [WpLJH10] attempted to identify influential users with a novel algorithm which they named TwitterRank. Their algorithm, an extension of the PageRank algorithm, measured the topic-sensitive influence of individual Twitter users by taking into account the topical similarity between different users and the link structure. They found that their TwitterRank algorithm outperformed other related algorithms such as PageRank, Topic-sensitive PageRank and ranking by indegree for identifying influential users. In the same study, they also investigated whether users that follow each other share similar topics, finding a high degree of homophily.

User influence on Twitter was also investigated by Cha et al. [CHBG10] who conducted an empirical analysis of influence patterns and of the different roles that users play in social media. They analyzed the Twitter network as a news spreading medium and studied the types and degrees of influence within the network as well as the dynamics of user influence across topics and over time. Concerning influence over time, they characterized the behaviors that made ordinary individuals gain high influence over a short period of time. The authors investigated directed links which determine the flow of information and indicate influence. The three measures of influence which they compared were *indegree* (the number of followers), *retweets* and *mentions*.

These measures represented three different types of a users' influence: *indegree* represented popularity, indicating the size of audience; *retweets* represented the content value of tweets, indicating an ability to generate content with pass-along value; and *mentions* represented the name value, indicating the ability to engage others in a conversation. The three types of influential users were examined with respect to their performance in spreading popular news topics.

The main result of their study was that having an active audience which retweeted and mentioned frequently was indicative for a user's influence, while a user's indegree revealed very little about his or her influence. Further, they found that the most influential users held influence over a variety of topics. Influence appeared to be gained by effort, such as limiting tweets to a single or a few topics.

Kim et al. [KJMO10] studied lists on Twitter and investigated whether they were useful for gaining insight about the characteristics and interests of users. In a first analysis of list names, they found that a large number of lists shared the same names. By aggregating the tweets of all users in a certain list, they found characteristics and interests that applied to all users in the list, even if these topics were not contained in an individual's tweets. Comparing the automatically identified interests with the interests which human evaluators assigned to the users, they found that the system using Twitter lists reflected the human perceived interests.

The work of [RDL10] attempted to characterize micro-blogs with topic models, with the goal of characterizing information needs which Twitter does not support at the moment. They implemented a partially supervised learning model (labeled LDA) which mapped content to four dimensions (substance, status, social and style), then characterized users and tweets using this model. They found that unmet information needs included better methods for finding and following new users and topics, and for filtering tweet feeds.

Macskassy [Mac12] studied social interactions on Twitter along three dimensions: user behaviors, characteristics of dialogues and characteristics of the social network that emerges from the interactions. In his study, he extracted dialogues from tweets and found that about half of the users interacted a lot while about 40% of users did not interact actively. Most of the dialogues were just between two people, even though the tweets were public. Furthermore, he found that most people either did not have dialogues or only spent about 5% to 10% of their general Twitter activity in direct interactions. The social network emerging from social interactions contained a giant component which was not very dense, but rather a set of tight and loosely connected clusters.

Wu et al. [WHMW11] investigated production, flow and consumption of information on Twitter, using Twitter lists to distinguish between elite users (celebrities, bloggers, media and representatives of formal organizations) and ordinary users. They examined the "two-step flow" theory of communication, which states that mass media influences the public only indirectly via an intermediate layer. This intermediate layer, consisting of users called opinion leaders, were categorized as ordinary users, but showed more connections and were also more exposed to media.

Almost 50% of the information originating from media passed to the public via opinion leaders, results which support the "two-step flow" theory. They also found that attention on Twitter was highly concentrated, that 50% of the total links in their dataset were created by only 20,000 users, that media produces most information and celebrities second most. Furthermore, they found a high degree of homophily in the relations and that different types of users tweeted about different topics. URLs broadcasted by different categories of users (and having different content) exhibited different lifespans, and also different content types had different lifespans. For example, videos and music had the longest lifespan, while content originating from media had the shortest lifespans.

2.2.2. Classification of Twitter Content

The work of Naaman et al. [NBL10] examined the activity of individuals on Twitter, conducting an exploratory study for creating a content-based categorization of tweet types. Using a combination of human coding and quantitative analysis, they characterized the content of tweets and investigated how the content varied by user characteristics, personal networks and usage patterns. They found two common types of Twitter users: *meformers* and *informers*. *Meformers* posted mostly "me now" tweets, while and *informers* posted tweets with an informational character.

Michelson and Macskassy [MM10] studied users' topics of interest by examining the entities which different users mention in their tweets. They presented a topic-profile for users: a list of the common, high-level topics found in their tweets. These topic-profiles characterised users' topics of interest by the categories of entities which were frequently found in their tweets.

The method they used was to find entities in tweets and then determine a common set of high-level categories covering the entities. To get encyclopedic knowledge about the entities and disambiguate them, they used Wikipedia as a knowledge base, and Wikipedia's user-defined categories were used to map the entities to the categories which they used to create the topic profile.

Sriram et al. [SFD+10] presented an approach to classify short text on Twitter. They argued that tweets are very short and therefore traditional methods such as bag-of-words have limited applicability. Developing a set of domain-specific features which they extracted from the Twitter user's profile and text, they focused on user intentions such as daily chatter, conversations, sharing information via links and reporting news. Their approach classified the tweets according to a predefined set of categories: news, events, opinions, deals, private messages. After collecting recent tweets and manually labeling them according to the best matching category, they used a Naive Bayes classifier for classification. Their results showed a high classification accuracy, outperforming their baseline of a traditional bag-of-words approach. Additional results were that users displayed a specific tweeting pattern and that users only tweeted within a limited amount of categories.

Hong and Davison [HD10] studied how topic models can be used for short text environments. They conducted experiments where they attempted to predict potential retweets and to classify Twitter users into topical categories. Studying how topic models can be trained on tweets, they examined different aggregation strategies, in order to discover whether they lead to different topic models and whether some strategies were faster than others. They found that different aggregation strategies yielded substantially different topic models and that in general, the model effectiveness was influenced by document length. Furthermore, they found that topic mixture distributions could help to improve classification.

2.2.3. Further Research on Twitter

This section focuses on further research which was conducted on Twitter. For example, researchers have examined anti-social psychological traits on Twitter [SBBP12], Twitter's impact on citizen journalism [Mur11] or the use of Twitter by politicians [LSAA11]; also the ethics of Twitter research has been addressed [Vie10].

The work of Sumner et al. [SBBP12] set out to predict the presence of Dark Triad personality traits via a linguistic analysis of tweets. The Dark Triad refers to the anti-social traits of narcissism, Machiavellianism and psychopathy. They explored to which extent it is possible to determine these traits based on Twitter use, by comparing the Dark Triad and Big Five personality traits of Twitter users with their user profile attributes and use of language in their tweets. Both a classification and a regression experiment were conducted: The classification was to identify individuals with high or low values of certain trait, the regression aimed to predict an individual's score on each of the personality traits.

For evaluation they used self-reported Dark Triad and Big Five personality trait scores and compared them to the values they obtained from the Twitter data. They found various statistically significant relationships. The most significant correlations were found between Dark Triad traits and language. Narcissism, for example, was positively correlated with @ and # characters and with the usage of words associated with sex. Machiavellian traits were positively correlated with swear words, anger and negative emotions, and negatively correlated with positive emotions and the use of "we". Psychopathic traits were positively correlated with swear words, anger, death and negative emotions as well as with filler words like, for example, "blah". The authors concluded that while models are not suitable for predicting the personality of individuals, they may be applied to large groups of users to investigate, for example, whether the presence of anti-social traits is rising or falling within a certain group.

The work of [WZO10] proposed a system for automatically generating personalized annotation tags to label a user's interests. In a keyword extraction task, they extracted words from the users' tweets: After preprocessing to remove different types of noise from the users' tweets, they performed TF-IDF ranking and TextRank to extract keywords from tweets. The top ranked keywords were used as tags for the user. Three human evaluators were employed to judge whether the top ranked keywords of the TF-IDF ranking and TextRank reflected the corresponding user's interests according to her tweets. Both ranking methods obtained a precision comparable to that of keyword extraction from web pages for content-targeted advertising, but TextRank outperformed TF-IDF in terms of precision. Also, there was a high variability among users.

Kietzmann et al. [KHMS11] presented a framework which used seven functional building blocks of different social media: identity, conversations, sharing, presence, relationships, reputation and groups. The authors stated that Twitter focuses more on conversation than identity.

The work of Murthy [Mur11] investigated the impact of Twitter on citizen journalism, specifically in the two cases of the US Airways flight 1549 and the Mumbai bomb blasts. He found indications for a rise of citizen journalism through the use of Twitter, but that the platform remained stratified by socioeconomic inequalities, that its usage was affected by the digital divide. Furthermore, he found that the attention of users gravitated towards traditional news media after a brief time of attention to the user who introduced the news.

The work of Skilters et al. [SKB+11] investigated pragmatic patters which characterize the construction of individual and collective identities on Twitter, studying discourse related to Latvian parliamentary elections. The aim of their study was to explore correlations between election results and the representations of political parties and their candidates on Twitter, and to explore the identity generation of politicians in pre-election communication. They conducted a topical analysis of election discussions and an analysis of hashtags and retweets, investigating richness of topics, channels of communication, frequency of mention and connotations and effects of tweets. Their data was statistically evaluated using the Pointwise Mutual Information algorithm and complemented with qualitative and quantitative content analysis.

The results indicated that the Twitter presence of individuals can be considered as "extended selves". The content analysis revealed the possibility of significant discrepancies in users' attitudes towards individual politicians and the political groups they belonged to. Frequent positive mention of an individual caused the perception of the significance of the relevant organization to fade to the background. Three factors were found contribute to the efficiency of political messages: the variety of thematic contexts, the frequency of mention, and positive connotations.

A study on Twitter's role as a new informal medium of communication at work was conducted by Zhao and Rosson [ZR09]. Studying the impact of Twitter on collaborative work, conducted several semi-structured phone interviews with employees of a large IT company. The participants were asked about their micro-blogging practices and their experiences of micro-blogging with co-workers. The study found that even among this small sample, Twitter was used in a wide variety of ways, and that Twitter could have a potential relational impact in three ways: person perception (constructing person schemas and background information to reduce social cognitive cost in interaction), common ground (increasing awareness of what the other person thinking about) and connectedness (a virtual feeling of proximity). Potential personal impacts were found to be work-relevant information sharing and expertise seeking.

How Twitter became an unofficial extension of the Eurovision song contest event was demonstrated in the work of [HHB12], which examined the expression of shared fandom on Twitter. The authors investigated the presence of a phenomenon called "audiencing" which is the *"public performance of belonging to the distributed audience for a shared media event"* [HHB12]. In order to discover how audiences expressed their fandom during the event and to characterize the networks and interaction between the participants, they examined several features, including tweeting patterns, -replies, retweets and the official event hashtags: #eurovision, #esc and #sbseurovision.

The study found an intersection between a transient audience and a loyal long-term audience in the #eurovision hashtag, indicating that Twitter could be seen as a technology for both long-term fandom and "audiencing". The authors concluded that Twitter acts as a platform which facilitates the connection and communion of fans and audiences.

Livne at al. [LSAA11] studied the use of the Twitter platform by politicians during elections in the United States. They analyzed differences between democrats, republicans and Tea Party candidates, investigating how election campaigns were expressed on Twitter. They proposed a method using language modeling for estimating content cohesiveness and divergence of individual Twitter users, including an analysis of text and graph mining techniques in order to characterize the features of different parties.

With the goal of characterizing the relation between network structure, tweet content and election results, regression models were employed to predict whether a specific candidate would be victorious in the elections. The study found a significant relationship between these factors and were able to predict the victory of individual candidates with an accuracy of 88%. The authors found that there was a significant difference in usage patters among the different parties and that, for example, conservatives candidates used Twitter more effectively than others. Tweets by members of the conservative party were more cohesive, and they exhibited a denser graph of connections.

2.3. Semantics and Pragmatics of Hashtags

Hashtags are strings of characters preceded by the hash (#) character and they are used on platforms like Twitter as descriptive labels or to build communities around particular topics [TR12]. The first introduction of the usage of hashtags was provided by Chris Messina in a blog post [Mes07a] (also see Section 1.5.4). Various aspects of semantics and pragmatics of hashtags have been investigated, including their diffusion dynamics (eg, [RTU13]), the different roles which hashtags hold (eg., [HTE10]), or their popularity peaks (eg., [LGRC12]).

Romero et al. [RTU13] studied hashtags in order to measure to which extent information popularity and social relations affect each other, in an attempt to bridge social and informational aspects together. They studied the interplay between the two, the extent to which they were related and whether they could predict each other.

In their work, they examined two decisive structures on Twitter: the graph structure of the social relations and the structure of topical affiliations. To build the graph of social relations, they used follower and communication (@-messages) relationships; to represent topical affiliations, they used hashtags. They conducted two experiments in order to identify key relationships between the two structures: predicting social relations with hashtags, predicting the future popularity of a hashtag by using social relations.

For the first experiment, predicting social links by examining the hashtags which were employed by users, they measured hashtag distance of Twitter users (for example via the smallest and largest hashtag that two users have in common). Using these measures, they tried to predict whether an edge was present between two arbitrary users in the social graph. The predictive power that their hashtag distance measures exhibited was high – up to 97% when taking into account the social subgraph that each hashtag defined.

In the second experiment, they examined the premise of viral marketing techniques, which states that edges in an existing social network may be used as bridges for information spread. In particular, they studied whether the structure of the social graph was related to what kind of topics would go viral in the future, by investigating to which extent a hashtag's future popularity could be predicted by the number of connections among its early adopters.

Their predictive algorithm took properties of the social graphs of early adopters as features and used them to predict the future popularity of the hashtag. To construct the different social graphs, they used uni-directional follower relations, reciprocal follower relations and @-message relations. Then they applied logistic regression to predict whether the number of Twitter users employing a hashtag would double in the future.

The study found that the structure of early adopter graphs had predictive power over the information diffusion process and that it could be used to predict the future popularity of a topic. However, the number of social connections was not linearly related with the future popularity of a hashtag. Instead, the hashtag became popular if the number of social relations of the early adopters was either very low or very high (i.e., where the graph of the initial adopters exhibited either very few or very many edges and singletons). These trends were true for short, medium and long term trends, as well as for varying amounts of early hashtag adopters. Romero et al. attempted to explain this result by the virality argument (for the case of many edges) and forces external to Twitter (for the case of few edges).

Wagner and Strohmaier [WS10] introduced the concept of "tweetonomy," a network-theoretic model of social awareness streams: a three-mode network, consisting of users, messages and resources. In their study, they explored whether the aggregation of messages in social awareness streams conveyed meaningful information about a certain domain. Introducing a set of stream-based measures targeted at systematically defining and comparing different stream aggregations, they found that different social awareness streams exhibited notable differences concerning the semantics which could be extracted from them. By transforming different aggregations of social awareness streams into lightweight, associative resource ontologies, they investigated if and what kind of knowledge could be acquired from them. The lightweight ontologies exposed how related two resources were, but did not contain any information about the semantics of relations.

The results of their study included that hashtag streams were rather robust against events such as New Years Eve, while user list streams were not. The authors concluded that hashtag streams are generally rather robust against external events, whereas user list stream aggregations are more susceptible to disturbances.

Letierce et al. [LPBD10] looked at three different Twitter hashtag streams based on the official hashtags of conferences. Their work was motivated by a human survey which the authors had conducted earlier and which had shown that most researchers wanted to communicate about their own research. Besides a detailed look at the content which users of the streams posted and how they did it, the work also examined structural properties of hashtag streams by investigating usage patterns and, specifically, user distributions.

