Elias Zeitfogel

Analysis of User Attention on Reddit

Master Thesis

Graz University of Technology

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

Supervisor: Univ.-Doz. Dipl-Ing. Dr.techn. Markus Strohmaier
Advisor: Dipl.-Ing. Philipp Singer

Graz, April 2014

Elias Zeitfogel

Benutzeraufmerksamkeitsanalyse auf Reddit

Masterarbeit

Technische Universität Graz

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

Begutachter : Univ.-Doz. Dipl-Ing. Dr.techn. Markus Strohmaier
Betreuer: Dipl.-Ing. Philipp Singer

Graz, April 2014

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 _____ _____

Datum Unterschrift

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

Abstract

The social news and content aggregator reddit has grown rapidly in recent years. Content on reddit is submitted exclusively by its users. These submissions can be voted and commented upon and are ranked accordingly to be displayed on the self-proclaimed "front page of the internet". The user interaction and attention that submissions receive is the main focus of this thesis. These attention patterns represent an intriguing field of study on reddit, analyzing them allows inferences regarding influencing factors for user behavior and provides a deeper understanding of what drives social interaction and shapes the dynamics of the social network. This thesis aims to explain the development and relations of the attention patterns of reddit's community. To that end, several indicators of attention are identified and thoroughly investigated in various dimensions, including a content categorization scheme. After establishing general characteristics of each indicator, it is revealed that the nature of attention varies greatly for different topics and types of content. A correlation experiment confirms these observations. A temporal analysis indicates a growing topical diversification that is broadly appreciated by the community, but also an increasingly excessive attraction of attention by certain kinds of content. Furthermore explicit and implicit factors influencing the attention patterns are determined. A classification experiment on submission titles hints at the formation of customary idioms and utilization of specific language in reddit's sub-communities. The titles are additionally exploited for expressive trend discovery and analysis. This work should lead to a better understanding of attention patterns on reddit and builds a foundation for subsequent research on user attention and interaction in systems with social collaborative filtering.

Kurzfassung

Reddit ist ein soziales Netzwerk, das der Aggregierung von Inhalten, die durch Benutzer eingereicht werden, dient. Die Nutzer von reddit können über jeglichen Inhalt abstimmen und ihn kommentieren. Aus dieser Interaktion wird eine Rangliste erstellt und auf der selbst ernannten "front page of the internet" prominent zur Schau gestellt. Diese Interaktion und Aufmerksamkeit, die Inhalte erfahren, stellen ein faszinierendes Forschungsgebiet auf reddit dar. Ihre Analyse ermöglicht Rückschlüsse auf Einflussfaktoren auf Nutzerverhalten und verschafft ein tieferes Verständnis darüber, was soziale Interaktion antreibt und dieses Netzwerk formt. Diese Arbeit versucht daher die Entwicklung und Beziehungen dieser Aufmerksamkeitsmuster besser zu erklären. Dazu werden verschiedene Aufmerksamkeitsindikatoren identifziert und untersucht. Es wird erkannt, dass die Erscheinungsform der Aufmerksamkeitsmuster je nach Thema und Art des Inhalts stark variieren kann. Ein Korrelationsexperiment bestätigt diese Beobachtungen. Durch eine zeitliche Analyse zeigt sich zudem eine wachsende thematische Diversität, die von den Benutzern sehr begrüßt wird. Andererseits konzentriert sich jedoch die meiste Aufmerskamkeit vornehmlich nur mehr auf wenige Arten von Inhalt. Zusätzlich wurden explizite und implizite Faktoren bestimmt, die diese Aufmerksamkeitsmuster beeinflussen. Ein Klassifikationsexperiment deutet auf die Bildung von eigenen Begriffen und spezifischer Ausdrucksweise in den einzelnen Untergemeinschaften der Plattform hin. Außerdem werden Ansätze für die Entdeckung von Trends und deren Analyse vorgestellt. Diese Arbeit soll zu einem besseren Verständnis der Aufmerksamkeitsmuster auf reddit führen und bildet einen Grundstein für weitere Forschung über Aufmerksamkeit und Interaktion von Benutzern in Systemen, die soziales kollaboratives Filtern implementieren.

Acknowledgements

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

I am very grateful for all the support, help and advice I received along the way of creating this thesis.

First of all, I want to thank my supervisor Dr. Markus Strohmaier for his ideas and inspirations and for being constantly available with advice and help, despite the long distances. Without his thoughtful guidance this work would not have been possible.

I wish to express my sincere gratitude for my academic advisor Dipl.Ing. Philipp Singer, whose frequent feedback during the research and writing process proved invaluable and who spent his free time thinking through and correcting my ideas.

For the additional input and the fruitful collaboration on the research paper "Evolution of Reddit: From the Front Page of the Internet to a Self-referential Community?", I want to thank my co-authors Philipp Singer, Fabian Flöck, Clemens Meinhart and Markus Strohmaier.

I wish to thank Clemens Meinhart for all the fun that was had during our studies, for the nights of coding and the countless conversations, but especially for his genuine friendship.

Finally, I want to thank my girlfriend Sophie for her patience and the constant motivation and my family for the never-ending support and providing me with the opportunity to carry out my studies.

Thank you!

Graz, April 2014 Elias Zeitfogel

Contents

List of Figures

List of Tables

Abbreviations

AI	**A**rtificial **I**ntelligence
AMA	**A**sk **M**e **A**nything
API	**A**pplication **P**rogramming **I**nterface
BNB	**B**inary **N**aive **B**ayes
CISPA	**C**yber **I**ntelligence **S**haring and **P**rotection **A**ct
DL	**D**iscussion **L**ist
EDT	**E**astern **D**aylight **T**ime
F2P	**F**ree **2**(to) **P**lay
GIF	**G**raphics **I**nterchange **F**ormat
IDF	**I**nverse **D**ocument **F**requency
MNB	**M**ultinomial **N**aive **B**ayes
NSFW	**N**ot **S**afe **F**or **W**ork
PDT	**P**acific **D**aylight **T**ime
PIPA	**P**ROTECT **I**ntellectual **P**roperty **A**ct
PPMC	**P**earson **P**roduct-**M**oment **C**orrelation
PRAW	**P**ython **R**eddit **A**PI **W**rapper
Q&A	**Q**uestion & **A**nswer
Redditor	**R**eddit **E**ditor
RES	**R**eddit **E**nhancement **S**uite
SOPA	**S**top **O**nline **P**iracy **A**ct
SROC	**S**pearman's **R**ank-**O**rder **C**orrelation
TF	**T**erm **F**requency
TF-IDF	**T**erm **F**requency - **I**nverse **D**ocument **F**requency
TIL	**T**oday **I** **L**earned
URL	**U**niform **R**esource **L**ocator
UTC	**T**emps **U**niversel **C**oordonné

1. Introduction

1.1. Motivation

Reddit was founded in 2005 by Steve Huffman and Alexis Ohanian. The small two-student-start-up was funded by Ycombinator[1] and has come a far way since. The self-titled "front page of the internet" has evolved into one of the largest online community portals in the world. The site currently exceeds 100 million unique visitors from over 196 countries each month, according to reddit (2013) itself.

On reddit[2], everyone can submit stories, link to other websites and engage in discussions on virtually any topic. Users are able to vote to express their appreciation or aversion for any content or comment, which directly shifts the ranking and placement on the site. Many other social networking sites, too, have some sort of mechanism to show support or liking for something (Facebook's *like*-button for example). But the possibility to actively display and perform dislike of virtually any content on the site is one aspect decisively separating reddit from any other online community. A negative voice arguable has a much stronger impact here, since reddit's ranking algorithm takes both positive and negative votes into account. Another important differentiating factor is expressed by the users autonomy and independence from fixed and predetermined structures through the ability to create their own sub-forums and communities.

The commercial as well as the political world have recognized reddit's value as a platform for sharing novel campaigns and promoting

1 `http://ycombinator.com/`

2 The name is uncapitalized as mandated by reddit's trademark guidelines found at `http://reddit.com/about/alien/`

their upcoming products. Well known Hollywood actors, like Keanu Reeves (2013) or Tom Hanks (2013), advertise their new movies and answer questions of users in return. One of the most popular entries on reddit was created by Barack Obama (2012) as part of his campaign for the presidential election in 2012.

Reddit's social community has already shown much interest in political and economical matters, eclipsing in a scheduled blackout of the site on January the 18th, 2012, in demonstration of the Stop Online Piracy Act and PROTECT Intellectual Property Act in the United States. This was a result of a movement started by countless reddit users which even organized themselves on the community portal in newly created sub-channels.[3]

The structure and mechanics of reddit, which are described in more detail in chapter 2, make its content easy distinguishable. The site also features a simple voting mechanism for readers to decide which parts of its content are worthwhile and which can be dismissed. Paired with a steady stream of new content and the ability to comment and discuss everything, the online portal has developed into an interesting field for intensive research regarding social interaction, user participation and attention patterns.

Attention patterns are especially intriguing. Analyzing these patterns allows inferences regarding influencing factors for user behavior and user activity and provides a deeper understanding of what drives social interaction and shapes the dynamics of the online social network.

Obtaining this information is valuable to judge the beneficing of a network or service to its users (Wagner, Rowe, et al. (2012)) and ultimately helps uncovering what makes a social platform thrive. Moreover, knowledge about user interactions enables the assessment of design choices and possibly presents contingencies for revision and enhancement (Benevenuto et al. (2009)). Attention patterns are also of notable importance for advertising and marketing strategies,

3 http://blog.reddit.com/2012/01/stopped-they-must-be-on-this-all.html

as they exhibit a definite impact on the dissemination of news and content in online social systems (Leskovec, Adamic, and Bernardo A. Huberman (2007)).

These topics are therefore of specific interest not only to politicians, political movements, designers, service providers and advertisers, but to anyone interested in sharing his or her opinion or promoting new concepts or even products. In fact, some of the most successful crowd-funding campaigns to date already included reddit in their marketing and promotion plans from the beginning. A prime example would be the virtual reality headset Oculus Rift[4]. And just recently Bill Gates (2014) took to reddit as a way of presenting and discussing his philanthropic endeavors.

Reddit is not just another social network and people new to it might struggle to understand its dynamics and find it difficult to apply their established strategies of getting the attention of its users. As of this writing, there are only very few scientific papers that study reddit thoroughly. This work aims to change this, especially regarding what drives user attention and user behavior in such systems, which type of content gets the how much and which type of attention and what deciding factors for popularity and success need to be considered.

1.2. Objectives

This thesis aims to set a first foundational step in providing a detailed overview of how attention works on reddit.

Score, as briefly described in section 2, is the primary measure used by reddit's sorting and ranking functionality. High score is vital for getting to the front page and in effect synonymously with being able to promote content and get one's submission recognized by a large audience.

4 http://oculusvr.com/

There are possible other indicators of attention that are not inherently exploited by reddit. For example, the amount of comments a submission receives. A high number of comments might indicate a lively discussion or at least that many user feel the need to share or say something about the submitted content. In some subreddits, an active discussion seems mandatory for making a post popular.

This leads to several new questions that need to be investigated. Part of the aforementioned overview can thus be accomplished by exploring which *indicators of attention* exist on reddit and subsequently by understanding their *characteristics* and how they *relate* to each other.

As new technologies evolve, the internet itself takes on new shapes and the interest of users on social networks might be changing as a consequence. New types of content or means of presentation are introduced, while others diminish. This work wants to provide insights into how these attention measures *developed over time*, what may have influenced this changes and, furthermore, if there are *distinct differences* in the attention and popularity *evoked by the various types of content* submitted to reddit.

While *a picture is worth a thousand words*, reddit submissions still rely on words and textual titles as the principal way of presenting the submitted content on its pages, even if there is another medium hidden behind it. A precise, sensational or even trenchant headline could be crucial to stand out of the mass of content. And could there be other conditions besides the *headline* that need to be satisfied for fruitful submissions?

Since reddit's ranking algorithm is dependent on when a submission gets its first votes, as explained in great detail by Salihefendic (2010), *time* could possibly have an important influence. Other ingredients for popularity might include the choice of subreddit where the submission is listed, since different sub-communities might perceive content differently. Therefore, a major part of this thesis is concerned with identifying the main *stimulating factors* that impact the popularity of submissions, how much and which type of attention they generate.

Considering the above mentioned objectives and the motivation for this thesis (see section 1.1) the following research questions arise:

i. Can attention indicators other than score be leveraged for better understanding of user attention on reddit?
ii. Which are these indicators, how are they characterized and do they correlate with each other?
iii. Did these indicators and the users' perception of submissions and content evolve over the time?
iv. Are these indicators and their relations globally applicable or limited to certain parts of the platform?

1.3. Contributions

The main contributions of this thesis are as follows:

i. Provision of a comprehensive analysis of user attention on reddit, granting deeper insight into the factors driving and influencing user behavior and user interaction.
ii. Introduction of corresponding attention indicators, elaborating their attributes and their relationships.
iii. Assessment of the evolution of an social online community's allocation of attention over time.
iv. Evaluation of universally and specifically valid attention characteristics and assessment of the impact of sub-communities and other distinguishing factors on user attention on reddit.

Secondary contributions include the presentation of a process for semi-automatic categorization of content, based on a single information variable.

This work also provides a detailed overview of the online community portal reddit. Its literary analysis exhibits an overview of successful social networks, present and past, the differences and similarities between them and what has already been examined and discovered about them and their users.

Additionally, a brief introduction for machine learning is given and the methods and results from this thesis supply a base for further research on attention characteristics and patterns both on reddit or other social online communities with collaborative filtering.

1.4. Thesis Outline

This thesis aims to provide an in-depth examination of how attention on reddit works, first by elaborating how attention can be measured, how positive or negative attention and popularity can be determined on reddit and further, amongst others, deriving possible issues that need to be considered for creating a successful submission.

The development and purpose of reddit are described in chapter 2, complete with technical details and the structural organization of the platform. The chapter also provides an overview of the community and its activities.

Chapter 3 establishes what has already been done regarding the research on social networks. This explains what is possible, what the strategies taken in this thesis are built upon and finally shows that no other work, to my knowledge, has examined the topic of this thesis to provide extensive information on attention and its stimulating factors on reddit.

Following in chapter 4 are descriptions and short introductions to the specific scientific approaches and methodologies used in this work and a brief summary and exposition of the data set consolidating all research in this thesis.

The processing steps and experiments conducted on the aforementioned data set are explicitly explained in great detail in chapter 5.

The results and findings of these experiments and their interpretations are evaluated in chapter 6. The results are then briefly discussed in their entirety in chapter 7 and a concise summary of what can

be inferred to answer the research questions is given in chapter 8 followed by possible threats to the validity of the approaches and results. Finally, potential subsequent work precipitated by ideas and approaches executed in this thesis is mentioned.

1.5. Collaborations

Several people are to mention for helping shape this thesis.

This work would not have been possible to this extent without the input of Jason Baumgartner, who provided the immense data set upon which every further analysis is based.

Some selected extracts of this thesis and subsequent research, based on the ideas and methodology used in this work, have been published in the paper "Evolution of Reddit: From the Front Page of the Internet to a Self-referential Community?" (2014) at the Web-Science Track at the 23rd International World Wide Web Conference in Seoul, South Korea. There, the co-authors Clemens Meinhart, Philipp Singer, Markus Strohmaier, Fabian Flöck and the author of this thesis studied how user submissions have evolved over time and how the community's allocation of attention and its perception of submissions have changed on reddit. The conclusions derived together reflect back into some sections of this thesis and influenced some of the strategies and consequently results presented later in this work.