The work of Huang et al. [HTE10] suggested that hashtagging in Twitter is more commonly used to join public discussions than to organize content for future retrieval. They stated that this kind of new tagging culture has created a completely new phenomenon, called *micro-meme*. A *micro-meme* is an emergent topic for which a hashtag is created. This hashtag is then widely used for a short period of time before disappearing again. While the *micro-meme* is active, its users form an asynchronous, massively-multiperson conversation around the topic. 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 for tag selection. This difference is due to the fact that users are influenced by the observation of the usage of micro-meme hashtags in other users' tweets, and they would have most likely not used the hashtag without this observation.

In their study, the authors used statistical metrics to describe activity of hashtags and to distinguish between two types of hashtags: conversational and organizational. Organizational hashtags facilitate discovery and access to resources at a later date, whereas conversational hashtags are an important part of the messages. The authors concluded that tags on Twitter are different to those in other social tagging systems (such as del.icio.us) in the way that users on Twitter are less likely to use tags for indexing messages for later retrieval, but rather use them for conversational purposes.

The role of hashtags was also investigated in [YSZM12]. Their study confirmed that a hashtag serves both as a tag of content and a symbol of community membership. By examining the factors of tagging content and joining communities, they found that the dual role of hashtags could be used to predict the users' behavior of adopting new hashtags.

The study investigated whether users were aware of this dual role, and whether the dual role influenced the behavior of adopting a hashtag. Based on different measures which attempted to quantify the affects that the dual role had on hashtag adoption, they predicted future adoption of hashtags. The factors concerning content tagging measured the relevance of a tag to the content of the tweet and the closeness to the user's preference. Measures related to community joining were prestige and influence of community members, calculated on the basis of the number of retweets, replies and mentions within the community of a hashtag's users. Apart from these factors, they also investigated several other factors that may influence hashtag adoption, but which are not relevant for investigating the dual role of hashtags (role-unspecific measures).

Logistic regression was used to predict the adoption of hashtags, using these measures as features. They found that all role-specific measures had significant predictive power to the future adoption of hashtags, indicating that the dual role of hashtags does in fact affect adoption. In order to investigate whether hashtag recommendation could be performed based on these features, they trained an SVM classifier for a binary classification task (whether a user will adopt a hashtag in future or not). Their best prediction model achieved an accuracy of about 80%, which indicated that it may be feasible to build recommender systems for hashtags.

Laniado and Mika [LM10] explored to what extent hashtags can be used as strong identifiers like URIs are used in the Semantic Web. They hypothesized that the intention behind introducing a new hashtag was to evolve it into a community symbol which helps to search and aggregate messages related to a certain topic, a function which is similar to that of shared URIs in the Semantic Web. In their work, they formalized a vector-space model for hashtags and 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.

They applied these metrics to a dataset acquired from Twitter and evaluated how well these metrics were able to identify hashtags which represent named entities and concepts found in Freebase, and that therefore constitute stable concepts with a unique identity. For evaluation they conducted a manual classification of a random sample of 257 hashtags and found that slightly more than half of the hashtags in the sample could be mapped to a Freebase entry. Their results indicated that the lexical usage of hashtags could indeed be used to identify hashtags which have the desirable properties of strong identifiers.

Bruns and Stieglitz [BS12] conducted a comparative study on the hashtags of about forty different cases such as elections, natural disasters and corporate crises, and identified various types of discussions which could be observed on Twitter. The study showed that thematic and contextual factors influenced original tweets, @-replies, retweets and URLs. Also, they found stable patterns of use in the context of different topics and events.

Based on the observation that popularity plays a major role in the dynamics of online systems, Lehmann et al. [LGRC12] set out to investigate popularity peaks of hashtags. Analyzing temporal, spacial and topical aspects of users' activity, they focused on spikes of collective attention in Twitter, specifically of those hashtags which exhibited a popularity peak during their observation period. They chose hashtags which were popular during their observation period, with at least 500 distinct users using them. The activity of those hashtags was investigated at the scale of days, and daily activity levels were analyzed. The peaks in the popularity of hashtags, specific features of a hashtag can now be related to its activity profile. Only events which were meaningful at this scale were examined.

Investigating activity profiles over time, they found three discrete classes of hashtags: hashtags with continuous activity (a constant level of daily activity), hashtags with periodic activity (a series of spikes spaced by one or more weeks) and hashtags with an activity concentrated around an isolated peak (unique events).

For the class of hashtags exhibiting an isolated peak, they further identified four groups of activity behavior, using the Expectation Maximization algorithm to learn an optimal Gaussian Mixture Model: anticipatory behavior (concentrated before and during the peak), unexpected events (concentrated during and after the peak), neither purely anticipatory nor purely reactive behavior (concentrated symmetrically around the peak) and events that were discussed only while they are happening (almost totally concentrated on the single day of the peak).

To provide semantic characterization of the hashtag classes, text mining techniques were employed. By grounding the words of the tweets in the WordNet semantic lexicon, they identified 18,000 distinct concepts which were associated with the hashtags. Concepts of WordNet (at depth 4) were used to link the four groups of activity behavior to semantic concepts. They found various semantic differences between the groups of activity: For example, hashtags with an activity concentration before and during the peak exhibited a greater reference to the concepts "social events" and "time period" than the other groups.

Tracking the propagation of the hashtags, they examined whether the different groups of hashtags exhibited different information propagation patterns. The results included that hashtags with an activity distributed symmetrically around the peak had a tendency to be retweeted more, suggesting a higher level of endogenous activity, while those hashtags associated with anticipatory behavior (activity concentrated before and during the peak) were less prone to viral spreading. The authors concluded that epidemic spreading plays a minor role in hashtag popularity, and that the spread of hashtags is driven mostly by exogenous factors.

Carter et al. [CTW11] proposed a method for tackling the problem of different hashtags being used in different languages to refer to the same event, which constitutes a difficulty for content analysis. A direct translation may not lead to the desired results, due to the problem that hashtags in one language do not match the hashtags used in another language. Their method for translating hashtags was based on methods from information retrieval and used translations of a hashtag profile to retrieve posts in the target language, of which hashtags were extracted and assumed to refer to the same topic as the hashtag under question. As proof of concept, they applied their method to the hashtag #33mineros which refers to the mining accident which occurred in August 2010 in a copper-gold mine in Chile. Their approach worked well on the one sample hashtag which they investigated, returning hashtags which were mostly suitable translations of the hashtag.

The work of Bruns and Burgess [BB11] investigated the use of hashtags in the formation of *ad hoc* publics in the context of political debate. While issue publics may only form post-hoc in many other environments, user-generated hashtags may be formed *ad hoc*. This ability is due to the fact that by including a hashtag in one's tweet brings the hashtag into being at the moment that it is first articulated. Furthermore, it is instantly disseminated to all of the tweet author's followers. Bruns and Burgess [BB11] argued that while not all hashtags represent *ad hoc* publics, this ability to form new hashtag communities constitutes the foundation for Twitter's recognition as an important tool for the discussion of current events. The authors conclude that by investigating the nature of the conversations within hashtag communities, researchers may trace the roles of individual participants and study how the community reacts to new stimuli. Their study provided some insights into the development of the community around the #spill hashtag, which was used to discuss the Labor leadership change in Australia in June 2010.

Bruns et al. [BBCS12] also studied the role of hashtags in crisis communication, specifically during the flood crisis which occurred in South East Queensland in 2011. They investigated the characteristics of sharing crisis information and dissemination of updates from authorities and normal citizens, and assessed the use of Twitter in such a crisis situation.

The hashtag #qldfloods became the central coordinating mechanism for user activity which was related to the floods crisis. The study found that the most visible account on #qldfloods was the Twitter account of the Queensland Police Service Media Unit, which was able to reach the audience effectively through their tweets. Their tweets were amplified with the help of other Twitter users via retweets: Tweets that contained information about the situation and advice, as well as news media stories and multimedia links were retweeted more often than tweets containing other information.

Overall, their findings demonstrated that Twitter played an important role in crisis communication and that the tweets authored by authorities were effective in terms of timeliness and informativeness, even though this role was still emerging and largely ad-hoc, with little planning.

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 Tsur and Rappoport [TR12] aimed to predict the exposure of a hashtag in a given time frame, while Romero et al. [RMK11] were interested in the temporal spreading patterns of hashtags.

Tsur and Rappoport [TR12] studied the effect of tweet content on information propagation. They attempted to predict the propagation of a new hashtag in the community within a certain time frame and investigated the way the content and structure of a hashtag drove its acceptance in the community. Using only global features such as hashtag content, global tweet features, graph topology features and global temporal features, they modeled the exposure and acceptance of a hashtag. The acceptance of a hashtag was captured by the normalized count of its acceptance in a time interval. For predicting the hashtag frequency after some time, they used a hybrid approach based on linear regression. They analyzed the contribution of the different feature types to the spread of hashtags, as well as the dependencies between the global features.

The results of their study included that a combination of the content, temporal and typological features performed best, but that also content aspects alone could be used as strong predictors. The authors concluded that content plays an important role in the acceptance of a hashtag by the community and that three main factors contributed to the acceptance of a hashtag: content, context and the social graph.

Romero et al. [RMK11] created a classification of hashtags by category, which was used as a starting point to create the dataset for this work. They identified eight broad categories in the 500 hashtags which were mentioned by most users within their dataset, each of the categories having at least 20 clear exemplars. They used manual annotation by both the authors and a group of independent annotators to assign the hashtags to the eight categories. The definitions of their categories can be seen in Appendix A. They found that the level of agreement among the annotators was high and that the categories showed identical behavior, independent of whether the classification was based on the authors' assignment, the independent group of annotators, or the intersection of both assignments.

The work of Yang and Leskovec [YL10] studied the spread patterns of Twitter hashtags and proposed a linear model which used the implicit network rather than the explicit network. The results of this study included that users with the most followers were not the most influential in propagating hashtags. Furthermore, the authors showed that the adoption of Twitter hashtags was governed by a large set of active users, each of which had relatively little influence.

Cunha et al. [CMC+11] presented a study of how hashtags are created, used and disseminated by members of information networks. They investigated the propagation of hashtags based on models for the analysis of the spread of linguistic innovations in speech communities, which are groups of people whose members linguistically influence another. The study identified aspects which were similar to those found in studies on offline speech. The authors state that hashtags were created similarly to the way lexical innovations take place when new terms are added to a language. Also, similar to the way a hashtag was spread and became popular (or not), new additions to language are either accepted or rejected.

The authors concluded that hashtags may serve as models for characterizing the propagation of linguistic forms, finding support for the existence of a "preferential attachment process" (also called the "rich-get-richer phenomenon") which states that the popularity of already popular words increases faster than the one of less popular words [CL06] [Sim55]. Also, they found a relationship between length of a tag and frequency of use (popular hashtags were shorter, on average) and hashtags containing underscores were found to be less popular.

This chapter has summarized the relevant related work regarding emergent semantics, the Twitter platform, and the semantics and pragmatics of hashtags. The next chapter introduces the dataset which I comprised for the experiments presented in this thesis.

3. Dataset

The dataset for this work was collected using Twitter's publicly available REST API. This chapter describes the concepts necessary for understanding the dataset, the dataset's structure as well as the process of data collection. It is organized as follows: Section 3.1 introduces the concepts of hashtag streams, social connections and stream audience, as well as the method which was introduced for ranking the users which are related to a stream. Section 3.2 gives an overview of the dataset's structure, and Section 3.3 describes the process of data collection as well the implementation of the crawler. The research questions and the experimental setups will be explained in Chapters 5 and 6.

3.1. Concepts

The dataset consists of hashtag streams, social connections and tweets posted by the top audience of each hashtag stream. In this section, these concepts are explained in detail.

3.1.1. Hashtag Streams

Hashtag streams constitute a special type of *social streams*, which are streams of data produced by users in online social environments, and are intended to be viewed by other users. A hashtag stream consists of all messages containing a specific hashtag and all resources and users related to these messages. Related resources include other hashtags, URLs and keywords. Specifically, they can be defined as a tuple consisting of users (U), messages (M), resources (R), a ternary relation (Y') between U,

M and R, and a function (ft) which assigns a temporal marker to each (Y') [WS10]. In this work, users who posted one or more messages in a hashtag stream are called *authors* of this stream.

3.1.2. Social Connections

Some of the pragmatic features used in this work (see Chapter 4) 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, users who hold both of these roles (i.e., have established a bidirectional link with an author) are called *friends*.

Specifically, the relations are defined as:

- User U_1 is a *follower* of user U_2 if there exists a unidirectional link from U_1 to U_2
- User U_1 is a *followee* of user U_2 if there exists a unidirectional link from U_2 to U_1
- User U_1 is a *friend* of user U_2 if there exists a bidirectional link between U_1 and U_2

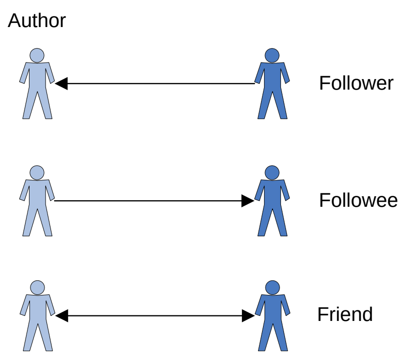


Figure 3.1.: Followers, followees and friends

3.1.3. User Ranking and Audience

As information consumption is driven by explicitly defined social networks in social online environments, the *audience* of a social stream can be estimated by analyzing the incoming and outgoing links of the authors who created the stream. Since the audience of a stream is potentially very large, the users which were related to each hashtag stream were ranked in the following way:

- **Top Authors:** A list of users ranked by the amount of tweets which they contributed to a specific hashtag stream.
- **Top Followers:** A list of users ranked by the number of authors that they follow.
- **Top Followees:** A list of users ranked by the number of authors that follow them.
- **Top Friends:** A list of users ranked by the number of reciprocal relationships that they have with the authors.

This ranking allows to determine the key audience members per hashtag stream. An exemplary illustration of the ranking is given in Figure 3.2, showing the ranking of stream friends. Table 3.1 shows an example of the top ten ranked authors, followers, followees and friends of the hashtag #ebay.

Table 3.1.: Top 10 authors, followers, followees and friends of the hashtag #ebay

Rank	Top Authors		Top Followers		Top Followees		Top Friends	
	User Id	#Tweets	User Id	#Authors	User Id	#Authors	User Id	#Authors
1	39553219	118	317895169	64	19709040	64	39553219	46
2	380016302	102	39553219	48	11522502	55	317895169	45
3	295544793	82	40957993	48	813286	51	40957993	44
4	414691532	77	43344858	45	39553219	46	19250651	42
5	254117479	73	254117479	44	317895169	45	43344858	42
6	59545760	65	19250651	43	40957993	44	18772703	41
7	44059194	54	18772703	42	15907720	43	254117479	36
8	214018568	47	481417356	42	14230524	42	27161746	35
9	308931718	27	28560724	38	15846407	42	219114054	35
10	161927644	23	49461885	38	18772703	42	18671234	34

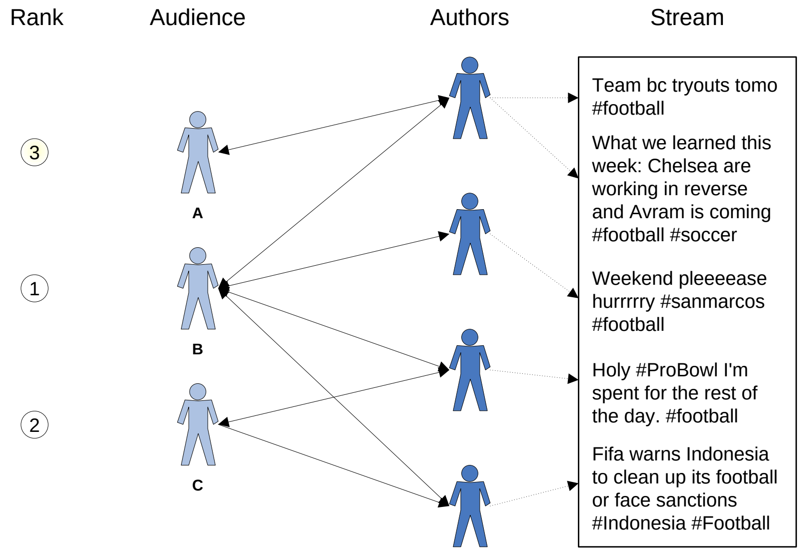


Figure 3.2.: Audience Ranking. To estimate the audience of a hashtag stream, the friends of the stream's authors are ranked by the number of authors which they are related with. In this example, the hashtag stream #football has four authors. User B is a friend of all four authors of the stream and is 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.

3.2. Structure of the Dataset

The complete dataset consists of three parts, each part representing a time frame of four weeks. Different time frames ensure that it is possible to 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. The starting dates of the time frames were March 4th (t_0), April 1st (t_1) and April 29th, 2012 (t_2). Table 3.2 depicts the number of tweets and relations between users that were collected during each time frame.

3.3. Data Collection

This section describes how the data was collected from Twitter. First, the process of hashtag selection is explained, then the process of the data collection is described and finally, a brief description of the crawler's implementation is given.

3.3.1. Hashtag Selection

Romero et al. [RMK11] conducted a user study and a classification experiment, identifying eight broad semantic categories of hashtags: *celebrity*, *games*, *idiom*, *movies/TV*, *music*, *political*, *sports* and *technology*. The starting point for the dataset of this work was 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. The definitions of the categories can be seen in Table A.1.

Table 3.2.: Description of the complete dataset

	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
Unique Followers	23,538,998	24,131,957	25,989,854
Unique Followees	10,073,580	10,649,087	11,355,138
Unique Friends	7,312,792	7,896,758	8,390,143
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
Total User Relations	92,958,706	92,654,511	103,658,310
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

For each category, ten hashtags were chosen at random (see Table 3.3). The random sample was biased towards active hashtag streams by re-sampling hashtags for which less than 1000 posts were found at the beginning of the data collection (March 4th, 2012). For those categories where less than ten hashtags were found that had more than 1000 posts (i.e., *games* and *celebrity*), the most active hashtags per category were selected (i.e., the hashtags for which we found the most posts).

Table 3.3.: 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

Trending Hashtags

For the time frames t_1 and t_2, the category *trending* was added as a ninth category. The *trending* category consisted of the ten most recent hashtags that Twitter had listed in "Worldwide Trends" (also see Section 1.5.3) at the start of t_1. However, these hashtags, along with the corresponding stream tweets, social connections and audience tweets which were only related to these hashtags (and not to any other stream) were removed from the dataset. The removal of this category was due to the fact that all trending topics were already inactive at the start of t_2. Table 3.4 depicts the trending topics which were crawled.

3.3.2. Overview of the Process

At the start of each time frame, the most recent tweets in English were retrieved for each hashtag using Twitter's public search API. Afterwards, the followers and followees of each user who had authored at least one message in any hashtag stream were retrieved. Finally, the user timelines

Table 3.4.: Trending hashtags streams retrieved at t_1 and t_2. All the trending streams were inactive four weeks later.

Hashtag	#Tweets t_1	#Tweets t_2
#22MillionGovernmentHookers	1441	0
#CongratsLarryforthebaby	1457	23
#falseliamisreal	1347	0
#halpgmh	1249	0
#Infinite2ndInvasion	1454	0
#JustinHowCouldYou	1387	2
#OnlyFakePeople	1486	92
#replaceATLsongswithAliKing	1474	0
#RIPmaz	1471	0
#slateronthebuzz	1462	139

of the top 100 authors, followers, followees and friends of each hashtag stream were crawled, retrieving up to 3,200 tweets of each user. These stream audience tweets were not used in the experiments described in this thesis, but were the basis for the experiments conducted in [WSPS13].

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, and in the subsequent two weeks of each timeframe, the audience tweets were retrieved. As this process was performed identically for each time frame, it was ensured that the social information crawled at t_0 was established previously to the tweets crawled at t_1 (and the social information of t_1 was established previously to the tweets of t_2). Figure 3.3 depicts this process.

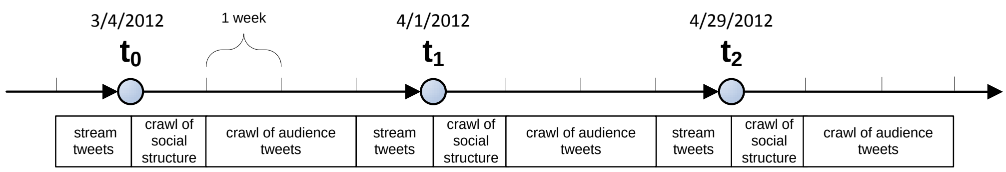


Figure 3.3.: Timeline of the data collection process

3.3.3. Implementation

The crawler was implemented in C# using the Twitterizer library (version 2.4) [Twi13g]. The data was stored in three separate SQLite databases (version 3.7.3), one for each time frame.

Twitter provides two different types of APIs: The REST API and the streaming APIs. The streaming APIs provide real-time access to tweets. Several streaming endpoints are provided, each for a different user case: public streams, user streams and site streams [Twi13a]. Because different snapshots of specific hashtag streams, including the social connections and the corresponding audience tweets, were required for this dataset, the REST API was better suited for the collection of the data. At the time of data collection, the REST API was at version 1.

The steps which were performed for each time frame during data collection are described in detail below.

Stream Tweets

For each hashtag, the most recent tweets were retrieved, up to a maximum of 1500 tweets, which is the Twitter API limit. The search call allows for a parameter to specify the language, and only English tweets were retrieved [Twi13e]. The following REST call was used:

- GET search: "Returns relevant tweets that match a specified query." [Twi13f]

For each tweet, the following information was stored:

- Tweet ID
- Date of creation
- Author ID
- Tweet text

Followers and Followees

For each user who had authored one or more tweets in any of the hashtag streams, the followers and followees were retrieved. If a user's profile was not private and an error was received from the API, the request was resent up to 20 times before ignoring the user and marking an error in the database. These users were recrawled later and at the end of the collection of each time frame, all users which had not set their profile to private were successfully crawled. For each user relation pair, the user IDs were stored.

The following REST calls were used:

- GET followersids: "Returns an array of numeric IDs for every user following the specified user." [Twi13f]
- GET friendsids: "Returns an array of numeric IDs for every user the specified user is following." [Twi13f]

Audience Tweets

After creating ranked lists of authors, followers, followees and friends, the user timelines of the top 100 users of each category were retrieved, up to a maximum of 3200 tweets (which is the API limit). For each tweet in the user timelines, the following information was stored:

- Tweet ID
- Date of creation
- Author ID
- Tweet text
- Whether the tweet is a retweet

The following REST call was used:

- GET statusesuser_timeline: "Returns the 20 most recent statuses posted by the authenticating user. It is also possible to request another user's timeline by using the screen_name or user_id parameter." [Twi13f]

Language Detection

For identifying the English tweets in the user timelines, the library DialogueMaster Babel [Zeu09] was used. This library is based on n-gram and word occurrence comparison and uses Wikipedia as primary source for statistics of character co-occurrences. For classification, the text to classify is tokenized and the resulting table is compared to all tables in the model. The most likely language is the one with the smallest distance to the text to classify. Figure 3.4 shows the demo interface of the library with a correctly classified tweet.

The library has an accuracy between 70% (for short Norwegian, Swedish and Danish models) and 99.8%, depending on the model and the length of the input text [Zeu09]. Tweets are difficult to classify as they consist of short and informal text, but as the task was not to correctly detect any language but only to detect whether the tweet was in English or in another language, the library performed well on a set of 2000 randomly selected and manually labeled tweets. On this set of tweets, consisting of 1000 English tweets and 1000 tweets which were in a different language, the library correctly classified 95.5% of the English tweets as "English" and 96.3% of the tweets in a different language as not English.

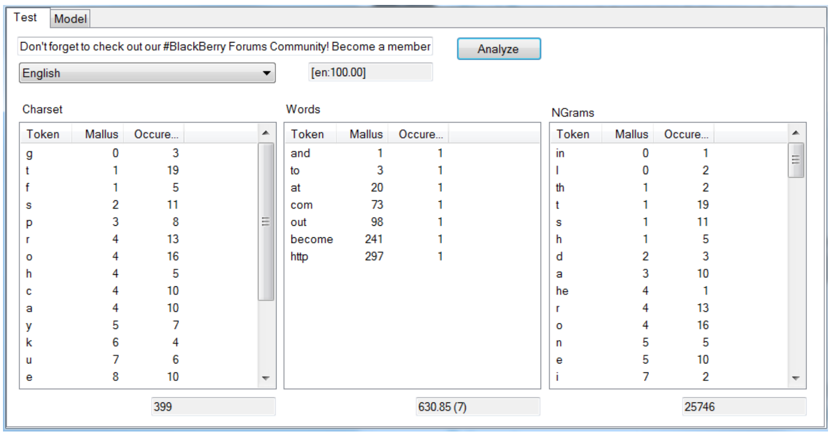


Figure 3.4.: Correctly classified English tweet in the demo interface of the Babel library

4. Methodology

The aim of this work is to explore the usage patterns of different types of hashtags in order to investigate whether these patterns are substantially different for different types of hashtags. For capturing the usage patterns of hashtag streams, several sets of features were developed. This chapter describes these measures as well as the terminology which was introduced to specify the different timeframes and combinations of features which were used in the experiments. The pragmatic measures are designed to capture the different social and message based structures of hashtag streams. In addition to pragmatic measures, also lexical measures were introduced for comparison. These lexical measures are based on a bag-of-words model using term frequency as weighting schema.

Pragmatic measures are further differentiated into static pragmatic measures and dynamic pragmatic measures. Static pragmatic measures capture the social structure and usage patterns of a hashtag at a specific point in time while dynamic pragmatic measures combine information from several time points.

This chapter is organized as follows: Section 4.1 goes into detail on the static pragmatic features, the dynamic pragmatic features are characterized in Section 4.2. Section 4.3 explains the lexical measures which were used to capture the co-occurring words. Finally, Section 4.4 specifies the terminology which was created to denote the different timeframes.

4.1. Static Pragmatic Measures

Static pragmatic features capture different aspects of a hashtag stream's usage patterns at specific points in time. This section describes the different sets of static pragmatic features which were introduced for this purpose.

4.1.1. Entropy Measures

Entropy measures help to characterize the distribution of the users involved in a hashtag stream (i.e. authors, followers, followees and friends), measuring their randomness. For each hashtag stream, the authors are ranked by the number of messages they published in that stream and the followers, followees and friends are ranked by the number of stream's authors they are related with (also see Section 3.1.3). A high entropy indicates that the users are equally distributed, while a low entropy suggests that the distribution is skewed.

The Shannon entropy is defined as follows, $\{x_1, ... x_n\}$ constituting the outcomes of a random variable X [BYRN11]:

$$H(X) = -\sum_{i=1}^{n} P(x_i) \log(P(x_i)) \tag{4.1}$$

In the case of the entropy measures, the random variable U refers to the total number of authors/followers/followees/friends per stream depending on which entropy (i.e., author-, follower-, followee-, friend-entropy) is computed. The probability of an author u_i in a stream is defined by the number of messages in the stream which were authored by this user, divided by the total number of messages. The probability of a follower/followee/friend u_i is defined by the number of authors of the stream this user is related with, divided by the total number of followers/followees/friends of all stream authors. Thus, the entropy measures are defined as:

$$H(U) = -\sum_{i=1}^{U} P(u_i) \log(P(u_i)) \tag{4.2}$$

Each of the entropy measures was normalized to the number of users in order to make them comparable. The *normalized entropy* $h(n)$ [CMS10] can be defined as:

$$h(n) = \frac{H(X(n)))}{\log n} \tag{4.3}$$

In the rest of this section, the different entropy measures, which are based on different kinds of user distributions, are discussed in detail.

Author Entropy

For each hashtag stream, the distribution of authors was computed, based on how many messages each author had contributed to the snapshot of the stream. A high *author entropy* indicates that the stream is created in a democratic way since all authors contribute equally much, while a low entropy suggests that a stream is dominated by few selected authors.

Follower Entropy

Knowing the authors of a hashtag stream, one can estimate the potential audience of a hashtag stream by creating a ranked list of the authors' followers. The rank of a follower depends on how many authors of the hashtag stream he or she follows. For each hashtag stream, the distribution of followers was computed, based on how many authors of the hashtag stream the user was following. A high *follower entropy* indicates that the followers do not focus their attention towards few authors but distribute it equally across all authors.

Followee Entropy

Knowing the authors of a hashtag stream, one can estimate the group of users from whom the authors of a hashtag stream may consume information (and may therefore be influenced by them) by creating a ranked list of the authors' followees. The rank of a followee depends on by how many authors of the hashtag stream he or she is followed. For each hashtag stream, the distribution of followees was computed, based on how many authors of the hashtag stream followed the user. A high *followee entropy* indicates that the authors do not focus their attention on a selected part of their audience.

Friend Entropy

Knowing the authors of a hashtag stream, one can estimate the acknowledged audience of a hashtag stream by creating a ranked list of authors' friends (i.e., users who follow and are followed by an author). The rank of a friend depends on how many authors of the hashtag stream he or she follows who also follow him or her back. For each hashtag stream, the distribution of friends was computed, based on how many authors of the hashtag stream the user had a reciprocal following relation with. A high *friend entropy* indicates that the friends do not focus their attention towards selected authors but distribute it equally across all authors.

4.1.2. Overlap Measures

The three *overlap measures* describe the overlap between the authors (A) and the followers, followees or friends (F) of a hashtag stream. It is defined as follows:

$$overlap(A, F) = \frac{A \cap F}{min(A, F)} \tag{4.4}$$

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 the active and passive parts of the community do not extensively overlap.

The overlap measures can be computed at different ranks. The rank indicates how many of the top followers/followee/friends are compared with the authors. For example, N@10 indicates that only the top ten followers, followee or friends are taken into account. The overlaps were calculated for N@10,000.

Author-Follower Overlap

The author-follower overlap measures the overlap between the top authors and the top followers of a hashtag stream. If the overlap is *one*, this indicates that all top authors are also top followers, which indicates that the stream is consumed and produced by the same users.

Author-Followee Overlap

The author-followee overlap measures the overlap between the top authors and the top followee of a hashtag stream. If the overlap is *one*, this indicates that all top authors are also top followees, which indicates that the stream is influenced and produced by the same users.

Author-Friend Overlap

The author-friend overlap measures the overlap between the top authors and the top friends of a hashtag stream. If the overlap is *one* this indicates that all top authors are also top followees, which indicates that the stream is consumed, produced and influenced by the same users.

4.1.3. Coverage Measures

Coverage measures characterize a hashtag stream via the nature of its messages. Four coverage measures were introduced: *informational coverage*, *conversational coverage*, *retweet coverage* and *hashtag coverage*. These measures are described in detail below.