Clemens Meinhart not only co-authored the previously mentioned paper, but also works on his own thesis at the time of this writing. While his thesis tries to find an answer to what reddit is and how it got there, by taking an in-depth look into reddits structure and its evolution over time, Clemens Meinhart and the author of this thesis both use the same data set as the footing of their respective research. To further facilitate the analysis of content on reddit, a categorization scheme that is shared across both works was developed. Domain normalization as a preliminary step to this scheme was constructed

in a joint programming session. Clemens Meinhart then built a list of the most used domains, which were assigned to certain types of content manually and individually by both parties. The author of this thesis constructed means for automatically comparing and merging these propositions. In the following, the categorization scheme was extensively discussed, re-evaluated and improved upon feedback from advisor Philipp Singer. Additionally, a modified variant of term frequency - inverse document frequency was utilized in Clemens Meinhart's and this work. However, this variant was applied in entirely different settings and was further adapted by the author of this thesis to fit to the objectives of this work. It is important to note that every other part of this thesis, including the experiments and results based on aforementioned assets, is a product of the independent and sole work of the author, Elias Zeitfogel.

2. Introduction to reddit

The internet is a vast source for information and data of any kind, may it be knowledge, news, videos or music. But it is also a platform for discussion and advice, a place where questions are answered and where different views and opinions are communicated and shared. Reddit is a community-driven online portal combining both of these aspects. The site's title still states its original mission, to be "the front page of the internet", a gateway to the (best) content available in the internet. To achieve this, users can submit content in two distinctively different ways as depicted in figure 2.1.

2.1. Submissions and Ranking

Users can either post content in the form of hyperlinks to web sites or start discussions via ordinary text posts (so called *self* posts). People registered on the site can then up- or down-vote these submissions and thus create a dynamically changing ranking of the "hottest" content based on votes and time.

This measure of popularity is called *score* on reddit. Score is calculated by considering any votes a submission gets and then subtracting downvotes from upvotes. Posts are sorted and ranked according to their score and displayed to the user on reddit's frontpage or on the many sub-pages described later in section 2.3. Freshness of the content has a big impact on the ranking; the score does not decrease over time but newer stories get rated higher. The first votes weigh much more than the following; this is explained in more detail by Salihefendic (2010).

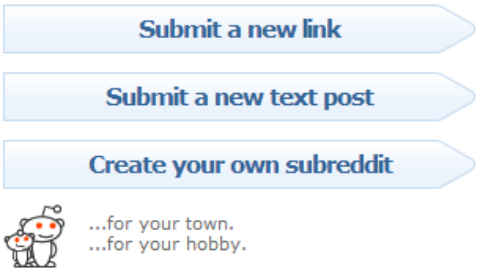


Figure 2.1.: The user interface for content creation on reddit. It was kept as simple as possible. The first two buttons are for submitting content to reddit, either external web-links or just text, the third button lets users create their own sub-communities.

The up- and down-voting system has the effect that stories with a large number of up votes, but roughly equally large number of down votes, might rank the same as low attention stories with just a few votes overall. This concept of controversy and engagement of users with a certain topic is therefore not reflected by the score.

But, reddit has another method of participation, namely the ability for users to comment on any submission or other comment. This feature enables another dimension of participation and the inclusion of a voting system for comments facilitates finding the best and presumably most qualitative ones for each submission, as seen in Figure 2.2. The up- and down-vote ability is symbolized by the ever-present and click-able upwards and downwards facing arrows.

2.2. Users

Users enjoy a high level of anonymity on reddit, besides a user-name and the *optional* recovery email-address, no further information is required to start submitting content and partake in the voting and commenting process.

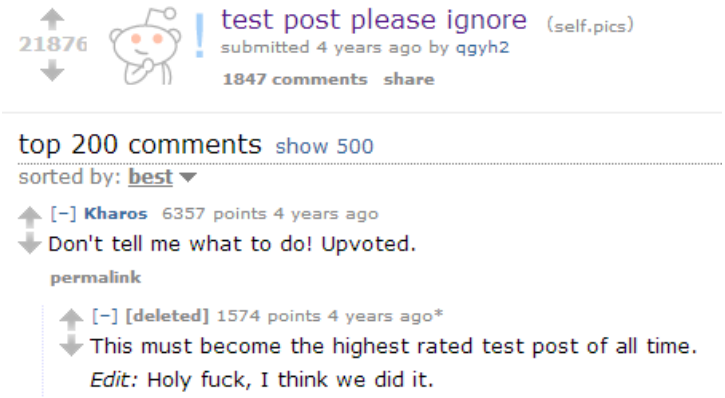


Figure 2.2.: A detailed view of a single submission with part of its comment stream. Notice that both the submission itself and every comment can be up- or down-voted using the small arrows.

Additionally, every news or topic can be read and viewed without logging in. The portal has a simple game-aspect implemented: users can earn *karma*[1] for every link submission or comment they make (but not for self posts). As simple as it sounds, this reddit specific virtual measure of reputation seems to work quite well at keeping users active, there are even external sites tracking and comparing *karma* statistics and development, including leaderboards and highscore lists.[2]

A standard user profile can be inspected in figure 2.3b, in this case the account Barack Obama used for his hugely popular post.

2.3. Subreddits

Reddit also utilizes concepts attributed to ordinary online community forums. To organize the broad range of content, a new feature was introduced: *subreddits*.

1 `http://en.reddit.com/wiki/faq`
2 `http://karmawhores.net/` provides further insight into the matter.

(a) A subreddit with short description, basic rule set and number of subscribers.

(b) A reddit user profile showing the link and comment karma generated.

Subreddits are basically mini-versions of reddit's frontpage but with a topical focus or interest. Since 2008, they are publicly available for any user to create his or her own sub-community.[3]

The user creating the subreddit is implicitly promoted to be the first *moderator* of this community and can equip his platorm with individual restrictions and regulations. A subreddit may only allow submissions concerning or discussing *barefoot running*[4] or a specific programming language, as depicted in figure 2.3a.

Founders can also assign one of three types to subreddits on creation,

- **public:** anyone can view and post new content
- **restricted:** anyone can view but only validated user can post new content
- **private:** both viewing and posting is restricted to validated users

3 http://blog.reddit.com/2008/03/make-your-own-reddit.html
4 http://reddit.com/r/BarefootRunning

Furthermore, every subreddit can have its own designated volunteer moderators responsible for the sub-community. Moderators have the ability to delete submissions that do not satisfy the rule-set of the community or even slightly change the look of *their* subreddit to create a fitting atmosphere for the topic at hand.

2.4. Frontpage

When a user first connects to reddit, he or she will be greeted by the frontpage (figure 2.4) with the latest and most popular submissions accumulated from a predefined set of subreddits, called *default subreddits*, including for example worldnews, technology and music.

The default subreddits can be found in the top menu of the page. Once a user is registered and logged in, he or she can start tailoring the reddit experience to his or her own interest and needs by adding (or removing) subreddits deemed fitting. This process is called subscribing and will add the submissions from a subreddit to the mix on the front page.

Besides viewing all subscribed subreddits together on the front page or just the content of a single subreddit, there is also the possibility to create *multireddits*, an intermediate step between a single subreddit and the front page. This feature can be used to combine a handful of subreddits, perhaps a couple of sports related subreddits, into a single page or view.

Popular subreddits with many subscribers even spawned their own cultures. Various modifications of language and unique acronyms are common, often not decipherable by the uninitiated. Likewise, there are different tastes of appreciation, in fact, what might be wildly favored in one community is sometimes even ridiculed in others.

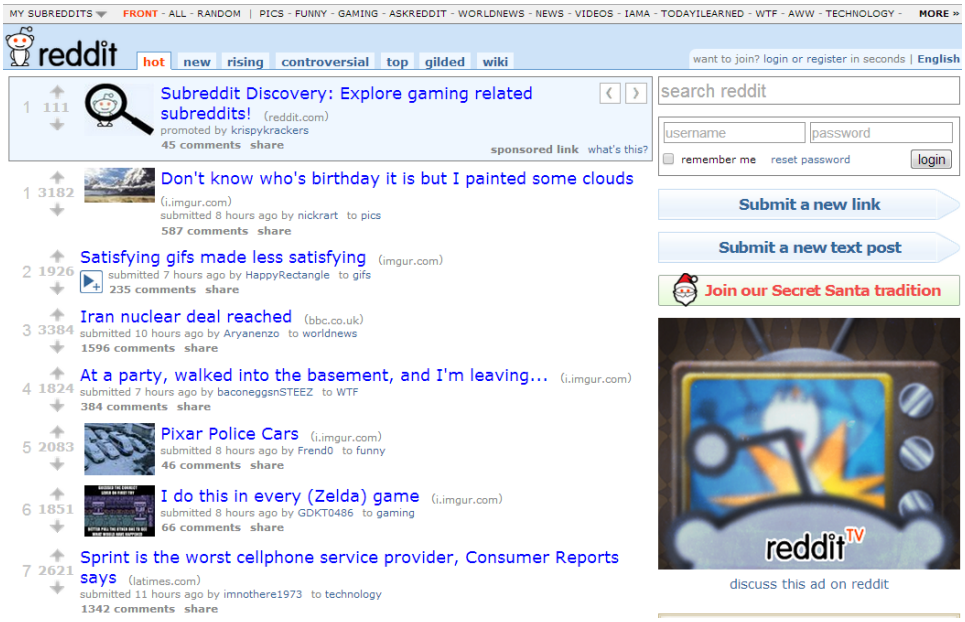


Figure 2.4.: The front page of the internet (reddit.com). The best submissions from the default subreddits plus one sponsored submission are shown and can be up- or down-voted directly from this view. The bar at the top lists the default subreddits and the user can browse his or her own chosen subreddits in the upper left corner. Below the first navigation bar, one can switch between displaying different rankings of the submissions, for example *hot*: currently highest rated submissions, *new*: newest submissions with few to none votes or *top*: highest rated submissions optionally for all time or just the last year, week, month and day. These display types are available for every subreddit. The search functionality, login and content submission is placed on the right side of the page.

2.5. Community

Generally, the tenor on reddit is very friendly and positive, many subreddits are created solely for the purpose of sharing. Sometimes for sharing experiences and knowledge as in the subreddit *IAmA* (short for "I Am A ... Ask Me Anything"), where people talk and answer questions about unique events in their lives, uncommon jobs and much more. Subreddits can generally by visited by adding *r/nameofthesubreddit* to reddit's uniform resource locator (URL) and are from now on annotated like this in this thesis (in the case of IAMA with *r/IAmA*). The aforementioned post of Barack Obama in section 1.1 was posted in this subreddit. This public Q&A was used to promote his presidential campaign in 2012.

But unlike other communities, for being the target of online advertisement, the reddit community wants something in return. In this case very personal information or reading what one of most powerful persons in the world might have to say (or write) about their concerns and worries. Having the opportunity to interact with people that are otherwise hard to reach at this level is very rare and something that makes reddit truly unique.

The topics and concepts covered by subreddits is immense (nevertheless, if something does not exist yet, it can easily be created in seconds). They range from news or sports to technology and gaming. There are subreddits for physics or chemistry and most programming languages and tools. They serve segregated as well as common discussion ground for movies, music or any kind of art and various genres. Countless informational subreddits, including *r/AskScience*, *r/AskReddit* or *r/todayilearned*, provide valuable and verifiable knowledge. Some encourage or only allow the submission of pictures, stretching from beautiful nature photographs, cute animals, funny events or adult content to online running gags, many of which were created on reddit and slowly trickle into other social networks and communities.

And then there are subreddits dedicated to helping other people,

organizing fundraisers, sending pizza to people in need[5] or motivating other people in their weight losing process. In the months before Christmas there is also a *Secret Santa* gift exchange within the community, which currently over one hundred thousand *redditors* (a synonym for reddit users, derived from the concatenation of *reddit editors*) take part in.

The users on reddit can generally be seen as very involved with everything going on around the site. Many successful and renowned posts reference or copy previously well received content and can lead to the creation of new subreddits or new internet phenomenons, so called memes[6], which often consist of a combination of pictures and variable text. Comments and the contained jokes have a tendency to be very self-referential as well; there is even a meme depicting an ironic view of the typical reddit user always referring to some aspects of current internet culture [7].

The author of this thesis co-authored a paper (2014) (mentioned in section 1.5) further investigating this trend of self-reference, which mostly manifests itself in growing focus on self created content. An uninitiated visitor may consequently have a hard time relating or understanding the purpose of some of the content. Despite this self-referential aspect reddit is an incredibly timely portal. News and anything else get collected so fast that reddit users are often confronted with boring feeds on any other social network or news platform they visit, since they already read it on reddit. A fact that in retrospective may have given reddit its name, as a combination or/and concatenation of *read it* and *edit*.

With a rising numbers of users reddit also quickly became the name giver for the now appropriately dubbed *reddit-effect*. This expression describes a phenomenon, where a popular submission links to another online service and the sudden spark of interest in this resource, combined with large numbers of users trying to access the service, causes it to be virtually unreachable or even crash. A problem that

5 `http://reddit.com/r/randomactsofpizza`
6 `http://en.wikipedia.org/wiki/Internet_meme`
7 `http://knowyourmeme.com/memes/internet-husband`

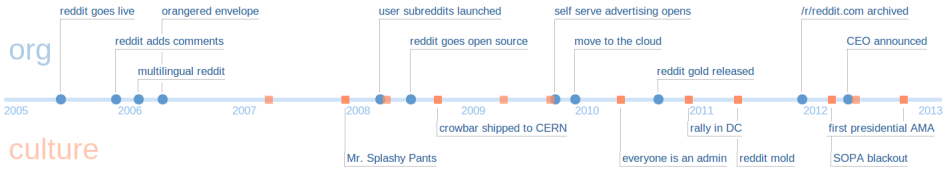


Figure 2.5.: Timeline of major milestones for reddit both in organizational structure and cultural events.

lead to the development of several reddit bots which automatically take snapshots of submitted online resources before the immense amount of traffic would overload their servers.[8]

It is also worth mentioning that reddits source code is publicly available for quite some time now, as announced by reddit (2008). The system is built on freely available tools or software libraries and everything can be reused. The only pieces of code, which are excluded from this, are mostly related to anti-spam or cheating protection. Only two years later an initiative was started by reddit (2010) to get support for the costly operation of their servers. For buying the virtual currency *reddit gold* users get premium features including more sophisticated filtering or comment tracking and the possibility to disable advertisement. A small sidebar on reddit tracks the daily progress as a proportion of the daily server cost.

As a side note, these publications and other important milestones or events can be found on reddits *about* site[9] where they are neatly formatted in a time-line as displayed in figure 2.5.

The phenomenon of reddit as a community and social news portal is hard to describe and one is advised to experience it him- or herself to get a more exhaustive impression of how reddit works and what it is about.

8 `http://reddit.com/r/snapshot_bot` is a particularly convenient bot, it can be activated without leaving reddit.

9 `http://reddit.com/about`

3. Related Work

Social networks and communities have been the target of research and scientific inspection ever since the rise of the internet providing easy means of communication for a large and growing audience. The first years saw interest in various and often rather small forums or early message boards. Later, much of the existing research has targeted the enormous user-maintained, collaborative online encyclopedia Wikipedia[1].

In recent years, social network portals like Twitter[2], a micro-blogging service that enables user to send publicly readable but limited text messages, have taken the spotlight for many works. The real-time mentality, the broad spectrum of users (from media and businesses to public and private persons) and the tremendous amount of users and activity provide countless opportunities to gain insights into communication behavior and social dynamics.

This sheer amount of research related to social networks and platforms taken together inspired some approaches engineered in this work to investigate certain attributes of reddit. Some important studies and their findings and contributions are shortly explained in section 3.1.