Informational Coverage

The *informational coverage* measure indicates how many messages of a stream have an informational purpose – i.e., contain a link. From the number of informational messages $|M_i|$ and the total number of messages of a stream $|M|$ one can compute the informational coverage of a stream which is defined as follows:

$$IC = \frac{|M_i|}{|M|} \tag{4.5}$$

Conversational Coverage

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 (e.g., through @replies). From the number of conversational messages $|M_c|$ and the total number of messages of a stream $|M|$, one can compute the conversational coverage of a stream, which is defined as follows:

$$CC = \frac{|M_c|}{|M|} \tag{4.6}$$

Retweet Coverage

The *retweet coverage* measures the percentage of messages which are retweets. Using the number of retweets $|M_r|$ and the total number of messages of a stream $|M|$, one can compute the retweet coverage of a stream, which is defined as follows:

$$RC = \frac{|M_r|}{|M|} \tag{4.7}$$

Hashtag Coverage

The *hashtag coverage* measures the mean number of hashtags per message in a stream. From the total number of hashtags $|H|$ and the total number of messages of a stream $|M|$, one can compute the hashtag coverage of a stream, which is defined as follows:

$$HC = \frac{|H|}{|M|} \tag{4.8}$$

4.2. Dynamic Pragmatic Measures

Dynamic pragmatic measures capture how a hashtag stream is used over time. These features require two time points for calculation as the difference between the distributions is captured. The dynamic pragmatic measures use the Kullback-Leibler divergence as basis.

4.2.1. Divergence Measures

To explore how the social structure of a hashtag stream changes over time, 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 is measured. 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.

The *Kullback-Leibler (KL) divergence* represents a natural distance measure between two probability distributions (A and B). The KL divergence is *zero* if the two distributions are identical and approaches infinity as they differ more and more. It is also known as *relative entropy* or *information* *divergence*. The KL divergence D_{KL} between two random variables A and B is defined as as follows [Shl07]:

$$D_{KL}(A||B) = \sum_i A(i) \log \frac{A(i)}{B(i)} \tag{4.9}$$

For the dynamic pragmatic features, a symmetric variation of the *Kullback-Leibler divergence* [JS01] is used:

$$\frac{1}{2} D_{KL}(A||B) + \frac{1}{2} D_{KL}(B||A) \tag{4.10}$$

The dynamic pragmatic features measure the symmetric KL divergence for the distributions of authors, followers, followees and friends. The KL divergence was calculated for the top 10,000 users as well as for the whole set of users for each user type and hashtag stream.

Temporal Author Dynamics

For measuring the change in author distributions, the KL divergence of the distributions of authors of a hashtag was calculated between two points in time. The authors distribution of a hashtag depends on the number of messages the author has contributed to the stream compared to the total number of messages. A high KL divergence for authors indicates that the authors who contribute to a hashtag stream fluctuate widely whereas a low KL divergence indicates that the stream is created by a stable community.

Temporal Follower Dynamics

For measuring the change in follower distributions, the KL divergence of the distributions of followers of a hashtag was calculated between two points in time. The follower distribution of a hashtag depends on the number of hashtag authors a user follows compared to the total number of hashtag-authors. A low KL divergence for followers indicates that there is a stable community of users consuming the hashtag stream.

Temporal Followee Dynamics

For measuring the change in followee distributions, the KL divergence of the distributions of followees of a hashtag was calculated between two points in time. The followee distribution of a hashtag depends on the number of hashtag authors a user is followed by compared to the total number of hashtag-authors. A low KL divergence for followees indicates that the hashtag stream is influenced by a stable set of users.

Temporal Friend Dynamics

For measuring the change in friend distributions, the KL divergence of the distributions of friends of a hashtag was calculated between two points in time. The friends distribution of a hashtag depends on the number of hashtag authors a user is friend with compared to the total number of hashtag-authors. A low KL divergence for friends indicates that there is a stable community of users which both influences and consumes the hashtag stream.

4.3. Lexical Measures

For the lexical measures, a vector-based method was used, which allows representing each hashtag stream as a vector of terms and uses term frequency as weighting schema. In this work, lexical measures are always computed for individual time points and are therefore static measures.

4.3.1. Term Frequency

Term frequency is based on the *Luhn assumption*, which states that *"the value, or weight, of a term k_i that occurs in a document d_j is simply proportional to the term frequency $f_{i,j}$. That is, the more often a term k_i occurs in the text of a document d_j, the higher its term frequency weight $TF_{i,j}$ is"* [BYRN11]. The assumption is based on the observation that terms which occur frequently in a document are important for describing its key topics. Thus, the term frequency (*TF*) weight formulation is defined as [BYRN11]:

$$TF_{i,j} = f_{i,j} \tag{4.11}$$

For the lexical features, *TF* is calculated for each hashtag stream.

4.4. Time Frames

As dynamic features require two points in time for calculation, a terminology was introduced to denote the different combinations of features regarding the timeframes.

Figure 4.1 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 analogously, with $t_{1\to2}$ containing the static features computed on data collected at t_1 and the dynamic measures computed on data collected at t_1 and t_2, and $t_{2\to1}$ containing the static features computed on data collected at t_2 and the dynamic measures computed on data collected at t_1 and t_2.

This chapter has introduced the pragmatic measures which were designed to capture the social and message based structures of hashtag streams,

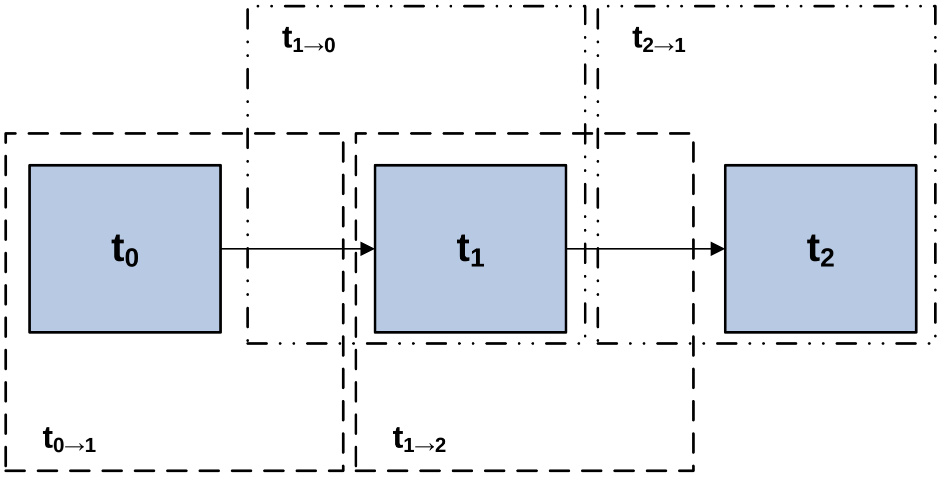


Figure 4.1.: Illustration of the time frames

as well as the lexical measures which were introduced for comparison. The following two chapters will describe the experimental setup and the results of the experiments which were conducted using the presented measures. These experiments quantify the association between the pragmatic measures and semantic categories of hashtags, and assess the utility of the pragmatic measures for classifying hashtag streams in their semantic categories.

5. Experiment 1: Usage Patterns of Hashtag Categories

The first experiment aims to explore to which extent usage patterns of hashtag streams in different semantic categories are indeed significantly different. It targets the first research question:

Do different semantic categories of hashtags reveal substantially different usage patterns?

The experiment was designed to explore the idiosyncrasies of hashtag usage within semantic categories, and to investigate whether there are significant differences between the categories. To this end, the distributions of the measures defined in Chapter 4 were compared across different hashtag types. This chapter describes the experimental setup and statistical methods used for the first experiment (Section 5.1) and reports the results obtained (Section 5.2).

5.1. Experimental Setup

For investigating whether different semantic categories of hashtags exhibit substantially different usage patterns, statistical standard tests were employed to reveal the similarities and differences concerning how they are used, by whom they are used, influenced or consumed, or for which purpose – e.g. chat or information sharing – they are primarily used. In this experiment, the timeframes $t_{0\rightarrow 1}$ and $t_{1\rightarrow 2}$ were used. This section first describes how the selection of hashtag streams was modified for the experiments, and then explains the statistical tests which were used to reveal the idiosyncrasies of the different semantic categories.

Removal of Hashtag Streams

For the experiments conducted in this work, hashtag streams which belonged to multiple categories were removed since the aim of the experiments was to learn what types of characteristics are useful for describing a semantic hashtag category. Concretely, the two hashtags #bsb and #mj were removed. Additionally, inactive hashtag streams (those where less than 300 posts where retrieved) were removed as estimating information theoretic measures is problematic if only few observations are available [PT96]. The most common solution is to restrict the measurements to situations where one has an adequate amount of data. Four inactive hashtags were found in the category *games* and seven in the category *celebrity*. The hashtag streams that were removed during this step were #e3, #mafiawars, #spymaster and #uncharted2 for the category *games*; in the category *celebrity* the streams #ashleytisdale, #brazilmissesdemi, #niley, #regis, #teamtaylor, #tilatequila and #weloveyoumiley were removed.

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 of this category is that topics related to celebrities have a shorter life span than topics related to other categories. The only other category where this problem partly occurred was the category *games*, where four of the ten original hashtag streams had to be removed. The final datasets which were used in the experiments of this work consist of 64 hashtag streams and seven semantic categories which were sufficiently active during the observation period.

Statistical Tests

To compare the pragmatic fingerprints of hashtags belonging to different semantic categories and to quantify the differences between categories, a statistical hypothesis test was chosen. The most commonly used method to evaluate the differences in means of two groups is the *t-test*. However, the *t-test* requires the data to be normally distributed within each group as well as a variance which is not reliably different between the groups. If the normality assumption is not met, a non-parametric alternative to the *t-test* has to be used [HL06].

The *Shapiro-Wilk-Test* can be used for testing whether a sample originates from a normally distributed population, which is a prerequisite for parametric tests like the *t-test* [DK09] [Roy95]. The test statistic W is given by

$$W = \frac{(\sum\limits_{i=1}^{n} a_i x_i)^2}{\sum\limits_{i=1}^{n} (x_i - \bar{x})^2} \tag{5.1}$$

where a_i are weighting constants, $\bar{x}$ is the mean of the sample, and x_i is the i-th order statistic of the sample (i.e., the i-th smallest value).

If the data does not satisfy the prerequisite of normality, data transformation can be used to achieve a normal distribution of skewed data. Percentages and ratios follow a binomial distribution, where the variance is a function of the mean. This dependency can be uncoupled by applying *arcsine transformation* to the data. *Arcsine transformation* (also called *angular transformation*) stretches values close to zero and 100 while compressing values close to the center [DK09]. It is defined as

$$y' = \arcsin(\sqrt{y/100}) \tag{5.2}$$

The *Shapiro-Wilk-Test* revealed that not all features were normally distributed, even after applying arcsine transformation to ratio measures, i.e., entropy, overlaps and the coverage measures which were expressed as a percentage. Therefore, a non-parametric test was required to test the differences in medians.

The *Mann-Whitney-Wilcoxon-Test* (also called *Mann–Whitney U Test* or *Wilcoxon Rank-Sum Test*) is a statistical hypothesis test for assessing whether one of two samples of independent observations tends to have larger values than the other, constituting a non-parametric alternative to the two-sample *t-test* [Ros09]. The test uses the sum of the ranks for observations from one of the samples as test statistic [Wil99].

The null hypothesis for this test states that there is no difference between the medians of the groups:

$$H_0 = \psi x - \psi y = \delta_0 \tag{5.3}$$

The two-sided alternative hypothesis for this test states that the samples come from different distributions – that there is a difference in the medians [UMA08]:

$$H_a = \psi x - \psi y \neq \delta_0 \tag{5.4}$$

As the *Mann-Whitney-Wilcoxon-Test* assumes equal variances, a *Levene's test* [HL06] for equality of variances was conducted. The test revealed that the variances could be assumed to be equal (the null hypothesis of equal variances did not have to be rejected).

With the data meeting the requirements for the *Mann-Whitney-Wilcoxon-Test*, a pairwise version of the test was conducted, testing all category pairs for statistically significant differences. For the significance criterion α, the threshold of 5% was chosen, which is a common choice for α [Dow11].

When multiple comparisons are performed, there is a need for adjusting the p-values in order to avoid a high amount of *type I* errors. *Type I* errors occur when the null hypothesis is true, but rejected (resulting in a false significant difference). The more hypotheses are tested in the multiple comparisons, the more likely it is for *type I* errors to occur.

In this experiment, the *Holm-Bonferroni* method was used for adjusting the p-values and counteract the problem of multiple comparisons. The *Holm-Bonferroni* method is a sequentially rejective method which has a prescribed level of significance protection against *type I* errors (for any combination of true hypotheses) [Hol79].

5.2. Results

This section presents the results from the empirical study on usage patterns of different semantic categories of hashtags.

Generally, the 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 5.1 depicts the measures that show statistically significant ($p < 0.05$) differences in both $t_{0\to1}$ and $t_{1\to2}$. The statistical differences in $t_{0\to1}$ are depicted in Table 5.2, along with their significance levels. The differences in $t_{1\to2}$ are displayed in Table 5.3.

Table 5.1.: This table depicts which features showed a statistically significant difference (with $p < 0.05$) for each pair of categories in both $t_{0\to1}$ and $t_{1\to2}$.

	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	

The results suggest that the pragmatics of hashtags are relatively stable over time, since the statistical tests were conducted on the first and the second timeframe ($t_{0\to1}$ and $t_{1\to2}$), and 26 significant pragmatic category differences were found to be significant in both $t_{0\to1}$ and $t_{1\to2}$. For $t_{0\to1}$, 35 pragmatic category differences were found to be statistically significant (with $p < 0.05$); in $t_{1\rightarrow 2}$, 33 differences were found. In total, 315 comparisons were made for each time frame (15 measures and 21 pairwise comparisons).

Table 5.2.: To assess the utility of different pragmatic features for differentiating between hashtag streams of different semantic categories, a pairwise Mann-Whitney-Wilcoxon-Test was conducted. This table shows which features showed a statistically significant difference (with $p < 0.05$) for each pair of categories in $t_{0\rightarrow 1}$. (significance levels: < 0.001 '***', < 0.01 '**', < 0.05 '*')

	games	**idioms**	**movies**	**music**	**political**	**sports**
idioms	kl_authors* informational* retweet**					
movies	kl_followers*	informational**				
music		informational**				
political	entropy_friend** kl_followers**	kl_authors** kl_followers*** kl_followees** informational*** hashtag*		entropy_friend* kl_followers*		
sports	kl_followers*	kl_authors** kl_followers** informational**			hashtag*	
technology	kl_followers*	overlap_authorfriend* kl_authors* kl_followers** kl_followees** kl_friends** informational*** retweet* hashtag*	kl_friends*	kl_followees* kl_friends*	overlap_authorfollower* overlap_authorfollowee* overlap_authorfriend*	

From the results of the tests, the conclusion can be drawn that some categories are better distinguishable than other categories, and that some measures can be used for distinguishing between many categories while other measures show no significant differences at all.

5.2.1. Comparison of Categories

Not surprisingly, the category which shows the most specific usage patterns is *idioms*: The hashtags of this category can be distinguished from all hashtags just by analyzing their pragmatic properties. Also, this category shows the most significant differences overall, a total of 19 differences. Hashtag streams of the category *idioms* exhibit a significantly lower informational coverage than hashtag streams of all other categories (see Figure 5.1(a)) and a significantly higher symmetric KL divergence for author's tweet-frequency distributions (see Figure 5.1(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 5.2(c) and 5.1(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 5.2(c) and 5.1(d)). This indicates that hashtag streams in the category *technology* have a stable social structure which changes little 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

Table 5.3.: To assess the utility of different pragmatic features for differentiating between hashtag streams of different semantic categories, a pairwise Mann-Whitney-Wilcoxon-Test was conducted. This table shows which features showed a statistically significant difference (with $p < 0.05$) for each pair of categories in $t_{1\rightarrow 2}$. (significance levels: < 0.001 '***', < 0.01 '**', < 0.05 '*')

	games	idioms	movies	music	political	sports
idioms	informational** retweet**					
movies		informational*** hashtag*				
music		informational***				
political	kl_followers** retweet*	kl_authors* kl_followers*** kl_followees** kl_friends* informational*** hashtag**				
sports	kl_followers**	entropy_author* kl_authors* kl_followers** kl_followees* informational*** hashtag*				
technology	kl_followers*	kl_authors* kl_followers** kl_followees** kl_friends** informational*** retweet** hashtag*	kl_friends kl_followees*	kl_friends*	overlap_authorfollower* overlap_authorfriend*	

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*. This indicates that the content of hashtag streams of the category *political* is more likely to be produced and consumed by the same people than content of technological hashtag streams.