Reddit itself, despite its age in the fast changing world of online communities (see section 1.1 and chapter 2 for more details), has not been a prominent target for researchers and many attributes and traits of the platform are still largely unknown. Yet, some preliminary work exists, mostly covering very specific characteristics or examining single features of the portal in great detail.

1 http://wikipedia.org/
2 http://twitter.com/

Their results are summarized in section 3.2 and their collectively slim coverage of reddit only further emphasizes the need for an broader insight in what drives user attention on reddit and which factors play important roles in deciding the success of submissions and content.

3.1. Social Network Analysis

The works and results described in this section are structured by the online platforms which are investigated. The social network in question and its basic attributes will be shortly elaborated. Table 3.1 and table 3.2 show all social networks that occur in this chapter, including the number of users and characteristics for each platform. The sources, where these user numbers were retrieved, are depicted in appendix A.1.

This overview helps to better understand why some networks or platforms inspired much more research and scientific interaction than others and why certain aspects could only be scrutinized on specific portals.

Additionally, considering the information found in chapter 2, the reader will be able to see which attributes of these networks differ from reddit and which are shared between the platforms thus allowing to adjust and re-evaluate some of the described research for this thesis.

Table 3.1.: The table provides a simple overview of the larger social networks that occur in the research described in this chapter. The features are extracted from the authors own experience with said social networks combined with descriptions from the scientific works researching them. Trustworthy user numbers are hard to come by, some of them are taken from the sites themselves or their press releases, some numbers are from web statistic or knowledge portals such as Wikipedia. For some networks with stark fluctuations in user numbers, the counts representing the portal at the time of most of the research described in this chapter are chosen. Although these numbers are simply based on recent reports (see appendix A.1), they are still included here, since their main purpose is to facilitate comparison and create a modest overview of the social networks.

Network		**Users**	**References**
Facebook	Elaborate user profiles, friend system, comments, posts, Like-button	1230m	Backstrom et al. (2011), Spiliotopoulos and Oakley (2013), Jensen and Dyrby (2013)
LinkedIn	Dynamic online CV, professional connections	300m	Benevenuto et al. (2009)
Twitter	Microblogging portal, follower system, retweets, hashtags	241m	Rao et al. (2010), Kwak et al. (2010), Duan et al. (2010), Pak and Paroubek (2010), Cha et al. (2010), Bakshy et al. (2011), Rowe, Angeletou, and Alani (2011), Cheng et al. (2011), Conover et al. (2011), Hong, Dan, and Davison (2011), Wagner, Singer, et al. (2013), Cohen and Ruths (2013)
Digg	News aggregator, user submissions and comments, positive voting	236m	Lerman (2006) , Wu and Bernardo A. Huberman (2007), Lerman (2007),Szabó and B. A. Huberman (2008), Lerman and Galstyan (2008), Y. Zhu (2009), Jamali and Rangwala (2009), Tang et al. (2011)
reddit	News aggregator, user submissions and vote-able comments, positive and negative voting, community created subreddits	115m	Van Mieghem (2011), Duggan and Smith (2013), Lakkaraju, McAuley, and Leskovec (2013), Gilbert (2013), Weninger, X. A. Zhu, and Han (2013)

Table 3.2.: The table provides a simple overview of the smaller social networks that occur in the research described in this chapter. Also note that the same limitations as in table 3.1 apply for user numbers and extracted features.

Network		**Users**	**References**
Orkut	Elaborate user profiles, friend system, comments, posts	66m	Benevenuto et al. (2009)
Sina Weibo	Chinese microblogging portal, follower system, retweets, hashtags, extensive set of emoticons	60m	Fan et al. (2013), Bao et al. (2013)
Myspace	Elaborate user profiles, media hosting and artist promotion	36m	Benevenuto et al. (2009)
4chan	Community board, fast paced and highly anonymous	25m	Bernstein et al. (2011)
Hi5	Elaborate user profiles, friend system, comments, posts	11m	Benevenuto et al. (2009)
Slashdot	Technology news aggregator, user submissions and rateable comments	3.7m	Gómez, Kaltenbrunner, and López (2008), Kunegis, Lommatzsch, and Bauckhage (2009), Kaltenbrunner, Gomez, and Lopez (2007)
Mendeley	Academic collaboration network	2.5m	Schöfegger et al. (2012)
Boards.ie	Irish community board	2.4m	Wagner, Rowe, et al. (2012)

Digg

Digg[3] is a social news aggregator very similar to reddit. In fact, Digg as a platform predates reddit by being launched as an experiment in 2004. The main concept is to let users submit content and then vote on it to decide what is newsworthy. Additionally, Digg implements an asymmetric friend system (similar to followers on Twitter).

Digg has a few key differences to reddit regarding its mechanics, which are discussed in more depth by Lerman (2006). Users on Digg can only submit links to external content and add a short description, but are not able to submit just textual content (*self* posts on reddit). The voting system also only allows for positive votes (*diggs*), while reddit supports both likes and dislikes of submissions via the up- and down-vote mechanic. Lastly, the recommendation algorithm on Digg takes a user's friend network into consideration, a feature called *social filtering* as explained by Lerman, whereas reddit has an *collaborative filtering* approach and provides more extensive and active individual filtering through its sub- and multi-reddits.

It is important to note that Digg has undergone several relaunches and some mechanics have changed over time and while their impact and influence on the online social network is mentioned in the results of following works, they may not be implemented in the same exact functionality any more.

Online platforms that enable users to vote on their content and directly influence which content is shown often provoke the search for more insight and possible improvements for this recommendation process. **Predicting the popularity of online content** by Szabó and B. A. Huberman (2008) models a process to predict the popularity of submissions on Digg up to 30 days ahead, only based on user activity and interaction with the same submissions in the first few hours after they were created. Predictions by several methods are compared, with linear regression proving to outperform others for the first few hours after submission time.

3 `http://digg.com/`

The authors uncover a stark fluctuation of activity on the portal based on time or weekday, weekends tend to have around 50% less user interactions on average. To accommodate for this changes and make further calculations largely independent from the specific time a submission was posted, Szabó and B. A. Huberman construct the metric *digg time* which progresses in votes cast instead of seconds. By observing the development of submissions promoted to the frontpage in digg time, a distinct separation of submissions into two clusters is achieved. A part of the submissions reaches their maximum popularity within the first hour after appearing on the site while the rest continues to grow in popularity for several days. The first cluster presumably indicates miscalculations by Digg's own prediction algorithm, since these stories were promoted but did not get any momentum, reinforcing the need for better prediction techniques.

Another paper considering the advantages and disadvantages of different prediction and recommendation techniques in online communities is **Social Networks and Social Information Filtering on Digg** by Lerman. The authors elaborate on the distinct differences between *collaborative filtering* and *social filtering* and showed the power and quality of social filtering on Digg.

Collaborative filtering is mainly used by companies trying to sell their products, listed examples in the paper note Amazon and their recommendation system. Collaborative recommendation systems first find other users who shared the same opinion on resources, documents or products as a target user. Then the system checks for items that are liked by the matching users but (ideally) have not been considered by the target user yet. Contrary, social filtering strongly relies on the social ties between users of a network, assuming that friends share the same interests anyway (otherwise they most likely would not be friends) and recommends or displays items that have been liked or commented by friends of the target user.

Since Digg implements an asymmetric friend system, meaning if Alice *friends* Bob, she can track the stories he submits and the submissions he votes on, but Bob can not see the same information on

Alice without also be-friending her. Through this directed graph Lerman are able to identify two groups of users on Digg, *celebrities* which are tracked by more users than they track themselves and *fans* with inverse tracking characteristics. Also users with very active friend networks on Digg seem to be largely responsible for what gets ranked to the top of the site. The authors mention about a third of all top ranked submissions was submitted by just a tiny fraction of users. This stems from the observed fact that users will both digg submissions which their friends posted and submissions their friends liked before them. This should be carefully evaluated as the authors mention the possibility for a *tyranny of the minority* where just a few users effectively control the content of the site.

A similar conclusion is drawn by Tang et al. (2011) with **Digging in the Digg Social News Website** where they inspect the social friend network on Digg. Nearly all users in this network are part of a single giant component and therefore closely connected, but only a very small fraction of this network shares a symmetric friendship connection. And, as noted before, a analogously small fraction is responsible for the majority of popular stories. The authors find that the friendship network is by far not the only way user access new stories, about half of all stories are spread by users outside the direct friend network, users discovering the stories on the frontpage after being promoted by the ranking algorithm. Only then do stories accumulate most of their votes, since they are now even displayed to users which are completely disconnected from the rest of the network.

It is also found by Lerman and Galstyan (2008) that stories are already highly visible after only a couple of votes. If the path a story takes while spreading through the social network graph on Digg covers more users that are not in direct relation to the author of the story (friends of friends) it tends to get more popular than otherwise. In other words, stories that just stay within the users own friend network are less likely to be successful in reaching the frontpage. In **Analysis of Social Voting Patterns on Digg** Lerman and Galstyan document how to predict popularity of submissions on Digg by

considering where the votes are coming from and propose a very detailed dissection of the friend system.

Jamali and Rangwala (2009) suggest a different approach to popularity prediction. Their paper **Digging Digg: Comment Mining, Popularity Prediction, and Social Network Analysis** closely scrutinizes the interest of users inferred from their comment activity. Entropy measures are utilized to describe the prevalent comment behavior, exposing that the average user is not very focused on certain categories on the site but tends to consume and observe a wide array of topics. They developed a model to predict the popularity of new submissions using the following features: *comment statistics*, *user interest peak*, *user feedback* and *community structure*. Interestingly, they observe that only marginally worse results can be achieved when limiting the training data to the first few hours after a submission has been published. This finding could be related to the results from **Novelty and collective attention** by Wu and Bernardo A. Huberman (2007), where connection between freshness of content and popularity is examined. The more attention a story gets, the faster it will spread as more and more people learn about it and pass it on. Over time the novelty will diminish, attention declines accordingly and the dissemination stops. Results show the novelty indeed decays very fast on Digg in just a matter of hours with a half-life period of hardly above a single hour.

On Digg, users get ranked according to their ability to submit stories which get featured on the frontpage and activity (number of submissions, votes they cast, comments they write). Some users are able to hold their top rank for months while many others come and go. This ranking is featured prominently on the site, presumable as an incentive to participate actively, and can be retrieved via the sites Application Programming Interface - API. **Dynamics of collaborative document rating systems** by Lerman (2007) models how the influence of single users will change over time by incorporating activity and the users friends network in the process.

While many of the works above are concerned with popularity prediction and specific characteristics of the voting system on digg,

one paper is set out to capture the bigger picture. **Measurement and Analysis of an Online Content Voting Network: A Case Study of Digg** by Y. Zhu (2009) aims at granting deeper understanding of how online networks with user submitted content and friend-guided voting mechanisms work. High level structural analysis is conducted, the user network shows low link symmetry and a weak correlation of in- and out-degree, which possibly results to the same as the discovery made by Lerman (2006) in "Social Networks and Social Information Filtering on Digg" with their user groups of celebrities and fans. Y. Zhu note a discrepancy in users ability to promote stories and discoveries about the voting power of users which are tracked by many other users are confirmed. Generally, the number of diggs a submission receives combined with the pace of when these diggs are accumulated are deciding factors for promotion to the frontpage, but do not necessarily guarantee it. The algorithm calculating the rank of submissions on Digg is not available to the public (unlike reddit, see section 2.5) and therefore some characteristics of it can only be assumed by reverse engineering. Y. Zhu differentiate between two filtering systems on Digg, one is dictated by the ranking algorithm and the other consists of the intrinsic friend network of each user. Deviant from other research the authors conclude that while both filter influence user interactions the ranking algorithm is much more powerful in doing so.

Slashdot

Slashdot[4] is a social portal for technology news of all sorts. Users post links to or short summaries of articles from external sources. The articles are organized into a small set of generic categories. Users can comment on any of these posts. The comments can then be rated by other users on a small scale (unlike the theoretically unlimited up- and down- votes a reddit comment can get) to ensure high-quality discussions and a friendly atmosphere and discourage unintended behavior such as spam.

4 http://slashdot.com/

With the absence of direct voting on submissions, the votes on comments and comments themselves are the primary source of user interaction on Slashdot. In **Statistical analysis of the social network and discussion threads in slashdot** Gómez, Kaltenbrunner, and López (2008) examine the properties of networks created from the comment streams on Slashdot. They create different networks, both undirected and directed, based on users commenting each other, since there is no other explicit link information available. The weight of the edges depends on the amount of comments the observed users exchanged. Both types of network reveal a giant single component containing most of the users of the network, not unusual for most social networks. They also find that, contrary to their expectations, not necessarily authors of posts but mainly regular users, who do not submit any content, are the most active users on Slashdot. Finally Gómez, Kaltenbrunner, and López speculate that for generating the most comments or attention a wide array of different opinions for the matter at hand is presumable a key factor.

Other factors influencing popularity and attention are analyzed in **Description and Prediction of Slashdot Activity** by Kaltenbrunner, Gomez, and Lopez (2007). The interaction of time and attention (expressed by comment activity) a post gets is considered and used to accurately predict when and how much attention a post receives depending on the time it was submitted.

Slashdot's *Zoo* feature augments the friend network with a functionality where user can annotate other users describing their sentiment towards them. **The slashdot zoo: mining a social network with negative edges** by Kunegis, Lommatzsch, and Bauckhage (2009) is looking at the user network created by the Zoo feature. This implies a network with negative edge weights, since users can not only be tagged as friends, but also as foes. The authors then study the resulting network on different levels, for instance on node level by creating a new popularity measure dubbed *Negative Rank* used to easily find users trolling the news platform. Their analysis concludes that the Slashdot Zoo network exhibits multiplicative transitivity, basically enforcing, as Kunegis, Lommatzsch, and Bauckhage put it, "the enemy of my enemy is my friend" principle.

Twitter

Twitter is a micro-blogging service where messages are limited to 140 characters. People can create an account on the platform and start writing about whatever they like (in consideration of the terms of service). The messages, so called *tweets*, are public and can be read by anyone. These tweets can be *re-tweeted* by other uses further spreading the message on the platform. To filter this continuous stream of content there are two general possibilities. For one, user can tag their literary outpourings with *hashtags*, the # -character followed by a mostly short and descriptive word or phrase summarizing or categorizing the content.

The other method of filtering or emphasizing content is provided by the follower system. Users can *follow* other users to keep up-to-date with whatever they are writing. It is worth mentioning that this is only a directed connection. The user that is being followed can only see who and how many user are following him, but is not automatically shown his followers tweets. This number of followers a user has combined with the directed graph network behind the involved accounts is often used a measure of popularity of a user and the produced content as stated by Kwak et al. (2010).

As already shortly mentioned above (see section 3) Twitter is a very popular choice among researchers for everything from social sciences to network theory and recommendation theories. This can be partly attributed to the easy-to-use and feature-rich API of Twitter, which facilitates crawling and generation of appropriate data sets, but the main reason still remains the sheer amount of users and variety of topics that are commented and discussed.

Observing a social online network in its entirety can be beneficial for discovering relations that would have remained concealed otherwise. **What is Twitter, a Social Network or a News Media?** by Kwak et al. provides an extensive overview for the Twitter platform as a whole. This is initiated by examining the topology of the directed network created by the follower mechanic. It is revealed that Twitter, similar to Digg, has quite a low proportion of bi-directional relations (though

not to the extent as the latter). In fact, two thirds of users are not followed by any of the users they follow themselves. Information about the time zone of users is interpreted as geographical location and shows that the more bi-directional relations a user has the larger grows the geographical distance of all users involved. Still, three quarters of users with bi-directional relations are within 3 hours of time difference of each other.