In general *sports*, *movies* and *music* were found to be the least distinguishable categories. *Movies* and *music* only show significant differences to the categories *idioms* and *technology*, *sports* only shows differences to *games* and *technology*.

The category pair showing the highest amount of measures with significant differences was *idioms-technology*, showing significant differences which occur in both $t_{0\to1}$ and $t_{1\to2}$ in seven measures. *Idioms-political* and *idioms-sports* were also very well distinguishable, with five and three significant differences, respectively. The category pairs where no significant differences were found for any measures in any time frame were *games-music*, *movies-music*, *movies-political*, *movies-sports*, *music-sports* and *sports-technology*.

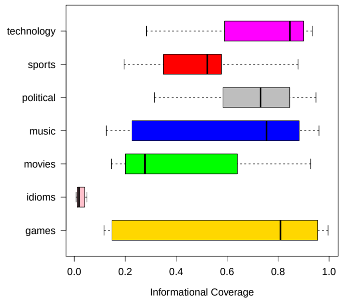
5.2.2. Comparison of Measures

Comparing the individual measures reveals that the most discriminative measures are 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). Figures 5.1 and 5.2 depict 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).

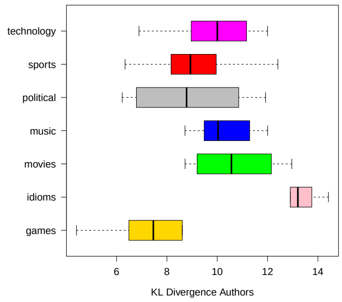
Although the overlap measures only helped to distinguish two categories (namely *technology* and *political*), it was the only significant difference that could be found for this pair of categories (for both time frames).

The entropy measures showed no significant differences occurring in both $t_{0\rightarrow 1}$ and $t_{1\rightarrow 2}$. However, in $t_{0\rightarrow 1}$ the normalized friend entropy could be used to distinguish between the categories *music* and *political*, and in $t_{1\rightarrow 2}$, the normalized author entropy could be used to distinguish between *idioms* and *sports*. The higher friend entropy of *music* in $t_{0\rightarrow 1}$ indicates that the contents of the streams in the category *music* may be consumed by a more equally distributed acknowledged audience than the content in the streams of the category *political*. The higher author entropy in the category *idioms* for $t_{1\rightarrow 2}$ might suggest that the different authors tend to contribute equally to the streams in the category *idioms*, while *sports* is dominated by selected authors.

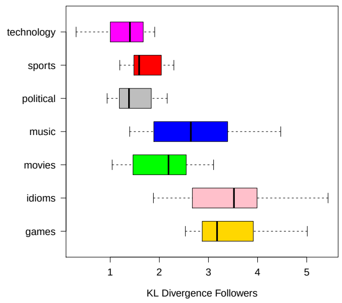
Some measures like the conversational coverage measure did not show any significant differences for any of the category pairs, for any time frame. This indicates that an equal amount of conversational activities take place in all hashtag streams.



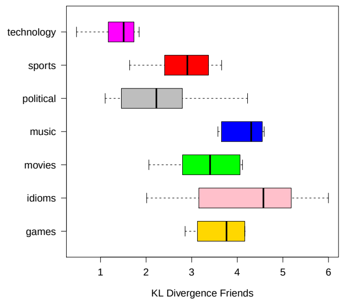
(a) Informational Coverage



(b) KL Divergence Authors

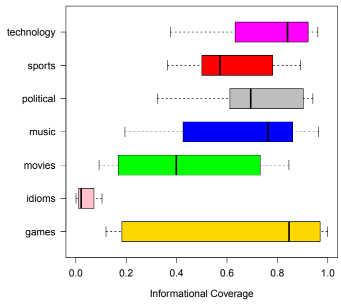


(c) KL Divergence Followers

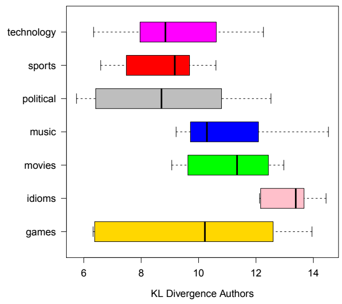


(d) KL Divergence Friends

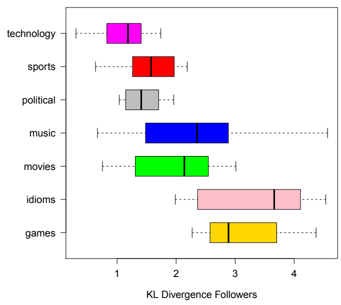
Figure 5.1.: Each plot shows the feature distribution of different categories of one of the 4 best pragmatic features for $t_{0\rightarrow 1}$. The box-plots show how each feature is distributed across hashtag streams of different categories. One can e.g. see from this figure that hashtag streams of the category idioms tend to be significantly less informational than hashtag streams of all other categories and that their authors change significantly more over time than authors of other streams. (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.



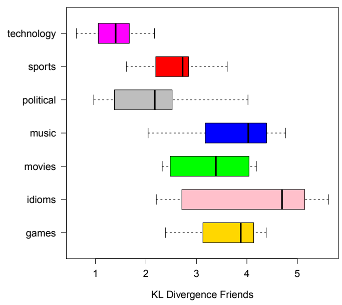
(a) Informational Coverage



(b) KL Divergence Authors



(c) KL Divergence Followers



(d) KL Divergence Friends

Figure 5.2.: Each plot shows the feature distribution of different categories of one of the 4 best pragmatic features for $t_{1\rightarrow 2}$. The box-plots show how each feature is distributed across hashtag streams of different categories. One can e.g. see from this figure that hashtag streams of the category idioms tend to be significantly less informational than hashtag streams of all other categories and that their authors change significantly more over time than authors of other streams. (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.

6. Experiment 2: Classification of Hashtag Streams

The aim of the second experiment is to explore whether social and structural properties of hashtag streams may help to automatically gauge the semantic category of hashtags. The experiment targets the second research question:

To what extent do pragmatic and lexical properties of hashtags help to predict the semantic category of a hashtag?

In order to answer this research question, hashtags were classified into their semantic categories by using pragmatic and lexical properties. To quantify the value of different pragmatic and lexical properties of hashtag streams for predicting their semantic category, a hashtag stream classification experiment was conducted and the performance of various classification models, trained with different sets of features, were systematically compared. The experiment was conducted in cooperation with Claudia Wagner and Philipp Singer.

This chapter is organized as follows: Section 6.1 describes the experimental setup, including the choice of classification methods and a description of the models. The results are reported in Section 6.2.

6.1. Experimental Setup

The aim of this experiment is to classify temporal snapshots of hashtag streams into their correct semantic categories (to which they were assigned in [RMK11]) just by analyzing how they are used over time. Then the performance of the pragmatically informed classifier is compared with the performance of a classifier informed by lexical features within this semantic multiclass classification task. Finally, the pragmatic and lexical features are combined in order to investigate whether pragmatic features may be used to supplement lexical features to achieve an improvement in performance. For this experiment, the timeframes $t_{1 \to 0}$ and $t_{2 \to 1}$ were used.

Grid search with varying hyperparameters was performed using Support Vector Machine (linear and RBF kernels) and an ensemble method with *extremely randomized trees*. Ensemble methods combine multiple separately trained classifiers (e.g., decision trees). The outputs of the individual classifiers are then combined when classifying an instance. The predictions of the ensemble of classifiers are often more accurate than the individual classifiers in the ensemble [MO11]. Even in the case that the averaging is not better than the performance of the best classifier in the ensemble, it still reduces the risk of making a poor selection, by averaging out unfortunate selections for the target variable [Pol06].

The method of *extremely randomized trees*, a computationally efficient tree-based ensemble method, was introduced by Geurts et al. [GEW06]. It can be used for classification and regression problems, and consists in strongly randomizing both the feature set and the cut-point choice when splitting a tree node. The *extremely randomized trees* method is similar to other decision tree-based ensemble method, such as *random forests*. The main difference is that *extremely randomized trees* chooses the cut-points randomly and that the whole training set is used to populate the trees, whereas *random forests* uses a subset of the training set (a bootstrap sample) to build the trees and chooses the cut-points according to the best split for the random subset of features [GEW06].

Since extremely randomized trees are a probabilistic method and perform slightly different in each run [GEW06], they were run ten times and the average scores are reported.

The features were standardized by subtracting the mean and scaling to unit variance. Stratified 6-fold cross-validation was used to train and test each classification model. 6-fold cross-validation was chosen due to the fact that the category with the least remaining active hashtags (*games*) contained six hashtags. The stratified 6-fold cross-validation ensures that in each fold, at least one hashtag of each category is in the validation set of that fold.

6.1.1. Classification Models

Since there are two different types of pragmatic features, static and dynamic ones, three separate classification models which were only informed by pragmatic information were trained and tested: a model using only static pragmatic features, a model using only dynamic pragmatic features and a combined model which uses static and dynamic pragmatic features. Additionally, a lexical model and a model which combines all pragmatic and lexical features were trained and tested. For each model, a baseline model using shuffled category labels was also created. These models are described in detail below. Additionally, Table 6.1 gives an overview of the different models, including which measures were used in the models and which timeframes they were trained and tested on.

Static Pragmatic Model: The static pragmatic classification model was trained and tested with only static pragmatic features on data collected at t_1 using stratified 6-fold cross-validation. The experiment was repeated on the data collected at t_2.

Dynamic Pragmatic Model: The dynamic pragmatic classification model was trained and tested with only dynamic pragmatic features on data collected at t_0 and t_1, using stratified 6-fold cross-validation. The computation of our dynamic features requires at least two time points. The experiment was repeated on data collected at t_1 and t_2.

Combined Pragmatic Model: The combined pragmatic classification model uses both static and dynamic pragmatic features. It was trained and tested on the data of $t_{1\to0}$ using stratified 6-fold cross-validation. Again, the experiment was repeated on the data of $t_{2\to1}$.

Lexical Model: The lexical classification model uses only the lexical features (i.e., TF weighted words). It was trained and tested on data from t_1

using stratified 6-fold cross-validation The experiment was repeated on data collected at t_2.

Combined Pragmatic and Lexical Model: Finally, a combined classification model was trained and tested, using all pragmatic and lexical features. The mixed classifier was trained and tested on the data of $t_{1\to 0}$ using stratified 6-fold cross-validation, then the experiment was repeated 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, a stacking method (see [HTF01]) was used: In the first step, a classification using lexical features alone was performed. It was trained and tested using leave-one-out cross-validation. For this first step, a SVM with linear kernel was used since it worked best for the lexical features. Secondly, the pragmatic features were combined with the resulting seven probability features which resulted from the previous classification model and which describe how likely each semantic category is for a certain stream, given its words.

Table 6.1.: Classification Models

Model	**Features**	**Time Frames**
Static Pragmatic Model	static pragmatic	t_1 t_2
Dynamic Pragmatic Model	dynamic pragmatic	t_0 to t_1 t_1 to t_2
Combined Pragmatic Model	static pragmatic, dynamic pragmatic	$t_{1\to 0}$ $t_{2\to 1}$
Lexical Model	lexical	t_1 t_2
Combined Pragmatic and Lexical Model	static pragmatic, dynamic pragmatic, lexical	$t_{1\to 0}$ $t_{2\to 1}$

Baseline Models

To get a fair baseline for the experiment, a control dataset was constructed by randomly shuffling the category labels of the 64 hashtag streams. The shuffling destroys the original relationship between the features and the semantic categories of hashtags, and the classifier attempts to use the features to classify the streams into their randomly assigned categories.

The baseline classifiers were created for each of the above classification models and were also trained and tested using stratified 6-fold cross-validation, and using grid search to determine the optimal parameters prior to training. For each baseline classifier, the random shuffling was repeated 100 times and the resulting average F1-score was used as baseline performance. The baseline classifiers test how well randomly assigned categories can be identified compared to the real semantic categories. One needs to note that a simpler random baseline, such as the *0-R* classifier [Wek13a], would be a weaker baseline than the one described above. The *0-R* classifier, which ignores the features and predicts the largest class for all instances, would result in $10/64$ correctly classified instances and a weighted average *F1-score* of 0.042.

6.1.2. Evaluation

For evaluating the classification performance, *precision*, *recall* and *F1-score*, which is the harmonic mean of precision and recall, were used as evaluation metrics. They are defined as follows [BYRN11]:

$$\text{Precision} = \frac{\text{true positives}}{\text{true positives} + \text{false positives}} \tag{6.1}$$

$$\text{Recall} = \frac{\text{true positives}}{\text{true positives} + \text{false negatives}} \tag{6.2}$$

$$\text{F1-Score} = 2 \times \frac{\text{precision} \times \text{recall}}{\text{precision} + \text{recall}} \tag{6.3}$$

6.1.3. Feature Ranking

To gain further insights into the impact of individual properties, the information gain (*IG*) of the features was analyzed with respect to the categories. The feature evaluation via information gain measures the "worth" of an attribute with respect to the class [Wek13b]. It is defined as follows (*H* denotes the entropy) [Wek13b]:

$$IG(Class, Attribute) = H(Class) - H(Class \mid Attribute) \qquad (6.4)$$

In the setting of this experiment, it measures how accurately a specific stream property P is able to predict stream's category C:

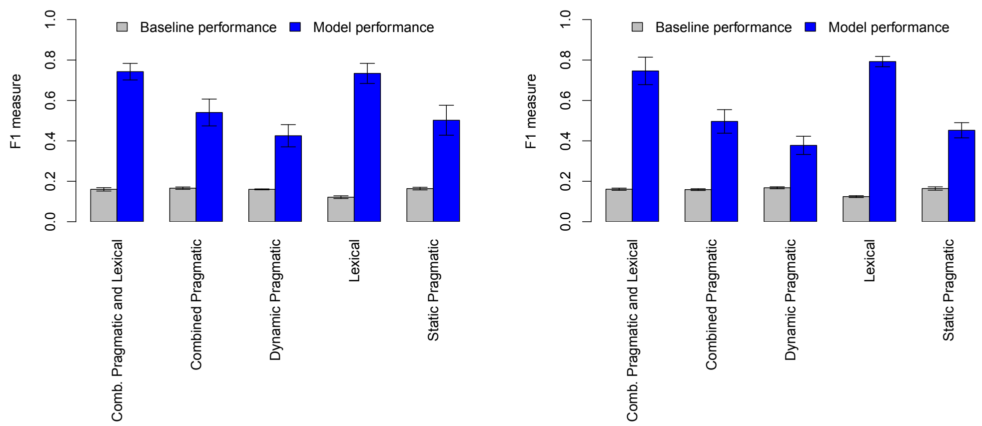
$$IG(C, P) = H(C) - H(C \mid P) \qquad (6.5)$$

6.2. Results

This section describes the results of the classification experiments and of the feature ranking.

6.2.1. Classification Results

The best performance was achieved with extremely randomized trees. Figure 6.1 shows the performance of this classifier trained with the 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 the baseline models. 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 6.1(a) we can see that for $t_{1\to0}$ the combination of lexical and pragmatic features performs slightly better than using lexical features alone. However, this overall improvement was not statistically significant and also could not be observed for $t_{2\to1}$.