Kwak et al. propose various metrics to evaluate the rank of a user within the Twitter network: number of followers, by PageRank (the algorithm Google uses to rank webpages for it's search engine initially described in *The PageRank Citation Ranking: Bringing Order to the Web.* by Page et al.) and by number of retweets. While the first two exhibit roughly the same results, mostly famous and otherwise influential people the last ranking consists mainly of media and press organizations. Given such a low reciprocity in the network, the authors assume the main purpose of Twitter to be an social information network rather than purely a social network. To better understand the information that is disseminated the trending topics on Twitter are compared with Google Trend[5] (popular keywords in search enquiries on Google) and CNN[6] headlines. The results let Kwak et al. conclude that the role of Twitter is "a media for breaking news".

Following investigations by Kwak et al. target the evolution of trending topics, including their life span and the power of retweets. According to the authors, any retweet is able to spread to about 1000 users, independent of the number of followers of the original tweet. Although retweets prove powerful in their capabilities to spread messages, **An Empirical Study on Learning to Rank of Tweets** by Duan et al. (2010) suggests an improvement on Twitters own tweet ranking, which defines the ranking and popularity of a message as its number of retweets.

The new ranking expands the original algorithm by adding content features and a user feature. The content features include tweet length

5 http://google.com/trends
6 http://cnn.com/

and whether the tweet contains an URL. The user feature is new model to determine the importance of a user, not, as often assumed, by follower count, but by the amount the user is listed by other users. Combining this additional features the authors are able to outperform previous work in this area.

A similar path to refining the original ranking algorithm is taken in **Predicting Popular Messages in Twitter** by Hong, Dan, and Davison (2011). They construct a classifier to predict whether a tweet will be retweeted and said popularity. Several features are added to significantly improve the results. These features consist of *content*, *topological*, *temporal* and *meta-data* features. To refine the content features both Term Frequency - Inverse Document Frequency (TF-IDF) and topic modeling approaches are used, while the topological features mainly consist of metrics that were also used in some other works described here, such as PageRank, bi-directional links and graph structure in general.

A different ranking, *TunkRank* has been proposed by Tunkelang (2009). This ranking calculates influence and power very similar to PageRank and is less dependent on the number of retweets, focussing more heavily on the follower network structure. A service providing TunkRank for Twitter users[7] both via web interface and API was created shortly after the initial publication but is now discontinued.

There are many more works dissecting nearly every aspect Twitter and some interesting approaches will be briefly presented here. **The Wisdom of the Audience: An Empirical Study of Social Semantics in Twitter Streams** by Wagner, Singer, et al. (2013) estimates the background knowledge of audiences on Twitter and it's importance for interpreting messages. **Classifying Latent User Attributes in Twitter** by Rao et al. (2010) builds a classifier to uncover information like gender, age and political orientation from messages on Twitter. Also both **Classifying Political Orientation on Twitter: It's Not Easy!** by Cohen and Ruths (2013) and **Predicting the Political Alignment of Twitter Users** by Conover et al. (2011) are concerned with the difficult task of identifying political alignment of Twitter

7 `http://tunkrank.com/`

users. **Twitter as a Corpus for Sentiment Analysis and Opinion Mining** by Pak and Paroubek (2010) on the other hand presents methods to extract the general sentiment of twitter messages. **Everyone's an Influencer: Quantifying Influence on Twitter** by Bakshy et al. (2011) and **Measuring User Influence in Twitter: The Million Follower Fallacy** by Cha et al. (2010) conduct exhaustive research on the characteristics of influence and user's repercussion. **Predicting Discussions on the Social Semantic Web** by Rowe, Angeletou, and Alani (2011) present methods to anticipate the amount of discussion a tweet will generate. **Predicting Reciprocity in Social Networks** by Cheng et al. (2011) investigates indicators for direct interactions between users, whether an interaction will be alternating or remain one-sided.

Facebook

Facebook[8] is one of the most popular and largest social networks to date. Every user has an extensive profile that can be customized with pictures and various private information. Users can befriend each other and view their friends activities. These activities may consist of posting status updates, sharing updates from others or liking content on the site. The purpose of status updates on Facebook is very similar to submitting content on reddit, even filtering in some sense can be achieved by either *liking* sites within Facebook (effectively subscribing to their content) or actively choosing to hide content from specific sources from the feed.

Unlike many other social networks described here, a friendship on Facebook is a bi-directional connection. This is necessary since much of the content on Facebook can be restricted to be only viewed by friends (or friends of friends for that matter). Some of the interactions are even entirely private, for example private messaging or restricted groups. So despite offering a rich API for data on both use of certain features of the site as well as underlying network structures, many

8 `http://facebook.com/`

interactions or at least their content are still hidden to researchers and can not be taken into account when examining this diverse social network.

The abundance of private information and photographs on Facebook create the impression of a rather user-centric platform, **Understanding Motivations for Facebook Use: Usage Metrics, Network Structure, and Privacy** by Spiliotopoulos and Oakley (2013) therefore aims to grant insight into the different reasons for people to visit Facebook. To achieve this, the authors combine commonly used network metrics, for example network size and diameter, average degree, average path length and clustering coefficient, with their own usage and gratification framework. They also conducted online surveys regarding gratification from Facebook usage and privacy concerns. The survey enabled the identification of the following seven major motivation factors: newsfeed, social network surfing, social investigation, content, photos, shared identities, social connection. To further improve on this results, demographic data about the user taking the survey is included in the study and helps creating a detailed mapping of user types to main motivation factors.

A large number of friends on Facebook has long been a prestigious feat. The question arises, if all these friendships are of equal importance. Backstrom et al. (2011) dedicated their work to answering how a user splits his or her attention over all participants of the networks. In **Center of Attention: How Facebook Users Allocate Attention across Friends**, the authors explain the balance between focusing attention on selected few versus evenly spread over the whole friend network. The attention is differentiated in two types interaction, attention by direct communication and attention by viewing. For a more detailed analysis the authors propose several types of interaction. They are as follows: *messages* as a direct private communication, *comments* and *wall posts* as a direct public communication, *profile views* and *photo views* as viewing. Some of the results also consider the *relationship status* of the people in question, an attribute rather unique to Facebook that not many other social networks can exploit.

Facebook is not only used by private persons but also by companies and various associations. In **Exploring Affordances Of Facebook As A Social Media Platform In Political Campaigning** by Jensen and Dyrby (2013) investigates what political parties expect in return from engaging in social networks. The authors first identify three main goals for political parties in approaching social media, namely *facilitation* of direct communication, dialogue and promotion of messages, *projection* of personality, authenticity and informality and *creation* of interaction and involvement with the target group. They then proceed to validate if these targets where achieved, what failed to be carried out efficiently and which unintended by-products emerged from the actions taken.

4chan

4chan[9] was launched in the second half of 2003. The original purpose was to create a copy of the popular Japanese Animé forum *Futuba Channel*[10] for English speaking people. Since then the highly anonymous and fast-living online discussion and image-sharing board has developed to serve a wide range of interests and is said to be one of the biggest influences for internet-culture.

Little to no restrictions posed on content that is submitted. The diverse range of content is divided into several sub-boards, similar to subreddits. The most popular sub-board, "Random", also called "/b/" is an endless stream of mostly images and playful or joking conversations about them, many of which are often thought to be responsible for the creation or further popularization of highly viral memes, for example *LOLcats*[11]. Furthermore, there is no ranking or recommendation of posts, the newest addition to the platform simply appears on top of the page, but disappears once the interest (in form of comments) fades off.

9 http://4chan.org/
10 http://www.2chan.net/
11 http://knowyourmeme.com/memes/lolcats

Content that is swiftly dismissed is in stark contrast to most other social networks, being able to revisit what happened days or even weeks ago often represents a prime feature - consider the timeline in Facebook. This unique characteristic inspired **4chan and /b/: An Analysis of Anonymity and Ephemerality in a Large Online Community** by Bernstein et al. (2011). The paper studies how an online community lacking nearly any sign of identity for its users paired with rapidly vanishing content is still able to to remain a largely popular and incredibly active community. Their analysis is concerned with posts from the already mentioned sub-board /b/ exclusively and starts with an attempt to capture the whole span of content submitted into nine categories describing the presumed purpose behind the submission. The categories range from *themed*, *sharing content*, asking for *advice* and encouraging *discussion* to *requests for action or items* and *meta*. Over 50% of all content observed fits into the categories themed, sharing content and advice.

The authors then continue to calculate the lifetimes of the threads in their dataset with interesting results. The shortest lifetime of a submission was only 28 seconds, whereas the longest thread was kept alive just above six hours. The median lifetime of posts was very small as well with barely four minutes before they disappear forever. The same calculations are done just for the first page which, by definition, is the most visible for the board, since no additional navigation steps are required. Bernstein et al. explain how users can control the flow of submissions by actively keeping threads alive or burying them and elaborate the role of the time of the day for both longevity of submissions and general activity on the portal.

On identity, below 10% of users chose an identifiable name for their posts, the rest preferred to stay anonymous. Still some signs originated in the 4chan community to define status or reputation. Most of these signs are only communicated as part of the submitted content. This high degree of anonymity is attributed by the authors to amplify the popularity and attention even very intimate topics or questions receive.

Sina Weibo

Sina Weibo, often just referred to as Weibo (Chinese for micro-blogging) due its enormous popularity, is, as the name suggests, a Chinese micro-blogging service very similar to Twitter and currently has over 50 million active users, despite being just over 4 years old. The platform borrows some design elements, most mentionable the 140 character limit and hashtags, from Twitter and merges them with some features from Facebook, such as a time-line of user actions, private messaging and the native inclusion of images, audio or video in posts. The service also cultivated a markedly comprehensive set of *emoticons*, small graphical elements denoting certain emotions, which now play a vital role in communicating on Sina Weibo.

Precisely this feature of conveying predefined emotions within text messages peaked the interest of Fan et al. (2013). In **Anger is More Influential Than Joy: Sentiment Correlation in Weibo** they write about the possible emotional influence happening between people using Sina Weibo. They build a classifier on top of categorizing emoticons used on the micro-blogging service and successfully annotate tweets to four different emotions: *joy*, *anger*, *sadness* and *disgust*.

Their findings suggest that anger has a much stronger influence on nearby nodes in the network than any other emotion. A fact proposing that angry sentiments or related news travel more quickly to other users and that their overall reach might be higher. Joy exhibits the same characteristics, just not as distinctive. Disgust and sadness on the other hand seem to have virtually no influence on even the closest neighbors, although Fan et al. experience a stronger sentiment correlation if users interact more frequently.

Just considering such user characteristics and their interaction networks could prove to be enough for certain prediction tasks according to Bao et al. (2013). Their paper **Popularity Prediction in Microblogging Network: A Case Study on Sina Weibo** explores the possibility of constructing a model to predict popularity of messages on Weibo without considering the actual content of the messages. The authors rather concentrate on the retweet path of a message and its

structural characteristics such as *link density* (ratio of the number of follower links to the number of all possible links) and *diffusion depth* (the longest occurring retweet path). By combining the path a tweet takes along with information about connectivity of the users on this path they are able to significantly improve on previous prediction results.

Various Social Networks

This section collects works concerning various social networks. They are treated together in a single section for mainly three reasons: (i) the social network in question has a structure or purpose very dissimilar from the previously observed networks, (ii) the social network has dramatically lower user base than the previously observed networks, (iii) the social network was not the exclusive target of the scientific work but rather a combination of different networks was used.

Hi5[12] is a social network founded in 2004 aiming specifically to be a means for social discovery, a place where people can meet and connect with new and interesting friends from all over the world.

LinkedIn[13] on the other hand is a social network for the professional world, linking specialists, managers and executives via job advertisements, company listing and direct connections. Most user profiles on LinkedIn are basically dynamically changing online curriculum vitaes.

Orkut[14] started as a social network with a *like*-button similar to Facebook and originally there was no privacy and everything was public. Over the time though more and more privacy restrictions were introduced. It is most widely used in Brazil and India.

Myspace[15] was one of the most popular social networks just a few years ago, but has steadily declined in usage statistics since. It is a

12 http://hi5.com/
13 http://linkedin.com/
14 http://orkut.com/
15 http://myspace.com/

social network service with focus on original music or now media in general, a place where bands and artists can promote their music.

Boards.ie[16] is a simple message board spanning a wide variety of topics. According to their own site, it is "Ireland's largest online community", with "almost 17,000" posts per day.

Mendeley[17] is a social network targeting academic users by providing management tools for collaboration, sharing, reference organization and research discovery.

Email-based discussion lists (DL) are not centralized social networks like the ones mentioned so far, but share some common features. Users can subscribe or unsubscribe and information gets send out to anyone on this list periodically.

The more information a online portal possesses about its users, the better it can serve them with tailored results and targeted services. The required information is often not explicitly declared by the users, but could be implicitly revealed through their actions. **Learning User Characteristics from Social Tagging Behavior** by Schöfegger et al. (2012) investigates how good the collective set of tags a user applies to documents can communicate background information about said user. To test their theory the authors successfully applied classification methods on data from Mendeley to determine the primary research discipline of each user just by tagging behavior. Using Term-Frequency Inverse Document Frequency or even binary values as feature representation significantly improved the performance of the classifier while restrictions proved only useful when applied to the target representations to reduce overlap of disciplines.

But what if the necessary information is not just hidden, but also simply not accessible through the provided API? To tackle this problem Benevenuto et al. (2009) reviewed *clickstream* data from a social network aggregator, combining the social networks(Orkut, Hi5, LinkedIn, Myspace) into a single account, with publicly available data through APIs. Their paper **Characterizing User Behavior in**

16 http://boards.ie/
17 http://mendeley.com/

Online Social Networks reports that over 92% of user interaction consists of browsing and can therefore not be learned from the public data alone. The authors advance to define several states of interaction with a social network and create detailed models of transition probabilities between these states and time spent in each state. Follow-up experiments then distinguish the activities by the interaction target: this can be a users own profile, friends or people outside their direct friend network. Benevenuto et al. derive that most content, especially media, is actually discovered outside the friend network.

A possible threat to any social network is the existence of so called *lurkers*. The definition of a lurker is somewhat dependent on the online community at hand, but generally the term describes an individual who does not participate or contribute but confines oneself to observing and consuming what the community offers. **Lurker Demographics: Countint the Silent** by Nonnecke and Preece (2000) studies the relation of lurkers to contributors in email-based discussion lists. The authors deduce that larger communities might have a tendency to comprise a higher number of lurkers than smaller ones, but the main topic or focus of the community has a more definitive impact on the lurker proportion. The effort it takes to participate in the community is another influencer, as most of the observed discussion lists with a low percentage of lurkers are also the ones with most activity and require more effort than others with higher lurker shares.

Many community forums try to accommodate for the diverse range of interests of its users by implementing various topic-specific sub-forums. These sub-forums spawn their own communities which not only appreciate and discuss different content, but presumable evaluate content differently. **Ignorance isn't bliss: An Empirical Analysis of Attention Patterns in Online Communities** by Wagner, Rowe, et al. (2012) finds that attention patterns are highly dependent on the community itself. Factors creating attention vary greatly depending on context and topic, the authors go as far as suggesting that global attention patterns may not exist.

Additionally, they conclude that aspects which impact if a discussion

starts at all differ from factors that impact the length of a discussion, a novel observation that did not get any exposure in previously mentioned work. To arrive at these results the five following sophisticated feature sets were constructed on data from Boards.ie: *user features*, *focus features*, *content features*, *community features* and *title features*. These sets contain supplementary comprehensive information, for example topic likelihood and topic entropy (focus) or posting time, informativeness and total text length (content).