(a) $t_{1\to0}$ (b) $t_{2\to1}$

Figure 6.1.: Weighted averaged F1-scores of different classification models trained and tested on $t_{1\to0}$ 6.1(a) and $t_{2\to1}$ 6.1(b) using 6-fold cross-validation. One can see from this figure that all models significantly outperform the baseline. Lexical features perform better than pragmatic features. In (a) we can see that the combination of lexical and pragmatic features performs slightly better than using lexical features alone.

6.2.2. Results by Category

Examining the F1-scores by category reveals striking differences in their performance. While the category *idioms* achieves an F1-score of 0.92 and 0.80 in $t_{1\to0}$ and $t_{2\to1}$, respectively, the category *music* has an F1-score of 0 in both time frames. This means that none of the hashtags streams of the category *music* could be correctly classified for any time frame. For this category, we can observe that the combined model performs significantly worse than the lexical model alone, demonstrating that supplementing lexical with pragmatic features can be counterproductive if the performance of pragmatic features is low for a specific category. Supplementing lexical features with the pragmatic features for categories which can be well distinguished by pragmatic features alone, on the other hand, may lead to substantial improvements in performance (in some cases), such as seen in $t_{1\to0}$ for the category *idioms*: The semantic model produces and F1-score of 0.69, while the combined model has an F1-score of 0.88.

The category *idioms*, which showed the best results of all categories in the classification experiments with pragmatic features, is not among the best distinguishable categories in the classification experiments with lexical features. This indicates that hashtags of the category *idioms* have more distinct usage patterns than other categories, but its tweets do not show more distinct lexical patterns than other categories.

Interestingly, hashtag streams of the category *games* are not well distinguishable from streams of other categories neither via their lexical fingerprint nor via their pragmatic fingerprint. This category had the lowest performance in the lexical and combined models, and the second lowest performance in the pragmatic models.

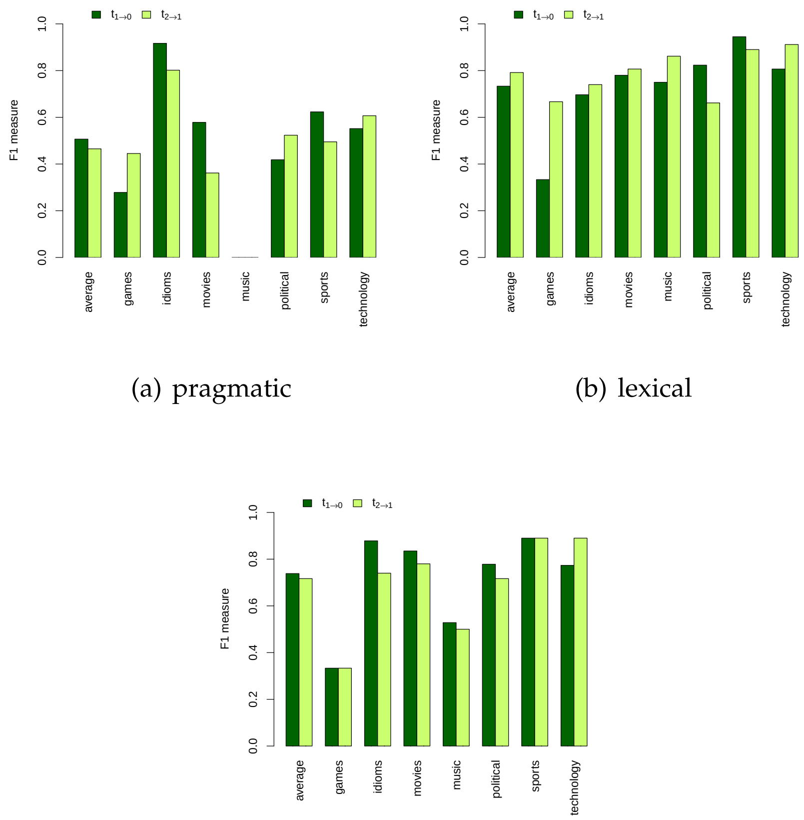
Figure 6.2 shows the performance (F1-scores) of the models by category. Precision and recall comparisons can be found in Appendix C.

6.2.3. Feature Ranking

In addition to the overall classification performance which can be achieved solely based on analyzing the pragmatics of hashtags, the impact of individual pragmatic features was also investigated. To evaluate the individual performance of the features, information gain (with respect to the categories) was used as a ranking criterion.

Table 6.2.: 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



(a) pragmatic

(b) lexical

(c) pragmatic and lexical

Figure 6.2.: Weighted averaged F1-scores of different classification models trained and tested on $t_{1\to0}$ 6.1(a) and $t_{2\to1}$ 6.1(b) using 6-fold cross-validation, by category.

The ranking was performed on $t_{1\to0}$ and $t_{2\to1}$ with stratified 6-fold cross-validation. Table 6.2 shows that the top five features (i.e., the pragmatic features which reveal most about the semantic category 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.

7. Discussion of Results

The results of this work show that pragmatic features indeed reveal information about hashtags' semantics and perform significantly better than the baseline classifiers. They can therefore be useful for the task of semantically annotating social media content. Not surprisingly, the 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.

Furthermore, the experiments show that pragmatic features may be useful for supplementing lexical features if lexical features alone are not sufficient. The results of the classification experiments show that the performance can slightly increase when combining pragmatic and lexical features. Even though the effect is not significant when using the combined performance of all categories, examining the performance of the individual categories reveals that a significant improvement may be achieved for certain categories. The reason for the lack of overall improvement might be due to the fact that, in this setup, lexical features alone already achieved good performance. In addition, the improvement of certain well-distinguishable categories is countered by the reduction in performance for other categories, where pragmatic features alone are not sufficient to create a distinction.

The classification results coincide with the results of the statistical significance tests. The category *idioms*, where most statistical significant differences were found, was also the category with the best precision, recall and F1-score in the classification experiments with pragmatic features.

The category *music*, which had the lowest precision, recall and F1-score in the classification experiments only exhibited statistically significant differences with two other categories (*idioms* and *technology*), and a difference was only found in one pragmatic feature (informational coverage and the KL divergence for friends, respectively).

The results of the statistical tests also coincide with the feature ranking via information gain. Ranking the properties by information gain showed that the most discriminative properties (the ones that showed a statistical significance in both $t_{0\to1}$ and $t_{1\to2}$ for the highest amount of category pairs) found in 5.2 were also the top ranked features (informational coverage and the KL divergences). The four pragmatic features which were in the top five ranked features for both $t_{1\to0}$ and $t_{2\to1}$ (i.e., informational coverage, hashtag coverage and the KL divergences for followers and friends) were also among the features which showed most statistically significant differences in both $t_{0\to1}$ and $t_{1\to2}$.

8. Conclusions and Implications

The results of this work suggest that the collective usage of hashtags reveals information about their semantics. 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. One of the main contributions of this work is the demonstration that semantic hashtag categories do indeed differ with respect to their usage patterns. Furthermore, this work shows how the usage information can be exploited for assigning semantic annotations to textual data streams.

Although the results show that lexical features perform 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 a wide range of different resource streams, including social video or image streams. This work has implications for researchers and practitioners interested in investigating the semantics of social media content.

8.1. Threats to Validity

Yin [Yin09] summarized four logical tests related to case study design quality, which have been commonly used to establish the quality of empirical social research. These design tests are *construct validity*, *internal validity*, *external validity* and *reliability*. This section describes potential threats to the validity of the experiments presented in this thesis, according to the four tests.

Construct validity refers *"to identifying correct operational measures for the concepts being studied"* [Yin09]. The pragmatic measures which were used in the experiments were designed to quantify a range of different usage patterns, both capturing patterns at specific points in time and describing their changes over time. The experiments presented in this thesis show that some of these pragmatic measures are well suited for distinguishing between and predicting semantic hashtag categories, while others are not. However, there may exist other pragmatic measures which could be equally, or even better, suited for characterizing usage patterns of hashtag streams. The *TF* weighting schema which was used for the lexical measures is a well established measure for describing key topics in a document [BYRN11]. While achieving the best lexical characterization of hashtag streams was not a goal of this work, other, more complex measures for characterizing the key topics might lead to a better performance of the lexical classification models.

Internal validity deals with *"seeking to establish a causal relationship"* [Yin09] and is mainly a concern for explanatory case studies. The experiments presented in this thesis do not attempt to infer causal relationships, but rather constitute a first step in exploring the idiosyncrasies and differences in usage patterns of semantic hashtag categories, and in exploring whether these patterns can be used to gauge the semantic category. Concerning the reduction of spurious statistical significances in the differences in usage patterns among semantic categories, the *Holm-Bonferroni* adjustment method was used to counteract the problem of multiple comparisons.

External validity refers to the generalizability of the study's findings, defining the *"domain to which a study's findings can be generalized"* [Yin09]. As the experiments presented in this work have focused on the investigation of hashtags on Twitter, the question whether the results are generalizable to other online platforms remains open. Suggestions for future work dealing with this aspect are given in Section 8.2.

Finally, *reliability* is defined as the *"demonstration that the operations of a study – such as the data collection procedures – can be repeated, with the same results"* [Yin09]. The experiments presented in this thesis were conducted identically on two separate timeframes, yielding similar results. This indicates that the results are independent of the time frame. However, the data for the time frames was collected in a time span of three months. Additional time frames, especially if collected after a longer period of time, would further strengthen the reliability of the experiments.

8.2. Limitations and Future Work

Further research is required to explore the relations between usage information and semantics, especially in domains where limited text is available. The experiments conducted in this work are a first step in this direction since the results show that hashtags of different semantic categories are indeed used in different ways.

Besides the potential applicability of pragmatic features to online platforms which have the distinct purpose of (exclusively) hosting multimedia streams, pragmatic features may also gain importance for semantically classifying content on Twitter. As Twitter now allows including multimedia content (photos and videos) in tweets, tweet text might lose some importance for identifying the topic of a tweet. Consider the tweet shown in Figure 8.1, which only consists of a photo and hashtags, and no ordinary text. While lexical features would not be able to draw much useful information from such a tweet, the performance of pragmatic features would not be impacted by the lack of text.

For further research, the pragmatic features presented in this work could be applied to multimedia-only platforms. But also other online platforms could be used to explore the relation between collective usage patterns of (hash)tags and their semantics. For example, Facebook recently introduced support for hashtags [Lin13], and would now constitute a suitable platform for further research on how hashtag stream usage patterns relate to their semantics. Another area where pragmatic features would have an advantage over lexical features would be in multi-lingual environments. Future research could include investigating the classification performance of models trained with pragmatic features of multi-lingual hashtag streams, and comparing it with the performance of lexical features.

Figure 8.1.: Example of a tweet which contains hashtags and media, but no ordinary text.

The results of this work are promising and show that pragmatic features allow classifying hashtag streams into their semantic categories solely based on their usage patterns. However, the results also show that for those categories which do not exhibit unique usage patterns (such as *music* or *games*), this approach does not work well. Further research in this direction would include taking into account more than seven semantic categories and exploring whether there are types of categories for which pragmatic usage patterns are especially useful.

Potential future research also includes further exploring the supplementation of lexical features with pragmatic features. For example, different approaches for combining lexical and pragmatic features could be investigated.

Appendix

Appendix A.

Hashtag Category Descriptions

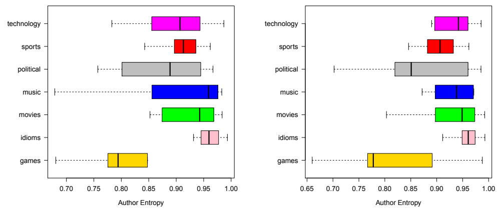
Table A.1.: This table shows definitions of the eight broad categories identified by Romero et al. [RMK11].

Category	Definition
Celebrity	The name of a person or group (e.g. music group) that is featured prominently in entertainment news. Political figures or commentators with a primarily political focus are not included. The name of the celebrity may be embedded in a longer hashtag refering to some event or fan group that involves celebrity. Note that many music groups have unusual names; these still count under the "celebrity" category.
Games	Names of computer, video, MMORPG, or Twitter-based games, as well as groups devoted to such games.
Idiom	A tag representing a conversational theme on Twitter, consisting of a concatenation of at least two common words. The concatenation can't include names of people or places, and the full phrase can't be a proper noun in itself (e.g. a title of a song/movie/organization). Names of days are allowed in the concatenation, because of the Twitter convention of forming hashtags involving names of days (e.g. MusicMonday). Abbreviations are allowed only if he full form also appears as a top hashtag (so this rules out hashtags including omg, wtf, lol, nsfw).
Movies/TV	Names of movies or TV shows, movie or TV studios, events involving a particular movie or TV show, or names of performers who have a movie or TV show specifically based around them. Names of people who have simply appeared on TV or in a movie do not count.

Music	Names of songs, albums, groups, movies or TV shows based around music, technology designed for playing music, or events involving any of these. Note that many music groups have unusual names; these still count under the “music” category.
Political	A hashtag that in your opinion often refers to a politically controversial topic. This can include a political figure, a political commentator, a political party or movement, a group on Twitter devoted to discussing a political cause, a location in the world that is the subject of controversial political discussion, or a topic or issue that is the subject of controversial political discussion. Note that this can include political hashtags oriented around countries other than the U.S.
Sports	Names of sports teams, leagues, athletes, particular sports or sporting events, fan groups devoted to sports, or references to news items specifically involving sports.
Technology	Names of Web sites, applications, or events specifically involving any of these.

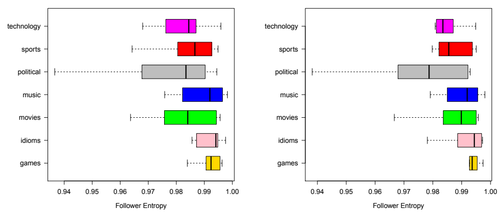
Appendix B.

Feature Distributions



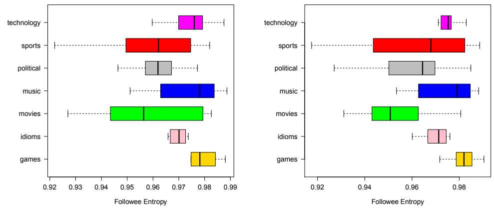
(a) $t_{0\to1}$ (b) $t_{1\to2}$.

Figure B.1.: Feature distribution of the normalized author entropy for $t_{0\to1}$ and $t_{1\to2}$.



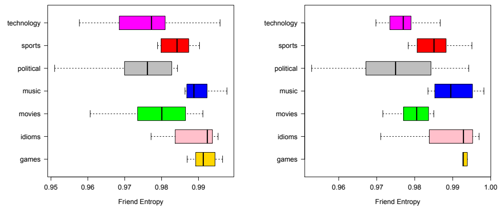
(a) $t_{0\to1}$ (b) $t_{1\to2}$.

Figure B.2.: Feature distribution of the normalized follower entropy for $t_{0\to1}$ and $t_{1\to2}$.



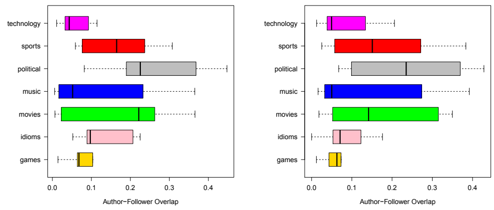
(a) $t_{0\to1}$ (b) $t_{1\to2}$.

Figure B.3.: Feature distribution of the normalized followee entropy for $t_{0\to1}$ and $t_{1\to2}$.