3.2. Reddit

The works described in this section either concentrate solely on reddit or their major contributions are derived from examining data and features on reddit. Some, for example "Widespread Underprovision on Reddit", engage in discussions for future work which could answer questions arising from their results and this thesis has incorporated some suggestions to enrich its empirical analysis and scientific methods describing attention on reddit in as many facets as possible.

Duggan and Smith (2013) recently conducted a survey regarding the internet use in the United States and presented the results in their paper **6% of Online Adults are reddit Users**. Findings include, as the title suggests, that roughly six percent of all adults in the United States who actively use internet services also visit reddit. A more detailed questionnaire revealed a strong bias towards younger male adults, in fact 15% of males aged 18 to 29 regularly use reddit in comparison with only five percent of female internet users in the same age group.

As briefly mentioned in section 1.1, the way reddit's voting system is implemented is a very unique feature for the social network compared to similar platforms. The results in section 5.5 will also show that voting is one of the most common types of interaction with the portal. Hence, it comes as no revelation that existing research has predominantly explored voting and the resulting score.

Once a submission is posted users can either up- or down-vote it to their liking and the score is updated accordingly. Van Mieghem (2011) modeled this process of vote accumulation as a general random walk. Their work **Human Psychology of Common Appraisal: The Reddit Score** studies the anatomy of score distribution on reddit. While the tail of the distribution would fit an exponential function, a mid section exhibits "power-law-like" characteristics, as the authors put it. This unexpected part of the distribution inspires further inquiries and reveals that as a submission receives more up-votes its down-votes will grow accordingly. A fact that is examined closely in the general sentiment analysis in section 5.5 of this work.

Voting behavior is presumably influenced by many factors. Besides the content of a submission, the presentation itself might be essential. Considering the time component in reddit's ranking algorithm, posting a submission when only few users are active could be detrimental. In **What's in a name? Understanding the Interplay between Titles, Content, and Communities in Social Media** Lakkaraju, McAuley, and Leskovec (2013) work to uncover the complex relationships between different components which play vital roles in determining the popularity of a submission. A statistical model is created taking into consideration the following aspects: *title* of the submission, *content* of the submission, *time* the submission was posted and the *community* (subreddit) the submission was posted to. Only image submissions which were posted several times are inspected to extract information from the discrepancies in titles, time and subreddits in relation to the obtained score. Two models, a *language* (for title features) and a *community* model (for everything else but title features) are constructed to learn the impact of each statistical metric. Lakkaraju, McAuley, and Leskovec find that the quality of the content itself is the most important factor for popularity, even though they do not propose or apply a measurement of quality. The choice of title is still important, characteristics such as length and descriptiveness as well as how good it is tailored for the target community remain crucial for success.

Equivalently intrigued by the voting behavior of reddit users, Gilbert (2013) explore a probable downside of collaborative filtering. **Widespread**

Underprovision on Reddit looks at how the social navigation mechanic implemented by reddit could eventually harm the sites ability to present only the most popular content. The problem here is that users need to vote on the exceedingly large stream of ingenuously new submissions. Only enough participation in this process can ensure a good separation of original and good content from the bad so that everybody can enjoy a frontpage where only the best submissions are displayed. If fewer users actually vote on these new submissions and a growing proportion of the user-base starts free-riding (enjoying the popular content without the effort of voting) reddit's capabilities of filtering the truly great and important content might diminish.

In a blog post Olson (2013b) is concerned with the development of reddit and gains first insights into reddit's evolution by studying the number of posts to subreddits. He finds a diversification of submissions into a growing number of distinct subreddits. Olson (2013a) also shares a guide on how to make a successful post on reddit based on a fraction of reddit's top rated submission. Similar to other works presented in this section, the author takes into account the submission time, received upvotes, domain of the external link of the submission and most used words in the title. Some of these factors will be expanded and investigated to a significantly larger extent in section 5.5 and section 5.8 of this thesis, in order to get a much more comprehensive view of the matter.

Research on online social networks is often limited by the dimensions of the available data, a social network provider may choose to refrain from disclosing everything that is collected. Technical constraints might also apply and prevent the collection of desired data in the first place. Gilbert therefore resorted to using pageview data from *imgur.com*[18], an image hosting service very popular on reddit, as an approximation of pageviews for image submissions on reddit. Although the data could prove to be inaccurate, since imgur.com has its own community and voting system and referrer information is not available, the authors reveal some interesting results.

18 `http://imgur.com/`

The number of pageviews popular submissions get is several magnitudes higher than the number of pageviews for new submissions, suggesting that only a small part of users are actually spending time looking through new submissions and potentially voting them. Direct implications of this are presumed by the authors to be that many submissions that could potentially have been popular never get popular and even popular submissions might have been ignored the first time they were posted. More than half of all popular submissions in their data set only get attention after being submitted at least twice, in some cases even five times.

Gilbert initiates a discussion to encourage further research providing first insight into several elements which could possibly explain this effect. The elements listed include the impact of the *inherent design* of reddit, the *voting mechanism* being no social interaction, *cross-posting* to larger subreddits, influence of the *title* and external circumstances like the *time of the day* when a submission was posted. The last two factors will be elaborated upon and are thoroughly investigated in section 5.6, section 5.8 and section 5.9 of this thesis.

The abovementioned works focused solely on submissions and their popularity, Weninger, X. A. Zhu, and Han (2013), on the other hand, observe submissions as merely a starting point for online discourse. Their study **An Exploration of Discussion Threads in Social News Sites: A Case Study of the Reddit Community** scrutinizes the evolution and dynamics of comment threads on reddit. The authors analyze the threads by both their development over time and the formation of hierarchical or topical structure. Their results show that many of the early comments are largely on the highest level in the hierarchy and often the very first comments to a submission already start sub-topics. More sub-topics can arise later in the thread, but the comments higher up in the hierarchy tend to receive more replies.

4. Methods and Materials

To answer the objectives posed in section 1.2 a thorough analysis of the available data is needed. The data used to achieve this endeavor is described in section 4.7. The other parts of this chapter provide a quick introduction and overview of machine learning in general and followed by more detailed descriptions of some of the scientific methods utilized to tackle the research objectives of this work.

4.1. Introduction to Machine Learning

Machine learning is closely tied with Artificial Intelligence (AI) and many theories and ideas are inseparable connected between these two wide-ranging concepts. To that end, it is not surprising that research in this areas did not start until the rise of computers and the digital age.

A first definition for machine learning is provided by Samuel (1959). According to him, machine learning is a "field of study that gives computers the ability to learn without being explicitly programmed ". Samuel played countless games of checkers against a computer and the machine collected all board positions and their outcomes (win/loss) to arrive at conclusions for better tactics and to eventually defeat its master.

A more general and formal description for machine learning is offered by Mitchell:

> A computer program is said to learn from experience E with respect to some class of tasks T and performance

> measure P, if its performance at tasks in T, as measured by P, improves with experience E.

(Mitchell, 1997)

This definition still holds true to Samuel's older approach. The task *T* can be interpreted as playing or winning a game of checkers, the experience *E* as the amount of games played and therefore board positions and outcomes collected, and performance measure *P* can be calculated as the probability that the learning machine will win the game.

Several types of machine learning have been delineated in scientific literature. Major categories include *supervised* learning, *unsupervised* learning, *semi-supervised* learning and *reinforcement* learning. In supervised learning the machine is taught valuable information about a task in advance before it can solve the task autonomously. Unsupervised learning algorithms, on the other hand, are not reliant on any prior knowledge. Semi-supervised learning combines both of the aforementioned systems to accommodate for specific tasks. Reinforcement learning is concerned with virtual agents exploring, based on the reaction to their actions within an unknown environment, how to maximize some sort of profit. The concepts behind supervised and unsupervised machine learning will be elaborated in section 4.1.1 and section 4.1.2, accordingly. Some of the examples used in these sections have been inspired by the Stanford machine learning course held by Ng (2014), available on the e-learning portal coursera[1].

Another area of machine learning has become very popular in the last years in the technology industry. *Recommender* systems are at the heart of large and successful companies like Amazon[2] or Netflix[3]. The online shopping giant uses recommender systems to determine which products might be bought together or which other products have been bought by customers. Based on this data Amazon tries to

1 http://coursera.org/
2 http://amazon.com/
3 http://netflix.com/

increase its profits by proposing products to buying customers. The online streaming portal Netflix aims to prolong its users interest in the service by suggesting new content based on previous ratings for movies or series. The collaborative filtering implemented in reddit to promote timely content with high score can also be viewed as a very simplistic recommendation system.

4.1.1. Supervised Learning

Supervised machine learning generally involves some manual input or other additional information that is provided for the machine to learn. This is comparable to a teacher - student situation, where the teacher offers initial guidance and detailed knowledge about a problem set until the student manages to solve similar problems on her or his own.

An important prerequisite for supervised learning is therefore the existence of already solved problems or correct answers. These known answers are often called a *training set*, since supervised learning algorithms utilize them to construct a model that ideally provides accurate solutions for new problems.

There are two very common applications for supervised learning, *regression* and *classification* problems, both of which will be described now in more detail.

Regression

Regression algorithms try to predict a continuous outcome value to a given input value. The prediction is based on already given input values and correct outputs for the problem.

A visual example for a linear regression problem and its utilization can be found in figure 4.1, inspired by a similar example by Ng (2014). Imagine the X-axis in all figures there to represent the size of swimming pools in a suburban area, the Y-axis the construction

price of those swimming pools and the red dots symbolize the actual swimming pools in the area. Someone wants to build a moderate sized pool (marked in blue on the X-axis in figure 4.1a), but does not know the cost or price for this undertaking. A linear regression algorithm calculates a fit for the data (green line) in figure 4.1b and in figure 4.1c the prediction for the price (dashed yellow line) can be inspected.

Algorithms solving regression problems mainly differ in the way they fit to the provided data. A linear fit as displayed in figure 4.1b might not always be suitable and other fitting models of non-linear nature are needed in that case.

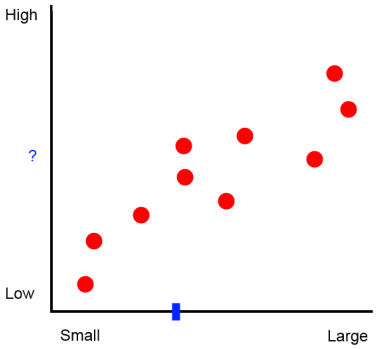
Regression analysis has a wide range of applications, from stock markets and real estate to biology and physics. It can be applied for virtually any situation where forecasting or predicting of outcome values based on previous experience is desired.

Classification

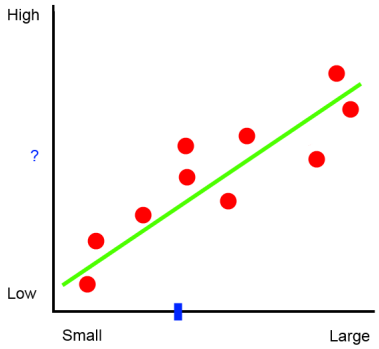
While regression algorithms predict a continuous outcome value to a given input, classification algorithms try to assign given inputs to distinct and predefined classes or labels. Once again, it is mandatory to already have knowledge about the problem prior to engaging the learning algorithm.

Consider a set of medical records about patients with tumors as the aforementioned knowledge. The patients and their tumors have been thoroughly examined and extensive data is provided in the medical records, depicting the tumors as either *harmful* or *harmless* for the patient. These two types of tumors serve as the distinct classes in this example taken and adapted from Ng (2014). After training on these records, classification algorithms could now help segregate the tumors of new patients between the two classes without the need of expensive and possibly invasive examinations.

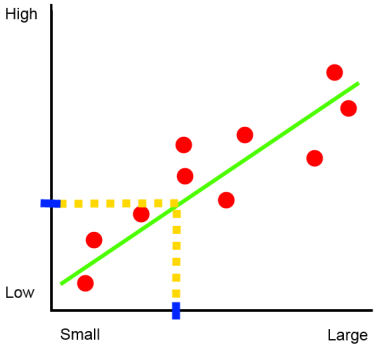
Figure 4.2 demonstrates a simplified two-label classification approach for the above mentioned scenario just based on tumor *size* as the



(a)Known answers and new problem



(b)Linear fit



(c)Prediction

Figure 4.1.: Figure 4.1a shows the data that is already known and the posed problem. A linear fit to this data is displayed in figure 4.1b. The result deducted via linear regression is illustrated in figure 4.1c.

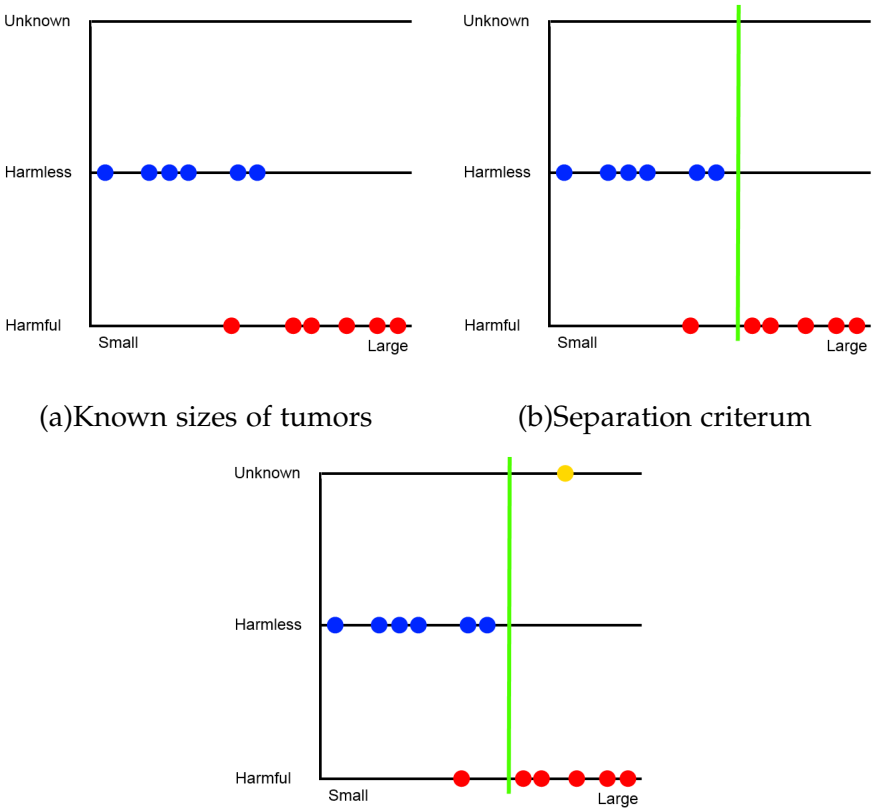
single feature, an example taken from . Algorithms for classification naturally also support multiple labels, expounded on the medical example that would mean more types of tumors, maybe different types of harmful tumors, some highly treatable, others not.

By adding additional features to the calculation, the quality of results can possibly be enhanced in otherwise difficult situations, as exemplified by figure 4.3. Here the seemingly inseparable classes of harmful and harmless tumors become easily divisible after adding a second dimension to the problem by considering another feature, namely the *age* of the patients.

Notice that even then, the separation line is not perfect, there are still some known tumors on the wrong sides of the separation line. Adding even more features could possibly resolve this issue and allow for a perfect classification. The question remains how trustworthy are results, if a perfect separation has not been found. To judge the implications and the quality of a classifier several performance metrics have been developed, they can be found section 4.3.

4.1.2. Unsupervised Learning

Unsupervised learning algorithms differ greatly from supervised learning in the general approach to problems. As can be guessed from the terminology, no guidance is provided before data is analysed. Picking up the student - teacher situation from before, the student is now on his own, the teacher just hands over a lot of data and leaves the room without further explanation. The student now has to try to find interesting aspects in the data. Unsupervised learning is often used to uncover properties and characteristics of data that is otherwise hidden, like inherent structures or hierarchies. Compared to the supervised classification algorithms, no classes are provided to which data should be matched to. Instead, a possible result of unsupervised algorithms could be the identification of such distinct classes that have not been established before.

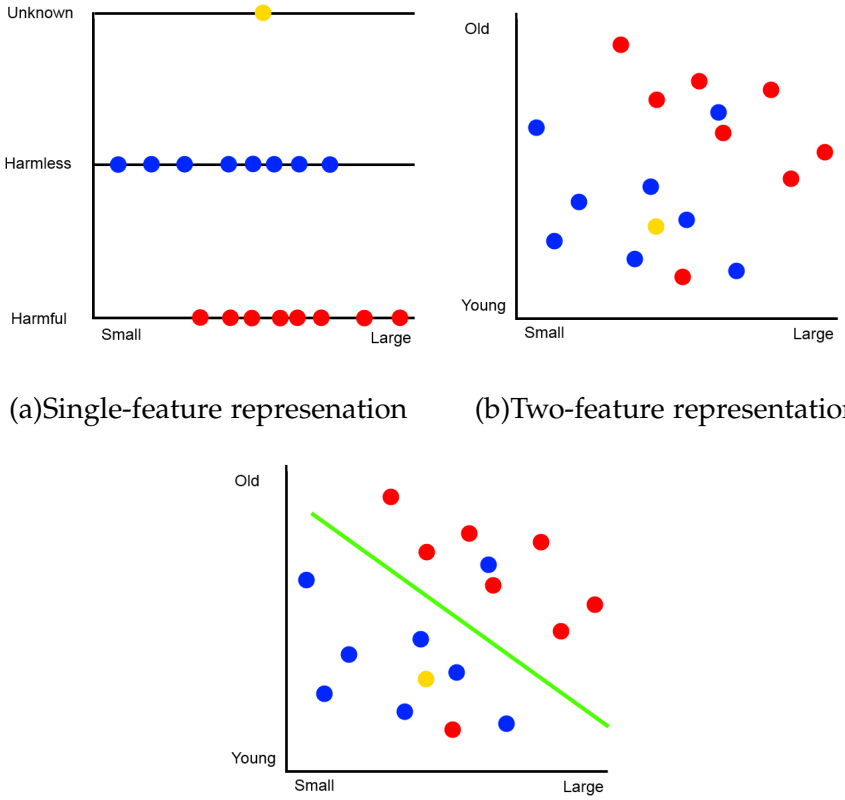


(a)Known sizes of tumors

(b)Separation criterum

(c)Classification of unknown tumor

Figure 4.2.: Figure 4.2a shows the harmless and harmful tumors and their known sizes taken from the medical records. In figure 4.2b a separation criterium (green line) is found by a classifier. Every tumor with a size smaller than this criterium (left side) is considered harmless by the classifier, every tumor larger larger than the criterium is considered harmful (right side). While not perfect, as evident by one harmful tumor left of the separation line, its the best result a simple single-feature classification algorithm can achieve here. A new unknown tumor (yellow dot) is handed to the classifier in figure 4.2c and is determined to be harmful due to its relative position to the separation criterium.



(a)Single-feature represenation (b)Two-feature representation

(c)Classification

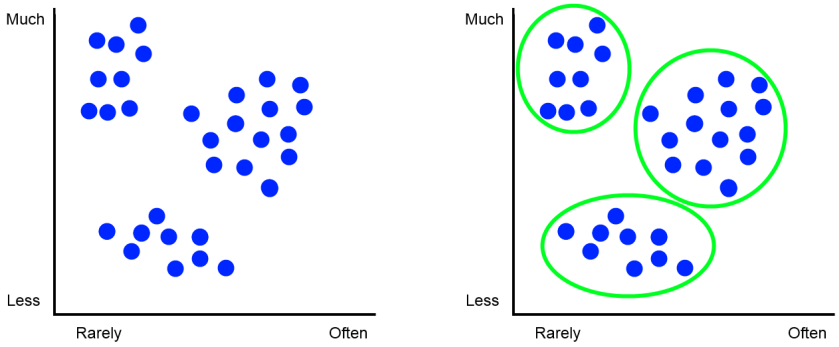
Figure 4.3.: Figure 4.3a illustrates a case, where it seems impossible to find a good separation between the two classes of harmful and harmless tumors, the unknown tumor can therefore not reliably be classified. Adding a second dimension (feature) to the problem, as shown in figure 4.3b, drastically simplifies this problem. The Y-axis denotes the new feature, the age of the patients with the displayed tumors. Figure 4.3c shows the now easily found separation line between harmful (above the green line) and harmless tumors (below the green line), the unknown tumor is classified to be harmless.

A common method of doing this is *clustering*. Clustering algorithms look through unlabeled data and try to separate it into cohesive batches. These batches, or clusters, can then be analyzed for prevalent properties, shared by any data point in a cluster. Popular applications of clustering include analysis of market segmentation, social networks or even astronomical photography. Examination of market segmentation, for example, could reveal groups of consumers which buy products from the same categories and have similar spending habits, both in amount and frequency. Marketing efforts can then be better streamlined for this new target groups. Looking at figure 4.4, one can see a simplified clustering approach on data with two features, the amount a client spends on her or his purchases on the Y-axis and the frequency of his purchases on the X-axis.

After such a separation into distinct clusters has been achieved, it is easy to designate new clients to one of the clusters, a straightforward way to do this would be a plain distance metric. However, labeling clusters still remains a difficult task, since automated labeling is often not easily interpreted by humans. This is especially true when performing text partitioning, where the most important common context for humans often greatly differs from what a machine perceives as the most influential common characteristic, as Baeza-Yates and Ribeiro-Neto (2011) note. To avoid issues with the interpretation of results, unsupervised text classification is therefore passed over in favor of supervised approaches in this thesis.

4.2. Naive Bayes Classification

As a method for exploring a possible community structure within reddit, this work relies on text classification. As already mentioned in section 1.2, text is the principal way how content is originally presented on the portal. If decisive separations in the way people communicate in different subreddits are possible and these results are conclusive enough, it can be derived that different sub-communities exist, even if they are just partitioned by their use of the language.



(a)Unexplored data (b)Recognized clusters

(c)Labeled data

Figure 4.4.: Figure 4.4a shows the collected client data, unlabeled, without any additional information. In figure 4.4b three distinct clusters are discovered by the algorithm and accordingly labeled in figure 4.4c.

Furthermore, this can help answer the research question regarding globally applicable attention indicators, since the existence of sub-communities could indicate the need to adjust ones means of presenting content in such a community to arrive at a successful and popular submission.

Naive Bayes classification is based on the theorem published by Bayes (1763). It is an entirely probabilistic approach. Given a set of documents which should be assigned to one of several classes, every possible combination of assignments is considered. An exemplary assignment would be that a document d is in a class c. The probabilities for all of these combinations are computed and for this calculation every document is represented by weighted vector $\vec{d}$, where every weight corresponds to a term in the document. Applying the Bayes theorem, the conditional probability that a retrieved document d represented by the weighted vector $\vec{d}$ is indeed in class c, $P(c|\vec{d})$, can be calculated as demonstrated by Baeza-Yates and Ribeiro-Neto (2011):

$$P(c|\vec{d}) = \frac{P(c)P(\vec{d}|c)}{P(\vec{d})} \tag{4.1}$$

Here, $P(c)$ is the probability that out of the aforementioned set of documents, a document in class c is drawn. Likewise, $P(\vec{d})$ is the probability that a drawn document would produce the weighted vector $\vec{d}$. The calculation of the probability $P(\vec{d}|c)$ differs for every type of naive Bayes classifiers.

To say in advance, independence among all terms in a document and therefore independence among all weights in the weighted vector is assumed to greatly simplify the computation. This is often not true for documents or texts in the real world, hence the prefix *naive* for this classifier.

A very basic idea for the weight vector representing a document would be to just value the weights for each term as 1 or 0, depending

if they appear in the document or not. In this *binary* naive Bayes classifier (BNB-classifier), a document d is assigned to class c, if and only if its score $S(d, c)$ is higher than any other document-class combinations. Baeza-Yates and Ribeiro-Neto calculate this score as

$$S(d, c) = \frac{P(c|\vec{d})}{P(\bar{c}|\vec{d})} \tag{4.2}$$

where, as described above, $P(c|\vec{d})$ is the probability that document d is in class c, $P(\bar{c}|\vec{d})$ is the opposing probability that document d is not in class c. Baeza-Yates and Ribeiro-Neto go on to substitute this probabilities with the definition mentioned earlier and arrive at roughly this estimation for the score:

$$S(d, c) \approx \frac{P(\vec{d}|c)}{P(\vec{d}|\bar{c})} \tag{4.3}$$

Now, finally to the computation of $P(\vec{d}|c)$ (and $P(\vec{d}|\bar{c})$ accordingly) for the BNB-classifier:

$$P(\vec{d}|c) = \prod_{t \in \vec{d}} P(t|c) \prod_{t \notin \vec{d}} P(\bar{t}|c) \tag{4.4}$$

where t is a term in document d and $P(t|c)$ denotes the probability of a term t appearing in an arbitrary document of class c, while $P(\bar{t}|c)$ consequently is the probability of a term t, which is not in document d, not appearing in an arbitrary document of class c. Other implementations of a naive Bayes classifier have altered this calculation, mostly due to using a differently derived weight vector. A more sophisticated approach than this binary weights is based on term frequency. The *multinomial* naive Bayes classifier (MNB-classifier) uses such an concept to calculate the weights in vector $\vec{d}$

based on how often the corresponding terms appear in a document. This is suggested by Baeza-Yates and Ribeiro-Neto to lead to better results, as more information about the documents is available. In this work *Term Frequency* and *Inverse Document Frequency*, as described in section 4.6.2, are combined for document weighting. While promising better results, on the other hand the computational effort increases significantly causing in much higher computation times compared to a binary method. Both, the binary and the multinomial naive Bayes classifiers are utilized side by side and thoroughly compared in this thesis.

4.3. Classification Performance Metrics

To compare above mentioned classification algorithms in section 4.2 regarding the best set of parameters and which one is most satisfactory for the task needed in this thesis, the performance of their results needs to be evaluated. Widely used metrics that are also suggested by Baeza-Yates and Ribeiro-Neto are *precision*, *recall* and the combination of both known as *F-measure*. It is important to note that all of these metrics measure the performance regarding only one class in a classification problem. Hence, it is often useful to calculate these metrics for every class and build the mean to get a proper overall performance measure of a classifier.

To better explain these metrics a set of documents, consisting of 20 documents, is assumed. The documents in this set cover a handful of topics, but each document is only covering precisely one of these topics. The topics can be regarded as the classes a classifier wants correctly appoint the documents to. One of these topics, topic c, is dealt with in ten of these documents.

4.3.1. Precision

The precision metric marks the fraction of all documents elected to a certain class c by a classifier that are actually in this class c. Drafted into a formula, precision for class c can be calculated like this:

$$P(c) = \frac{n_{correct}}{n_{assigned}} \tag{4.5}$$

where $n_{correct}$ is the number of documents in the intersection of the documents rightfully in class c and the documents being assigned to class c by the classifier. Furthermore $n_{assigned}$ is the number of the documents that are assigned to class c by the classifier.

So considering the aforementioned set of documents, assume a classifier concludes that twelve documents of the set deal with topic c ($n_{assigned}$). But only eight of these twelve documents do in fact cover said topic, four of them were wrongly assigned by the classifier. These eight documents are the intersection ($n_{correct}$) that was mentioned above. The precision for topic c is then

$$P(c) = \frac{8}{12} \tag{4.6}$$

$$P(c) = 0.75 \tag{4.7}$$

4.3.2. Recall

The fraction of all documents of a certain class c that are assigned correctly by a classifier forms the recall metric, mathematically defined as

$$R(c) = \frac{n_{correct}}{n_c} \tag{4.8}$$

where $n_{correct}$ is again the number of documents in the intersection of the documents rightfully in class c and the documents being assigned to class c by the classifier. The denominator is now the number of documents in class c, namely n_c.

Continuing the example, again, eight documents are correctly assigned ($n_{correct}$). The total number of documents in class c (n_c) is known to be ten (see section 4.3). The recall of the classifier is

$$R(c) = \frac{8}{10} \tag{4.9}$$

$$R(c) = 0.8 \tag{4.10}$$

4.3.3. F-Measure

With precision and recall there are now already two metrics evaluating the performance of a classifier. To simplify comparison between various classifiers even further, the two metrics can be combined into a F-measure regarding class c as follows

$$F_x(c) = \frac{(x^2 + 1)P(c)R(c)}{x^2 P(c) + R(c)} \tag{4.11}$$

Here, x is a weight that can be used to shift the importance in which precision and recall influence the F-measure. This is convenient when someone for example wants to focus on high precision classifiers but does not want to completely disregard the recall.

One of the most used variants, according to Baeza-Yates and Ribeiro-Neto (2011), is the F_1-measure, with $x = 1$:

$$F_1(c) = \frac{2P(c)R(c)}{P(c) + R(c)} \tag{4.12}$$

which basically is the harmonic mean of precision and recall and will also be adopted for classifier evaluation in this thesis.

Inserting the already calculated values for precision and recall from the examples in the corresponding sections, following value for the F_1-measure for class c can be obtained:

$$F_1(c) = \frac{2 * 0.75 * 0.8}{0.75 + 0.8} \tag{4.13}$$

$$F_1(c) \approx 0.77 \tag{4.14}$$

4.4. Correlation measures

One of the research questions in section 1.2 poses the question if *attention indicators correlate with each other*. To that end, this and the next section propose two very common and widely used ways of determining the relationship of two variables.

4.4.1. Pearson Product-Moment Correlation

Generally speaking, correlation measures are designed to determine the best fit over the data of two variables and calculate how far the data points are away from this ideal fit. Specifically, the *Pearson Product-Moment Correlation* (PPMC) weighs the strength of the *linear* relationship between two variables.

Originally, the idea for the formula was introduced by Bravais (1846), but only Pearson (1896) was able to finally prove it to be the best fit to a provided arbitrary bi-variate data. The *Pearson correlation coefficient* is derived from dividing the *co-variance* of two variables by the multiplication of their respective standard deviations. The formula goes as follows:

$$\rho_{x,y} = \frac{cov(X,Y)}{\sigma_x \sigma_y}$$
$$\text{where} \quad cov(X,Y) = E[(X-\mu_x)(Y-\mu_y)]$$
$$\text{so} \quad \rho_{x,y} = \frac{E[(X-\mu_x)(Y-\mu_y)]}{\sigma_x \sigma_y}$$

X and Y are the two variables in question, $cov(X,Y)$ is the co-variance of the variables and σ_x and σ_y are the corresponding standard deviations. E is the expectation and μ_x, μ_y are the means of X, Y respectively. $\rho_{x,y}$ is the resulting measure of correlation, also denoted as r, the Pearson correlation coefficient.