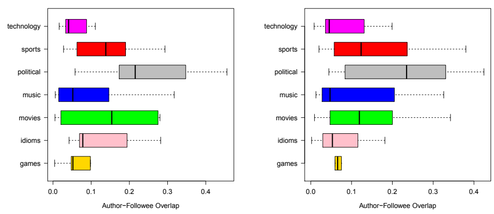
(a) $t_{0\to1}$ (b) $t_{1\to2}$.

Figure B.4.: Feature distribution of the normalized friend entropy for $t_{0\to1}$ and $t_{1\to2}$.



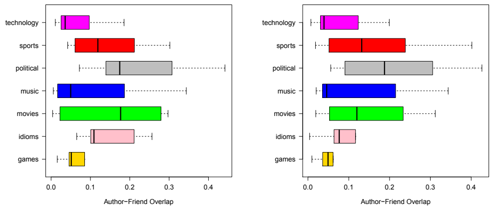
(a) $t_{0\to1}$ (b) $t_{1\to2}$.

Figure B.5.: Feature distribution of the author-follower overlap for $t_{0\to1}$ and $t_{1\to2}$.



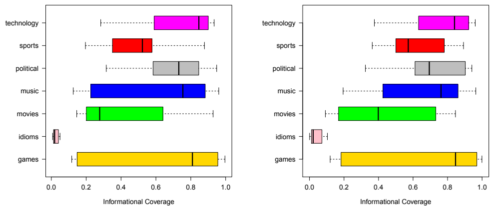
(a) $t_{0\to1}$ (b) $t_{1\to2}$.

Figure B.6.: Feature distribution of the author-followee overlap for $t_{0\to1}$ and $t_{1\to2}$.



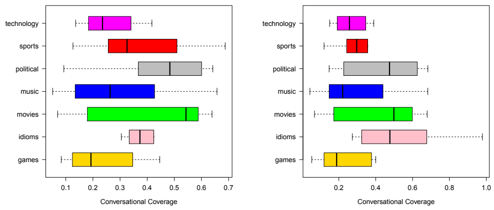
(a) $t_{0\to1}$ (b) $t_{1\to2}$.

Figure B.7.: Feature distribution of the author-friend overlap for $t_{0\to1}$ and $t_{1\to2}$.



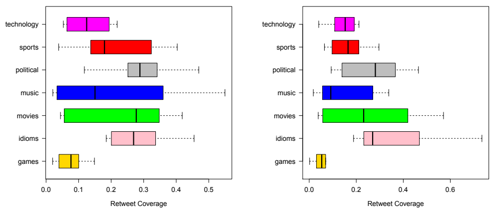
(a) $t_{0\to1}$ (b) $t_{1\to2}$.

Figure B.8.: Feature distribution of the informational coverage for $t_{0\to1}$ and $t_{1\to2}$.



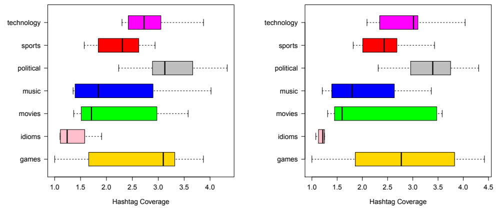
(a) $t_{0\to1}$ (b) $t_{1\to2}$.

Figure B.9.: Feature distribution of the conversational coverage for $t_{0\to1}$ and $t_{1\to2}$.



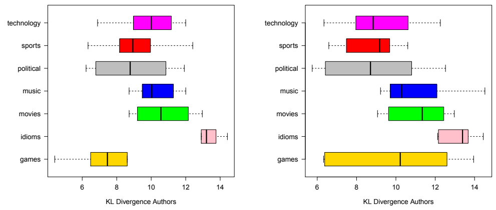
(a) $t_{0\to1}$ (b) $t_{1\to2}$.

Figure B.10.: Feature distribution of the retweet coverage for $t_{0\to1}$ and $t_{1\to2}$.



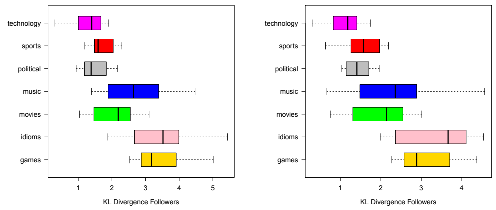
(a) $t_{0\to1}$ (b) $t_{1\to2}$.

Figure B.11.: Feature distribution of the hashtag coverage for $t_{0\to1}$ and $t_{1\to2}$.



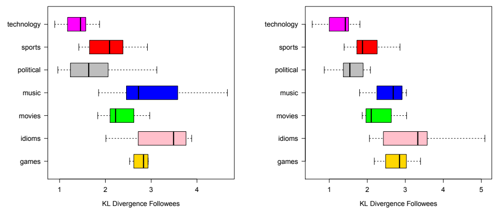
(a) $t_{0\to1}$ (b) $t_{1\to2}$.

Figure B.12.: Feature distribution of the temporal author dynamics for $t_{0\to1}$ and $t_{1\to2}$.



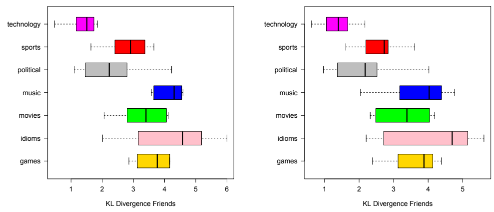
(a) $t_{0\to1}$ (b) $t_{1\to2}$.

Figure B.13.: Feature distribution of the temporal follower dynamics for $t_{0\to1}$ and $t_{1\to2}$.



(a) $t_{0\to1}$ (b) $t_{1\to2}$.

Figure B.14.: Feature distribution of the temporal followee dynamics for $t_{0\to1}$ and $t_{1\to2}$.

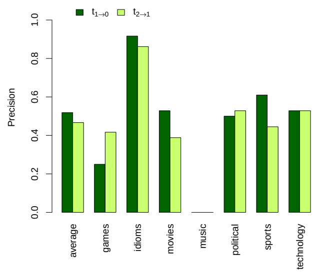


(a) $t_{0\to1}$ (b) $t_{1\to2}$.

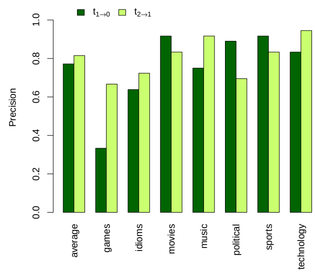
Figure B.15.: Feature distribution of the temporal friend dynamics for $t_{0\to1}$ and $t_{1\to2}$.

Appendix C.

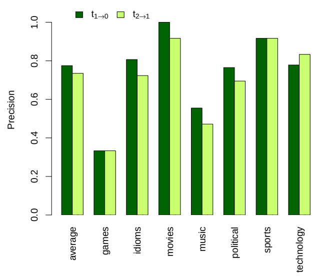
Precision and Recall by Category



(a) pragmatic

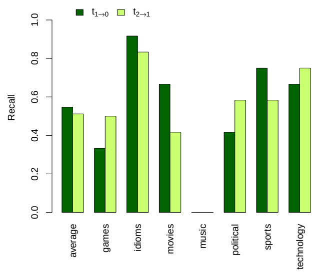


(b) lexical

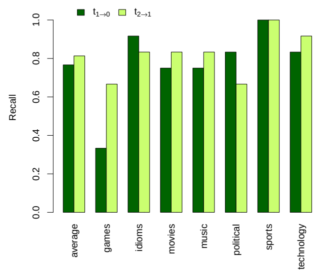


(c) pragmatic and lexical

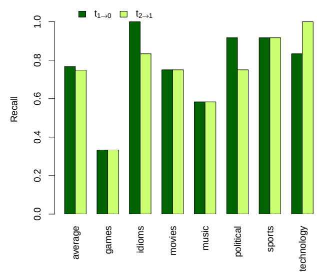
Figure C.1.: Weighted averaged precision of different classification models trained and tested on $t_{1\to0}$ and $t_{2\to1}$ using 6-fold cross-validation



(a) pragmatic



(b) lexical



(c) pragmatic and lexical

Figure C.2.: Weighted averaged recall of different classification models trained and tested on $t_{1\to0}$ and $t_{2\to1}$ using 6-fold cross-validation

Bibliography

[ACMO+04] Karl Aberer, Philippe Cudré-Mauroux, Aris M Ouksel, Tiziana Catarci, Mohand-Said Hacid, Arantza Illarramendi, Vipul Kashyap, Massimo Mecella, Eduardo Mena, Erich J Neuhold, et al. Emergent semantics principles and issues. In *Proceedings of the Database Systems for Advances Applications, 9th International Conference*, 2004.

[AK10] Satyen Abrol and Latifur Khan. Tweethood: Agglomerative clustering on fuzzy k-closest friends with variable depth for location mining. In *Proceedings of the 2010 IEEE 2nd International Conference on Social Computing*, 2010.

[Bac99] Kent Bach. The semantics-pragmatics distinction: What is it and why it matters. *The Semantics/Pragmatics Interface from Different Points of View*, 1999.

[BB11] Axel Bruns and Jean E. Burgess. The use of twitter hashtags in the formation of ad hoc publics. In *Proceedings of the 6th European Consortium for Political Research General Conference*, 2011.

[BBCS12] Axel Bruns, Jean E Burgess, Kate Crawford, and Frances Shaw. #qldfloods and qpsmedia: Crisis communication on twitter in the 2011 south east queensland floods. *ARC Centre of Excellence for Creative Industries and Innovation, Queensland University of Technology*, 2012.

[BKH+11] Dominik Benz, Christian Körner, Andreas Hotho, Gerd Stumme, and Markus Strohmaier. One tag to bind them all: Measuring term abstractness in social metadata. In *Proceedings of the 8th Extended Semantic Web Conference*, 2011.

[BS12] Axel Bruns and Stefan Stieglitz. Quantitative approaches to comparing communication patterns on twitter. *Journal of Technology in Human Services*, 2012.

[BSGHP06] Amanda C Brandone, Sara J Salkind, Roberta Michnick Golinkoff, and Kathy Hirsh-Pasek. *Language Development*. National Association of School Psychologists, 2006.

[Bur12] Jean Burgess. The iphone moment, the apple brand and the creative consumer : From "hackability and usability" to cultural generativity. In *Studying Mobile Media: Cultural Technologies, Mobile Communication, and the iPhone*. Routledge, 2012.

[BYRN11] Ricardo Baeza-Yates and Berthier Ribeiro-Neto. *Modern information retrieval*. Pearson Education Limited, 2011.

[CBHS08] Ciro Cattuto, Dominik Benz, Andreas Hotho, and Gerd Stumme. Semantic grounding of tag relatedness in social bookmarking systems. In *Proceedings of the 7th International Semantic Web Conference*, 2008.

[CHBG10] Meeyoung Cha, Hamed Haddadi, Fabricio Benevenuto, and Krishna P Gummadi. Measuring user influence in twitter: The million follower fallacy. In *Proceedings of the 4th International AAAI Conference on Weblogs and Social Media*, 2010.

[CKJ11] Iván Cantador, Ioannis Konstas, and Joemon M Jose. Categorising social tags to improve folksonomy-based recommendations. *Web Semantics: Science, Services and Agents on the World Wide Web*, 2011.

[CL06] Fan Rong K Chung and Linyuan Lu. *Complex graphs and networks*. American Mathematical Soc., 2006.

[CMC+11] Evandro Cunha, Gabriel Magno, Giovanni Comarela, Virgilio Almeida, Marcos André Gonçalves, and Fabrício Benevenuto. Analyzing the dynamic evolution of hashtags on twitter: a language-based approach. In *Proceedings of the Workshop on Language in Social Media*, 2011.

[CMG09] Meeyoung Cha, Alan Mislove, and Krishna P Gummadi. A measurement-driven analysis of information propagation in the flickr social network. In *Proceedings of the 18th International Conference on World Wide Web*, 2009.

[CMS10] Bernat Corominas-Murtra and Ricard V Solé. Universality of zipf's law. *Physical Review E*, 2010.

[Con98] James Conant. Wittgenstein on meaning and use. *Philosophical Investigations*, 1998.

[CR93] B Cialdini Robert. Influence: Science and practice. *New York, NY: Harper Collins*, 1993.

[CTW11] S. Carter, E. Tsagkias, and W. Weerkamp. Twitter hashtags: Joint translation and clustering. In *Proceedings of the 3rd International Conference on Web Science*, 2011.

[Cut02] Joan Cutting. *Pragmatics and discourse: A resource book for students*. Psychology Press, 2002.

[DK09] Carsten F Dormann and Ingolf Kühn. Angewandte statistik für die biologischen wissenschaften. *Helmholtz Zentrum für Umweltforschung-UFZ*, 2009.

[Dow11] Allen B Downey. *Think stats*. O'Reilly Media, 2011.

[EMS+10] Martin Ebner, Herbert Mühlburger, Sandra Schaffert, Mandy Schiefner, Wolfgang Reinhardt, and Steve Wheeler. Getting granular on twitter: tweets from a conference and their limited usefulness for non-participants. In *Key Competencies in the Knowledge Society*. Springer, 2010.

[Fas90] Ralph W Fasold. *The Sociolinguistics of Language: An Introduction to Sociolinguistics*. B. Blackwell, 1990.

[FRH11] Victoria Fromkin, Robert Rodman, and Nina M Hyams. *An introduction to language*. Wadsworth Publishing Company, 2011.

[G+95] Thomas R Gruber et al. Toward principles for the design of ontologies used for knowledge sharing. *International journal of human computer studies*, 1995.

[GEW06] Pierre Geurts, Damien Ernst, and Louis Wehenkel. Extremely randomized trees. *Machine learning*, 2006.

[GH06] Scott Golder and Bernardo A. Huberman. The structure of collaborative tagging systems. *Journal of Information Science*, 2006.

[Hal13] Harry Halpin. *Social Semantics - The Search for Meaning on the Web*. Springer, 2013.

[HD10] Liangjie Hong and Brian D. Davison. Empirical study of topic modeling in twitter. In *Proceedings of the IGKDD Workshop on Social Media Analytics*, 2010.

[HGM06] Paul Heymann and Hector Garcia-Molina. Collaborative creation of communal hierarchical taxonomies in social tagging systems. Technical report, Computer Science Department, 2006.

[HH08] Patrick J Hayes and Harry Halpin. In defense of ambiguity. *International Journal on Semantic Web and Information Systems (IJSWIS)*, 2008.

[HHB12] Tim Highfield, Stephen Harrington, and Axel Bruns. Twitter as a technology for audiencing and fandom: the# eurovision phenomenon. *Information, Communication & Society*, 2012.

[HL06] Thomas Hill and Paweĺ Lewicki. *Statistics: methods and applications: a comprehensive reference for science, industry, and data mining*. StatSoft, Inc., 2006.

[Hol79] Sture Holm. A simple sequentially rejective multiple test procedure. *Scandinavian journal of statistics*, 1979.

[HRS07] Harry Halpin, Valentin Robu, and Hana Shepherd. The complex dynamics of collaborative tagging. In *Proceedings of the 16th international conference on World Wide Web*, 2007.

[HRW09] Bernardo A. Huberman, Daniel M. Romero, and Fang Wu. Social networks that matter: Twitter under the microscope. *First Monday*, 2009.

[HT09] Harry Halpin and Henry S. Thompson. Social Meaning on the Web: From Wittgenstein to Search Engines. *IEEE Intelligent Systems*, 2009.

[HTE10] Jeff Huang, Katherine M. Thornton, and Efthimis N. Efthimiadis. Conversational tagging in twitter. In *Proceedings of the 21st ACM Conference on Hypertext and Hypermedia*, 2010.