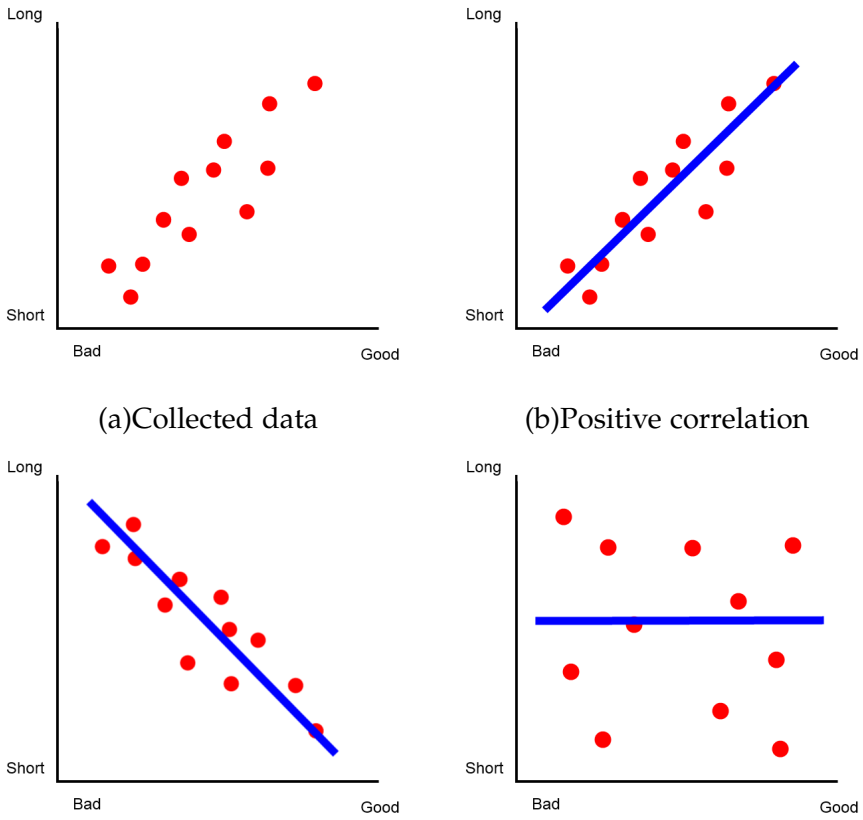
For a sample the values can be empirically estimated and substituted:

$$r_{x,y} = \frac{\sum_i (x_i - \overline{x})(y_i - \overline{y})}{\sqrt{\sum_i (x_i - \overline{x})^2 \sum_i (y_i - \overline{y})^2}}$$

Here x_i, y_i are the collected values in the sample, $\overline{x}$ and $\overline{y}$ the empirical means of the the variables X and Y in the sample.

The correlation coefficient, r, can take any value between -1 and $+1$. Both -1 and $+1$ would subsequently mean a perfect negative (-1) or positive $(+1)$ correlation between the two variables, whereas 0 would indicate no correlation at all. The farther away r is from 0 in either direction, the stronger the correlation and the less is the deviation of the data points from the line of best fit. A positive correlation would furthermore imply that as the value of one variable rises, the value of the other variable would rise too. Contrary, in a negative correlation, as the value of one variable rises, the value of the other variable decreases. Note that the formula takes no account for dependent variables, so the interpretation of the result is up to the observer.

By demonstrating a simple example, one might be able to get a better understanding of the process. Someone might speculate that



(a)Collected data (b)Positive correlation

(c)Negative correlation (d)No correlation

Figure 4.5.: The axes in all figures represent the two variables *leg length* (y-axis) and *average marathon placement* (x-axis). In figure 4.5a the red dots account for the athletes in the study. Figure 4.5b shows a positive correlation of the two variables ($r > 0$), longer legs mean better (higher) placements. Figure 4.5c displays a negative correlation ($r < 0$) and figure 4.5d depicts a case of no correlation (r would be close to or exactly 0 here).

(a)Non-linear relationship

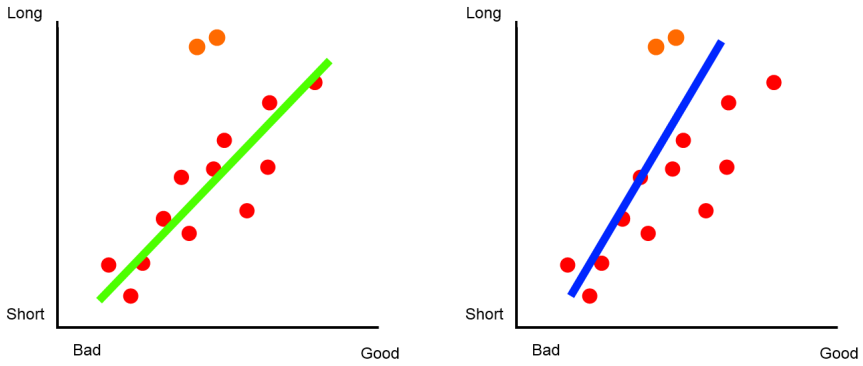
(b)Non-linear monotonic relationship

Figure 4.6.: Figure 4.6a shows example data that exhibits a *non-linear* relationship, figure 4.6b shows a non-linear relationship of *monotonic* nature.

people with longer legs can run better and faster consistently. This person further suspects that athletes with longer legs place higher in marathons. So this person starts collecting data from the athlete's profiles as well as their marathon statistics and comes up with values for their leg length and average placement achieved. These values are shown in figure 4.5a. Assuming that the blue line in figure 4.5b is the line of perfect fit, a positive correlation is found, r is presumable somewhere above 0.5, depicting a strong correlation, showing that indeed, longer legs imply better placements in races. For completeness, figures 4.5c and 4.5d present a negative and no correlation.

The second limitation of the PPMC is its lack of robustness against outliers. Outliers or extreme values can often be regarded as exceptions in the data, whose importance, for the overall result in such a correlation test, is often negligible. But since all data points are treated equally, just a few outliers could easily skew the result, as illustrated in figure 4.7.

There are two limitations to this approach of finding associations between two variables. The obvious one was already stated in the



(a)Outliers ignored (b)Outliers considered

Figure 4.7.: Outliers (in orange) are introduced to the example from figure 4.5. Figure 4.7a shows the hypothetical fit when outliers would be ignored, figure 4.7b displays how the fit is skewed by considering them for the calculation.

introduction of this section, the PPMC is designed to only determine *linear* relationships, any *non-linear* relationship in data is either poorly indicated or not recognized at all. Considering figure 4.6, an example for a non-linear correlation of two variables, it is easy to understand the inaptitude of linear approaches to provide satisfying results for this problem.

Removing of this outliers is a valid option, but any manipulation of the original data is rarely desirable. Nevertheless, finding an appropriate rule set for removal is often a highly sensible task, especially as the data gets more complex.

4.4.2. Spearman Rank-Order Correlation

To tackle the drawbacks of the PPMC, Spearman (1904) adapted the existing approach by Bravais and Pearson by replacing the actual measurements of variables with their ranks. Considering the empirical formula for the PPMC:

$$r_{x,y} = \frac{\sum_i (x_i - \overline{x})(y_i - \overline{y})}{\sqrt{\sum_i (x_i - \overline{x})^2 \sum_i (y_i - \overline{y})^2}}$$

The measurements are now replaced with their ranks in the sample:

$$r_s = \frac{\sum_i (rank(x_i) - \overline{rank_x})(rank(y_i) - \overline{rank_y})}{\sqrt{\sum_i (rank(x_i) - \overline{rank_x})^2 \sum_i ((rank(y_i) - \overline{rank_y})^2}}$$

The Spearman Rank-Order Correlation (SROC) coefficient is denoted as r_s, $rank(x)$ is the rank of the measurement of the variable X in the sample and $\overline{rank_x}$ the mean of those ranks.

In the case of tied ranks, the average of the occupied ranks is adopted for all affected measurements. If there are no tied ranks, a simplified calculation can be applied:

$$r_s = 1 - \frac{6 \sum_i d_i^2}{n(n^2 - 1)}$$

The formula utilizes the distance between ranks of measurements of X and Y, as d_i is the difference of $rank(x_i)$ and $rank(y_i)$ and n is the total number of measurements or ranks.

This exploitation of ranks mitigates the influence of outliers or distributional disparities in the data, as extreme values are now only considered by their rank and not their actual distance from the rest of the data. Additionally, associations of non-linear nature between variables can be more accurately detected, as long as they are *monotonic* (see figure 4.6b).

As with the earlier PPMC coefficient, the SROC coefficient r_s can range from -1 to $+1$, where, again, 0 indicates no correlation and $+1$ or -1 express a perfect positive or negative monotonic correlation.

4.4.3. Significance of Results

Statistical significance is a measure to determine if an obtained result is a mere product of happenstance, as discussed by Goodman (1999). In this case the obtained result is the correlation of two variables claimed by a correlation coefficient. To proof this claim, it has been established to invalidate the counterclaim. A hypothesis test is performed, assuming the *null hypothesis* that no correlation between the variables exist. The p value is then the probability that the observed results would also be obtained under the null hypothesis. If the probability p is below a certain threshold, the null hypothesis can be rejected and the observed claim is validated. In literature this threshold for the p value is commonly set at 0.05 or 0.01.

4.5. Entropy measures

Purely empirical observations are not sufficient for the analysis of the evolution of attention and user behavior on reddit. Entropy measures are used to better describe the equality (or inequality) of the distribution of attention on the observed submissions at a given point in time. These can then be combined to facilitate a greater understanding of the development over the time captured by the data set.

The *Shannon Entropy* has first been introduced by Shannon (1948) to characterize the information value of messages, defined as:

$$H(X) = -\sum_{i=1}^{n} p_i \log p_i \tag{4.15}$$

H is therefore "the entropy of the set of probabilities p_i,...,p_n" in the random variable X. The resulting entropy can also be normalized over the length of input to allow comparability over different

domains and is described by Corominas-Murtra and Solé (2010) as follows:

$$h(n) = \frac{H(X(n))}{\log n} \tag{4.16}$$

Several notes can be made about the resulting entropy from both approaches:

- H will take a positive value or 0, while h is bound between 0 and 1.
- If and only if all but one probability are 0, both H and h are 0, indicating perfect inequality of distribution.
- H is the possible maximum $logn$ for n observed probabilities if all of them are identical, indicating perfect equality of distribution. In this case, h will be 1.

4.6. Lexical Analysis

As already suggested in section 1.2, imposed through the design of the system itself, the submission title might play a vital role in determining the attention a post receives on reddit. Similarly, the text of a self post can stimulate the collective perception of a submission. Two different measures for lexical analysis of submission title and submission text are presented in this section.

4.6.1. Length of Text

The length of text is defined as the number of characters a string contains. This measure can be calculated quickly and can therefore be used in more complex analysis without demanding too much additional computational power. Still, the length of a title or submission text can expose important information about the overall attitude

towards a post. In this thesis, the length of a string is also presumed to be indicating the ease of reading said string. The higher the number of characters, the more time a user needs to read a submission title or text.

The assumption of this work is that, if specific communities within reddit exist, they react differently to this measure, for example, a scientifically interested sub-community could value long, descriptive titles more than a sub-community for sharing funny pictures. A short and captivating title could be better suited for this community. Thus, aforementioned information also includes knowledge about the acceptance of different types of titles in different communities.

4.6.2. TF-IDF

Table 4.1.: A set of documents and their contents.

Document	**Content**
document 1	A very very long text is very very long text.
document 2	Very long books are a very long read.
document 3	Long text makes books full of long text, also text.

Term Frequency - Inverse Document Frequency (TF-IDF) is a widely used scheme for term weighting. Baeza-Yates and Ribeiro-Neto (2011) even consider it to be the most popular term weighting scheme in information retrieval to date. TF-IDF determines the importance of terms in a document for describing and differentiating said document in an arbitrary set of documents. Its major appliance is the assessment of usefulness of documents to given query terms in information retrieval systems.

As the name suggests, TF-IDF is comprised of two initially separate weighting schemes: *Term Frequency*(TF) weights and *Inverse Document Frequency* (IDF) weights. The next sections elaborate on these systems and the fruits of their combination.

Term Frequency

Term frequency weights are a very straightforward approach. According to Luhn (1957), who first proposed it, the weight of a term is simply the number of times it appears in the text of a given document. The number of appearances is also called *frequency* for the remainder of this thesis. This implies that the more often a term appears in a document, the higher its frequency and the higher its TF weight. Successive research has adapted this idea, mathematically the TF weight is described by Baeza-Yates and Ribeiro-Neto (2011) as

$$tf_{i,j} = \begin{cases} 1 + log f_{i,j} & \text{if} \quad f_{i,j} > 0 \\ 0 & \text{otherwise} \end{cases} \tag{4.17}$$

where $tf_{i,j}$ is the TF weight of term i in document j and $f_{i,j}$ the frequency of term i in document j. The *log* is applied for smoothing and furthermore makes the TF weight easier comparable with the IDF weight outlined in the next section.

Considering an exemplary set of documents in table 4.1, a brief illustration of the weight calculation in table 4.2 emphasizes the simplicity of this approach.

Table 4.2.: A minimalistic example for the calculation of TF weights is shown (log in base 2). Table 4.1 displays the corresponding set of documents.

Terms	$f_{i,1}$	$f_{i,2}$	$f_{i,3}$	$tf_{i,1}$	$tf_{i,2}$	$tf_{i,3}$
very	4	2	0	3	2	0
long	2	2	2	2	2	2
text	2	0	3	2	0	2.585
read	0	1	0	0	1	0
...						

Inverse Document Frequency

Another foundation for term weighting was laid by Sparck Jones (1988) with her interpretation of term specificity, *Inverse Document Frequency* (IDF). The relation of term *specificity* and *exhaustivity* of document descriptions is likewise discussed by Baeza-Yates and Ribeiro-Neto (2011). The specificity of terms is defined as the capability of a term to describe the subject of a document. Terms covering a broader range of subjects can be easily applied to more topics and are assumed to be used more often to describe documents. Thus, terms that are utilized often in descriptions are less specific. Inversely, rarely picked terms tend to be more specific. Imagine an example with the two terms "food" and "pizza", the latter being clearly more specific. Which one would appear more often in the documents descriptions of a collection of restaurant reviews?

Exhaustivity, however, is described as how well a document description (consisting of several terms) embodies all topics found in a document. A description containing more terms would therefore provide greater exhaustivity, since every topic could be assigned a term in the description. Yet, if a description consists of too many terms, a query could retrieve this document (based on the exhaustive description) without it being ultimately relevant for the query. Additionally, ample descriptions are more likely to include less specific terms, as they are employed more often. Interpreting this line of thought in a statistical manner leads to the following conclusions:

- *the exhaustivity of a document description can be quantified by the number of index terms it contains*
- *the specificity of a term can be quantified as an inverse function of the number of documents in which it occurs*

(Baeza-Yates and Ribeiro-Neto, 2011)

The IDF weight gauging term specificity is consequently defined as

$$idf_i = log\frac{N}{n_i} \tag{4.18}$$

where idf_i is the IDF weight of term i, N is the total number of documents in the set and n_i the number of documents in the set in which the term i occurs. Note that now the whole document is considered as its own description to evaluate the specificity of all terms in a collection. For smoothing around extreme values of n_i, Baeza-Yates and Ribeiro-Neto suggest a slight alteration of the above formula:

$$idf_i = log(1 + \frac{N}{n_i}) \tag{4.19}$$

Table 4.3 continues the example started earlier and illustrates the calculation of the original IDF weights for the exemplary set of documents.

Table 4.3.: A minimalistic example for the calculation of IDF weights is shown (log in base 2), the total number of documents, N, is 3. Table 4.1 displays the corresponding set of documents. The IDF approach considers less used terms to be more specific, "read" only appears in one document and in consequence has the highest IDF weight. The term "long" appears in all three documents and is not considered to be any specific at all.

Terms	n_i	idf_i
very	2	0.18
long	3	0
text	2	0.18
read	1	0.48
...		