[HTF01] Trevor Hastie, Robert Tibshirani, and Jerome Friedman. *The Elements of Statistical Learning*. Springer New York Inc., 2001.

[JS01] Don H Johnson and Sinan Sinanovic. Symmetrizing the kullback-leibler distance. *IEEE Transactions on Information Theory*, 2001.

[JSFT07] Akshay Java, Xiaodan Song, Tim Finin, and Belle Tseng. 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*, 2007.

[KBS+10] Christian Körner, Dominik Benz, Markus Strohmaier, Andreas Hotho, and Gerd Stumme. Stop thinking, start tagging - tag semantics emerge from collaborative verbosity. In *Proceedings of the 19th International Conference on World Wide Web*, 2010.

[KHMS11] Jan H Kietzmann, Kristopher Hermkens, Ian P McCarthy, and Bruno S Silvestre. Social media? get serious! understanding the functional building blocks of social media. *Business Horizons*, 2011.

[KJMO10] Dongwoo Kim, Yohan Jo, Il-Chul Moon, and Alice Oh. Analysis of twitter lists as a potential source for discovering latent characteristics of users. In *Proceedings of the Workshop* *on Microblogging at the ACM Conference on Human Factors in Computer Systems*, 2010.

[KLPM10] Haewoon Kwak, Changhyun Lee, Hosung Park, and Sue Moon. What is twitter, a social network or a news media? In *Proceedings of the 19th International Conference on World wide Web*, 2010.

[KNT10] Ravi Kumar, Jasmine Novak, and Andrew Tomkins. Structure and evolution of online social networks. In *Link Mining: Models, Algorithms, and Applications*. Springer, 2010.

[LBK09] Jure Leskovec, Lars Backstrom, and Jon Kleinberg. Meme-tracking and the dynamics of the news cycle. In *Proceedings of the 15th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, 2009.

[LGRC12] Janette Lehmann, Bruno Gonçalves, José J Ramasco, and Ciro Cattuto. Dynamical classes of collective attention in twitter. In *Proceedings of the 21st International Conference on World Wide Web*, 2012.

[Lin13] Greg Lindley. http://newsroom.fb.com/News/633/Public-Conversations-on-Facebook, 2013. [Online; accessed on 2013-08-19].

[Liv04] Paul Livingston. 'meaning is use'in the tractatus. *Philosophical investigations*, 2004.

[LM10] David Laniado and Peter Mika. Making sense of twitter. In *Proceedings of the 9th International Semantic Web Conference*, 2010.

[LPBD10] Julie Letierce, Alexandre Passant, John Breslin, and Stefan Decker. Understanding how twitter is used to widely spread scientific messages. In *Proceedings of the Web Science Conference 2010: Extending the Frontiers of Society On-Line*, 2010.

[LSAA11] Avishay Livne, Matthew P Simmons, Eytan Adar, and Lada A Adamic. The party is over here: Structure and content in the 2010 election. *ICWSM. 5th*, 2011.

[Mac12] Sofus A. Macskassy. On the study of social interactions in twitter. In *Proceedings of the 6th International AAAI Conference on Weblogs and Social Media*, 2012.

[MCM+09] Benjamin Markines, Ciro Cattuto, Filippo Menczer, Dominik Benz, Andreas Hotho, and Gerd Stumme. Evaluating similarity measures for emergent semantics of social tagging. In *Proceedings of the 18th International Conference on World Wide Web*, 2009.

[Mes07a] Chris 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. [Online; accessed on 2013-08-19].

[Mes07b] Chris Messina. New microsyntax for twitter: three pointers and the slasher. http://factoryjoe.com/blog/2009/11/08/slashtags/, 2007. [Online; accessed on 2013-08-19].

[Mes07c] Chris Messina. Twitter hashtags for emergency coordination and disaster relief. http://factoryjoe.com/blog/2007/10/22/twitter-hashtags-for-emergency-coordination-and-disaster-relief/, 2007. [Online; accessed on 2013-08-19].

[Mik07a] Peter Mika. Ontologies are us: A unified model of social networks and semantics. *Web Semantics*, 2007.

[Mik07b] Alexander Mikroyannidis. Toward a social semantic web. *Computer*, 2007.

[MM10] Matthew Michelson and Sofus A. Macskassy. Discovering users' topics of interest on twitter: a first look. In *Proceedings of the 4th Workshop on Analytics for Noisy Unstructured Text Data*, 2010.

[MO11] Richard Maclin and David Opitz. Popular ensemble methods: An empirical study. *Journal Of Artificial Intelligence Research*, 2011.

[MSLC01] Miller McPherson, Lynn Smith-Lovin, and James M Cook. Birds of a feather: Homophily in social networks. *Annual review of sociology*, 2001.

[Mur11] Dhiraj Murthy. Twitter: Microphone for the masses? *Media, Culture & Society*, 2011.

[NBL10] Mor Naaman, Jeffrey Boase, and Chih-Hui Lai. Is it really about me?: message content in social awareness streams. In *Proceedings of the 2010 ACM Conference on Computer Supported Cooperative Work*, 2010.

[NM08a] Michael G. Noll and Christoph Meinel. Exploring social annotations for web document classification. In *Proceedings of the 23rd Annual ACM Symposium on Applied Computing*, 2008.

[NM08b] Michael G. Noll and Christoph Meinel. The metadata triumvirate: Social annotations, anchor texts and search queries. In *Proceedings of the 7th IEEE/WIC/ACM International Conference on Web Intelligence*, 2008.

[oP12] Stanford Encyclopedia of Philosophy. http://plato.stanford.edu/entries/pragmatics/, August 2012. [Online; accessed on 2013-08-19].

[Ore07] Tim Oreilly. What is web 2.0: Design patterns and business models for the next generation of software. *Communications & strategies*, 2007.

[OSvZ09] Simon Overell, Börkur Sigurbjörnsson, and Roelof van Zwol. Classifying tags using open content resources. In *Proceedings of the 2nd ACM International Conference on Web Search and Data Mining*, 2009.

[Per10] Sarah Perez. http://readwrite.com/2010/09/14/twitter_is_not_a_social_network_says_twitter_exec, 2010. [Online; accessed on 2013-08-19].

[PLG10] Anon Plangprasopchok, Kristina Lerman, and Lise Getoor. Growing a tree in the forest: constructing folksonomies by integrating structured metadata. In *Proceedings of the 16th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, 2010.

[Pol06] Robi Polikar. Ensemble based systems in decision making. *Circuits and Systems Magazine, IEEE*, 2006.

[PT96] Stefano Panzeri and Alessandro Treves. Analytical estimates of limited sampling biases in different information measures. *Network: Computation in Neural Systems*, 1996.

[PWSS13] Lisa Posch, Claudia Wagner, Philipp Singer, and Markus Strohmaier. Meaning as collective use: predicting semantic hashtag categories on twitter. In *Proceedings of the 22nd international conference on World Wide Web Companion*, 2013.

[RDL10] Daniel Ramage, Susan Dumais, and Dan Liebling. Characterizing microblogs with topic models. In *Proceedings of the 4th International AAAI Conference on Weblogs and Social Media*, 2010.

[Reh10] Sabina Rehman. Wittgenstein's language-games, stoppard's building-blocks and context-learning in a corpus. *Skase Journal of Literary Studies*, 2010.

[RMK11] Daniel M. Romero, Brendan Meeder, and Jon 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*, 2011.

[Ros09] Sheldon M Ross. *Introduction to probability and statistics for engineers and scientists*. Academic Press, 2009.

[Roy95] Patrick Royston. Remark as r94: A remark on algorithm as 181: The w-test for normality. *Journal of the Royal Statistical Society. Series C (Applied Statistics)*, 1995.

[RTU13] Daniel M. Romero, Chenhao Tan, and Johan Ugander. On the interplay between social and topical structure. In *Proceedings of the 7th International AAAI Conference on Weblogs and Social Media*, 2013.

[SBBP12] Chris Sumner, Alison Byers, Rachel Boochever, and Gregory J. Park. Predicting dark triad personality traits from twitter usage and a linguistic analysis of tweets. In *Proceedings of the 11th International Conference on Machine Learning and Applications*, 2012.

[Sch06] Patrick Schmitz. Inducing ontology from flickr tags. In *Proceedings of the Workshop on Collaborative Tagging at the 15th International Conference on World Wide Web*, 2006.

[Sem12] Semiocast. http://semiocast.com/en/publications/2012_07_30_Twitter_reaches_half_a_billion_accounts_140m_in_the_US, 2012. [Online; accessed on 2013-08-19].

[SFD+10] Bharath Sriram, Dave Fuhry, Engin Demir, Hakan Ferhatosmanoglu, and Murat Demirbas. 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*, 2010.

[SHB+12] Markus Strohmaier, Denis Helic, Dominik Benz, Christian Körner, and Roman Kern. Evaluation of folksonomy induction algorithms. *Transactions on Intelligent Systems and Technology*, 2012.

[Shl07] Jonathon Shlens. Notes on kullback-leibler divergence and likelihood theory. *System Neurobiology Laboratory, Salk Institute for Biological Studies, California*, 2007.

[Sim55] Herbert A Simon. On a class of skew distribution functions. *Biometrika*, 1955.

[SKB+11] Jurgis Skilters, Monika Kreile, Uldis Bojars, Inta Brikse, Janis Pencis, and Laura Uzule. The pragmatics of political messages in twitter communication. In *Proceedings of the 8th Extended Semantic Web Conference*, 2011.

[SKK10] Markus Strohmaier, Christian Körner, and Roman Kern. Why do users tag? detecting users' motivation for tagging in social tagging systems. In *Proceedings of the International AAAI Conference on Weblogs and Social Media*, 2010.

[SLR+06] Shilad Sen, Shyong K. Lam, Al Mamunur Rashid, Dan Cosley, Dan Frankowski, Jeremy Osterhouse, F. Maxwell Harper, and John Riedl. tagging, communities, vocabulary, evolution. In *Proceedings of the 20th Anniversary Conference on Computer Supported Cooperative Work*, 2006.

[Soc13] American Dialect Society. http://www.americandialect.org/hashtag-2012, 2013. [Online; accessed on 2013-08-19].

[SP11] Kate Starbird and Leysia Palen. Voluntweeters: Self-organizing by digital volunteers in times of crisis. In *Proceedings of the 2011 Annual Conference on Human factors in Computing SSystems*, 2011.

[TR12] Oren Tsur and Ari Rappoport. What's in a hashtag?: content based prediction of the spread of ideas in microblogging communities. In *Proceedings of the 5th ACM International Conference on Web Search and Data Mining*, 2012.

[Tra09] Jennifer Trant. Studying Social Tagging and Folksonomy: A Review and Framework. *Journal of Digital Information*, 2009.

[Twi08] Twitter. https://blog.twitter.com/2008/twitter-trends-tip, 2008. [Online; accessed on 2013-08-19].

[Twi09] Twitter. https://blog.twitter.com/2009/whats-happening, 2009. [Online; accessed on 2013-08-19].

[Twi10] Twitter. https://blog.twitter.com/2010/measuring-tweets, 2010. [Online; accessed on 2013-08-19].

[Twi12] Twitter. http://blog.twitter.com/2012/03/twitter-turns-six.html, 2012. [Online; accessed on 2013-08-19].

[Twi13a] Twitter. https://dev.twitter.com/docs, 2013. [Online; accessed on 2013-08-19].

[Twi13b] Twitter. https://twitter.com/about, 2013. [Online; accessed on 2013-08-19].

[Twi13c] Twitter. https://business.twitter.com/twitter-101, 2013. [Online; accessed on 2013-08-19].

[Twi13d] Twitter. https://business.twitter.com/whos-twitter, 2013. [Online; accessed on 2013-08-19].

[Twi13e] Twitter. https://dev.twitter.com/docs/api/1/get/search, 2013. [Online; accessed on 2013-08-19].

[Twi13f] Twitter. https://dev.twitter.com/docs/api/1, 2013. [Online; accessed on 2013-08-19].

[Twi13g] Twitterizer. https://github.com/Twitterizer/Twitterizer, 2013. [Online; accessed on 2013-08-19].

[UMA08] María Dolores Ugarte, Ana F Militino, and Alan T Arnholt. *Probability and Statistics with R*. CRC Press, 2008.

[Vie10] Sarah Vieweg. The ethics of twitter research. In *Proceedings of the Workshop on Revisiting Ethics in the Facebook Era: Challenges in Emerging CSCW Research*, 2010.

[VW07] Thomas Vander Wal. Folksonomy coinage and definition. http://vanderwal.net/folksonomy.html, 2007. [Online; accessed on 2013-08-19].

[Wek13a] Weka. http://weka.wikispaces.com/ZeroR, 2013. [Online; accessed on 2013-08-19].

[Wek13b] Weka. http://weka.sourceforge.net/doc.dev/weka/attributeSelection/InfoGainAttributeEval.html, 2013. [Online; accessed on 2013-08-19].

[WHMW11] Shaomei Wu, Jake M. Hofman, Winter A. Mason, and Duncan J. Watts. Who says what to whom on twitter. In *Proceedings of the 20th International Conference on World Wide Web*, 2011.

[Wil99] Christopher J Wild. *Chance encounters: A first course in data analysis and inference*. John Wiley & Sons, 1999.

[Wit21] Ludwig Wittgenstein. *Tractatus logico-philosophicus*. Routledge & Paul, 1921.

[Wit53] Ludwig Wittgenstein. *Philosophical Investigations*. Blackwell Publishers, 1953.

[Wit10] Ludwig Wittgenstein. *Philosophical investigations*. Wiley-Blackwell, 2010.

[WNHT09] Xiao Wu, Chong-Wah Ngo, Alexander G Hauptmann, and Hung-Khoon Tan. Real-time near-duplicate elimination for web video search with content and context. *IEEE Transactions on Multimedia*, 2009.

[WpLJH10] Jianshu Weng, Ee peng Lim, Jing Jiang, and Qi He. Twitterrank: Finding topic-sensitive influential twitterers. In *Proceedings of the 3rd ACM International Conference on Web Search and Data Mining*, 2010.

[WS10] Claudia Wagner and Markus Strohmaier. The wisdom in tweetonomies: Acquiring latent conceptual structures from social awareness streams. In *Proceedings of the Semantic Search Workshop at the 19th International Conference on World Wide Web*, 2010.

[WSPS13] Claudia Wagner, Philipp Singer, Lisa Posch, and Markus Strohmaier. The wisdom of the audience: An empirical study of social semantics in twitter streams. In *The Semantic Web: Semantics and Big Data*. Springer, 2013.

[WZO10] Wei Wu, Bin Zhang, and Mari Ostendorf. Automatic generation of personalized annotation tags for twitter users. In

Proceedings of the 2010 Annual Conference of the North American Chapter of the Association for Computational Linguistics, 2010.

[Yin09] Robert K Yin. *Case study research: Design and methods*. Sage, 2009.

[YL10] Jaewon Yang and Jure Leskovec. Modeling information diffusion in implicit networks. In *Proceedings of the 2010 IEEE International Conference on Data Mining*, 2010.

[YSZM12] Lei Yang, Tao Sun, Ming Zhang, and Qiaozhu Mei. We know what @you #tag: does the dual role affect hashtag adoption? In *Proceedings of the 21st International Conference on World Wide Web*, 2012.

[Zeu09] Carsten Zeumer. `http://www.codeproject.com/Articles/43198/Detect-a-written-text-s-language`, 2009. [Online; accessed on 2013-08-19].

[ZR09] Dejin Zhao and Mary Beth Rosson. How and why people twitter: the role that micro-blogging plays in informal communication at work. In *Proceedings of the ACM 2009 International Conference on Supporting Group Work*, 2009.