Term Frequency - Inverse Document Frequency

Combining TF weights and IDF weights was first recommended by Salton and Yang (1973). Taken above mentioned definitions, TF-IDF weights can be calculated by multiplying both weights as explained by Baeza-Yates and Ribeiro-Neto (2011):

$$w_{i,j} = \begin{cases} (1 + log f_{i,j}) * log(\frac{N}{n_i}) & \text{if} \quad f_{i,j} > 0 \\ 0 & \text{otherwise} \end{cases} \tag{4.20}$$

Several variants of TF-IDF weights exist in the scientific literature, both TF and IDF weights are often adjusted to match the field of application.

The benefits of consolidating the two systems are obvious. Higher weights are initially assigned to terms that occur repeatedly within a document, they most likely indicate the main topics that are being discussed. Nevertheless, these weights are regulated by their terms relative frequency in all other documents of the set, rarely overall occurring words are more specific and thus a better differentiator from other documents. In the end, high weights only remain for terms that exhibit two noticeable characteristics:

- the terms appear frequently in the currently observed document
- the terms appear infrequently in all other documents of the set

In view of this, it becomes obvious that TF-IDF weights can be calculated without any prior knowledge of the documents contents. This deems the weighting scheme to be especially well suited for the evaluation of unfamiliar documents, making it a prime candidate for, to the best of the authors knowledge, first large scale analysis of reddit's submission titles over five years.

In table 4.4 the last calculation step to obtain the final TF-IDF weights in the small example accompanying this section can be observed.

4.7. Dataset

This section will shortly the describe the acquisition of the data used for the research done in this thesis. Additionally, a brief overview of

the data set itself, including a detailed description of its parameters, is given.

4.7.1. Collection

Initial crawling of reddit's API[4] with the help of my advisor Philipp Singer and analysis of the acquired data proved fruitful. But soon the limitations of the API seemed to hamper any enthusiastic aims of providing truly comprehensive insights into attention patterns on reddit based on more than tiny sample sizes.

After first contact on the subreddit */r/TheoryOfReddit*, Jason Baumgartner was so kind to provide an immense dataset that was used for all research done in this thesis. Jason Baumgartner, also known as /u/stuck_in_the_matrix on reddit, is the owner of *RedditAnalytics*[5], which is work in progress during the writing of this thesis, but aims to be a portal offering exhaustive data, both archived and live, about reddit to academic researchers. Without his generous contribution many of the observations presented later in this thesis would not have been possible to this extent.

Table 4.4.: The resulting TF-IDF weights obtained from the TF weights in table 4.2 and the IDF weights in table 4.3 are displayed. The term "very" is the most useful term to index document 1, "read" for document 2 and "text" for document 3 (compare with table 4.1 and note that some of the terms not shown in this example might have higher weights).

Terms	$tf_{i,1}$	$tf_{i,2}$	$tf_{i,3}$	idf_i	$w_{i,1}$	$w_{i,2}$	$w_{i,3}$
very	3	2	0	0.18	0.54	0.36	0
long	2	2	2	0	0	0	0
text	2	0	2.585	0.18	0.36	0	0.47
read	0	1	0	0.48	0	0.48	0
...							

4 http://reddit.com/dev/api

5 http://redditanalytics.com

4.7.2. Scope of the data set

While originally spanning from late 2007 to mid 2013, the data was limited to all posts submitted between January 2008 and December 2012 for the sake of having complete years for more intuitive analysis and viewing.

The resulting data consists of close to 60 million user submissions generated over five years. The information accumulated in a single submission is listed in section 4.7.3. The data has only been collected after the submissions have been frozen by the systems, no change to the values of a submission is possible after this point. Figure 4.8 shows a preliminary dissection of the available data. The raw numbers collected from the data are available in table 4.5.

Already note a few interesting aspects, as their implications will be discussed later in more detail. In total, there are close to three times as much posts pointing to external links than there are self posts. Links going to nearly 2 million different domains are posted, but the top 100 domains, including the fictive domain *self*, which covers all self posts, (about 0.00005% of all domains) cover over 69% of all submissions. The effect is similar with subreddits, but by far not as strongly developed. Here, the top 504 subreddits (about 0.8% of all subreddits) are the target of roughly 82% of all submissions.

Table 4.5.: Raw numbers extracted from the dataset complemented with statistics. The top 504 subreddits are the subreddits which have a minimum of 10.000 submissions.

Submissions	58.874.227		
Link posts	43.894.520	74.6	% of all submissions
Self posts	14.979.707	25.4	% of all submissions
Unique authors	4.910.850	12	posts on average
Distinctive domains	1.841.239		
Posts to top 100 domains	40.772.856	69.3	% of all submissions
Distinctive subreddits	125.662		
Posts in top 504 subreddits	48.191.547	81.9	% of all submissions

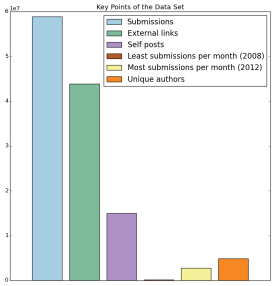
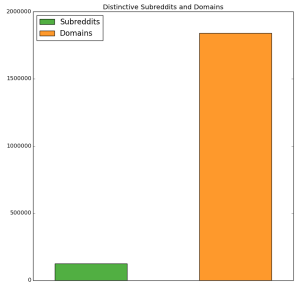
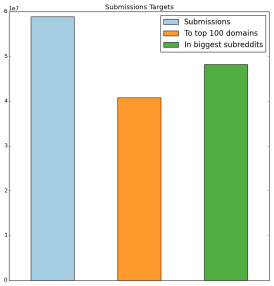


Figure 4.8.: Key points of the data set are shown. Submissions denotes the total number of submissions in the dataset. External links are submissions that submitted a link, self posts on the contrary are submissions that were text posts on reddit (see chapter 2 for more details on this). Followed by the submission count in reddit's weakest and strongest month during the time span of the dataset. To put these counts in perspective, the number of unique users submitting content to reddit is displayed.



(a) Number of distinguishable subreddits and domains



(b) Target of submissions by top domains and subreddits

Figure 4.9.: In figure 4.9a the number distinctive subreddits and domains collected from the data set are shown. Figure 4.9b shows how many submissions are either including a link to one of the top 100 domains and/or are posted in one of the biggest subreddits. The biggest subreddits are defined as having over 10.000 submissions, the data set counts 504 of those.

Figure 4.10.: Submissions on reddit are shown. The labels mark some of the important data points for this thesis. The upper submission links to an external picture, while the lower submission is a text post (self submission). The blue box provides additional information for a submission. Some, not self-explanatory, labeled elements are elaborated in table 4.6.

4.7.3. Description of submission data

As mentioned in the previous sections, the data set consists of nearly 60 million submissions. These submissions also represent the closed data points for the following research in this work. All feasible information is encoded in the submissions themselves, there is no additional meta data.

To get a better understanding of the information that is accessible per data point, take a look at the submission in figure 4.10. Table 4.6 explains some of the annotations and lists data that is available but not perceivable by looking at the figure.

Additional data about a submission is given, but not used in this thesis. Some of this information could proof useful in future research though.

Table 4.6.: Some of the information provided by a single submission data point and their explanations. See figure 4.10 for a more complete collection.

Domain (external)	Domain of a posted external link
Domain (self.subreddit)	Internal domain of a self post, consisting of self (dot) subreddit
Text	Text is only availabe for self posts, posts to links can not include text
Number of comments	Only the number of comments is provided, but not the comments themselves
Time	Time of creation in UTC
Link	Exact link that was posted
NSFW Flag	A flag denoting a submission as "Not Safe For Work"

5. Experimental Setup

In this chapter the experiments conducted on the available data (see section 4.7) are explained thoroughly. An overview of the general approach for all experiments and investigations is given in section 5.1. To answer the research questions posed in section 1.2, the scientific methods described in chapter 4 have been assembled and adapted to fit the needs of this thesis. Which methods are used for which examinations and their aptitude to answer the research questions is discussed section 5.2. The remainder of this chapter serves the presentation of the actual experiments.

5.1. Preliminary differentiation

Attention on reddit is examined in several dimensions. Two of these dimensions are motivated directly by intrinsic features of the online portal and can therefore easily be found in the data itself. These two dimensions are the *subreddits* to which submissions are posted and the *domains* to which the those submissions link to. The third dimension specifically involves the *type of content* submitted to reddit and has to be constructed beforehand.

5.1.1. Subreddits

Subreddits are comparable to moderated subforums in many other community board. They were designed to cover distinct topics and often have a rule set defining which type of content or submissions are acceptable to the community at hand. These rules are generally

checked by moderators, who have to power to filter non conforming submissions. The users reading, voting and commenting on a subreddit are assumed to have certain expectations of the content that is submitted.

Consider the two subreddits *r/aww*[1] and *r/politics*[2]. While the first is mostly about sharing cute pictures or videos of pets, the latter is concerned with political news and information in thethe United States. It seems obvious that these two sub-communities value different content and maybe even interact differently with the portal itself, the behavior and appreciation of users might be influenced by the subreddit they are currently observing. Subreddits therefore form one of the dimensions in which attention on reddit is analyzed.

5.1.2. Domains

Domains provide a catchy and more human-readable way of addressing services, networks or machines on the internet, as they are used to mask the underlying Internet Protocol. The names of these domains are often chosen in a manner to convey the owner or purpose of the addressed service or application.

For example, the domain *quickmeme.com*[3] points to a service, where users can *quickly* create and host their own *memes*. Domains consist of several levels separated by dots. The leftmost part of such a domain is the so-called "top-level domain", they consist of either generic acronyms, such as *com* (Commercial), *org* (Organization) and *info* (Information) or country-codes, such as *uk* (United Kingdom) and *at* (Austria). In the case of the aforementioned example, the top-level domain is *com*. The second level of the domain often already illustrates the purpose of the service, as described earlier, but more levels are possible.the

1 http://reddit.com/r/aww
2 http://reddit.com/r/politics
3 http://quickmeme.com/

As it happens, many services offer their own link shortening services, probably influenced by the vast popularity of micro-blogging services like Twitter. The much shorter domain *qkme.com* is still pointing to *quickmeme.com*. Additionally, some services use different domains to host their media content, any photographs on the social image hoster *Flickr*[4] are actually served from the domain *farm1.staticflickr.com* (or similar). These domain names are therefore consolidated to facilitate observations for the purpose of this thesis.

The consolidations include the merging of domains ultimately pointing to the same service with different top-level or second-level domains, unifying hosting and shortened URLs and generalization of domain names that only differ by a coded prefix (like the language codes in the different blogspot domains). An exemplary list of these operations can be found in appendix A.2.

Submission can easily be categorized by the domains they point to. The existence of self posts, self-written text, not hyperlinking other content, enables a clear distinction between content self-published directly on reddit and submissions pointing to external content.

For this reason, self posts, which technically have no domain, are labeled by reddit with the fictive domain *self.subreddit* and are regarded to be of the domain *self* for this thesis. It is important to reiterate that no self posts, but only link posts can accumulate karma (see section 2.2 for more details), as this fact could possible influence the motivation for posting a submission in one of the two ways as well as the resulting perception of the submission.

After this preparatory work, the domains of submissions can be utilized to identify the amount of external and internal content, prominent external targets of submissions and how much and what kind of attention they receive. Consequently, domains constitute another dimension in which examinations of attention can be differentiated.

4 `http://flickr.com/`

5.1.3. Type of Content

Both subreddits and domains are very specific points of views for the aggregation and development of attention. Building a foundation for a better overall understanding of recognition and perception of content on reddit requires to somehow classify the *types of content* that are submitted, regardless of particular subreddits or domains. In section 5.3 a method is presented how to divide submissions into various content categories, based on the available information. This categorization scheme has been developed together with Clemens Meinhart, as mentioned in section 1.5.

5.1.4. Scope of the Experiments

While most experiments are performed on complete data set as described in section 4.7, some are only conducted on specific subsets of the submissions. Experiments specifically comparing subreddits just consider subreddits with over 10.000 submissions, still around 82% of all submissions.

Likewise, some experiments involving domains, especially the categorization of content in section 5.3, evaluate only data from the top 100 domains, which cover roughly 70% of all submissions. See table 4.5 for the exact numbers.

Some of the investigations are limited to the 20 subreddits or domains with the most submissions overall, as they build a good base to survey general user behavior on reddit while keeping the extent of the discussed parameters manageable. The notation *top 20 subreddits* and *top 20 domains* always refers to these 20 subreddits and domains throughout this thesis, with the exception of the classification experiment in section 5.9 where the top 20 subreddits are only calculated from 2012.

5.2. Research Questions

In this section a recapitulation of the research questions posed in section 1.2 is provided together with a brief overview of how these will be tackled by the methods already demonstrated in chapter 4.

i. Can attention indicators other than score be leveraged for better understanding of user attention on reddit?

After examining the available data and some preliminary empirical observations, other indicators for attention have been determined. They can differ greatly between distinct subreddits, domains and types of content. A description of these indicators is found in section 5.4.

i. Which are these indicators, how are they characterized and do they correlate with each other?

Section 5.5 explains how basic characteristics were identified and section 5.6 clarifies how the examination of synergies and relations of the previously discovered attention indicators is constructed.

iii. Did these indicators and the users' perception of submissions and content evolve over the time?

Section 5.7 illustrates several ways how the development of the users perception of content can be tracked over time, including entropy measurements to gauge the equality of distribution of said attention metrics.

iv. Are these indicators and their relations globally applicable or limited to certain parts of the platform?

Combining the results of the experiments and observations in section 5.5 and section 5.6 allows to answer if stark discrepancies in the attention indicators exist, when only considering specific segments of reddit in comparison to the platform as a whole.

Furthermore, two approaches were taken to discuss if implicit or explicit factors, influencing and inducing these attention indicators, can be identified and evaluated.

Implicit factors are expected to be found in the structure of reddit and in a submission itself. Therefore, the combination of the title of a submission with the subreddit it was posted to is assumed to play an important role in the detection of these factors. The experiment to validate such claims is described in section 5.9.

Explicit factors could also be expressed as external circumstances that are not intrinsic to a submission or reddit as a portal, but rather factors that influence attention indicators outside the scope of presentation, content or (sub-)communities. Time, to be more precisely, the time of the day a submission was posted, is suspected to be such a factor, section 5.8 outlines the steps taken to estimate the influence of the submission time.

Above mentioned implicit factors, submission titles and subreddits, are additionally observed over the course of a year to determine fluctuations in word usage to possibly recognize trending terms. To complement this analysis, the coefficients of the classification experiment in section 5.9 are re-utilized for the same purpose. The construction of both approaches is defined in section 5.10.

5.3. Types of Content

Adding another dimension aside from the actual domains and subreddits was necessary to provide an overview of users perception and attention towards content on reddit on a more abstract level. To this end, the available information in the data set was used to craft a categorization scheme for submissions. This was done in collaboration with Clemens Meinhart, as described in section 1.5.

It is very difficult to identify the content of a submission automatically. Subreddits often have rules how submissions should be posted and what they should cover content-wise. But it is not clear if this is

consistently enforced and how many subreddits actually have such regulations, as their descriptions and rule sets can only be retrieved manually. The title of a submission can also indicate the type of content. But again, it is not guaranteed that this is case for a significant part of all submissions. Domains, on the other hand, can easily serve the objective to classify the content of a submission. A domain represents the type of content behind the hyperlink of a submission on a very high level of abstraction. The domain points to the service on the web and in many instances the purpose of this service can be clearly determined. This allows to assign a type of content to each service or domain.

Assuming one of the predefined types of content is *video*, a brief example helps to better understand the reasoning behind this idea. An arbitrary submission links to a video clip on the domain *youtube.com*. After visiting this domain manually on the web, it becomes evident that the domain points to *Youtube*[5], a popular and well-known video platform. The type of content promoted by above mentioned submission is therefore assigned to the content category *video*. So by visiting the service behind the domain an estimation for its purpose can be established.

Six types of content were constructed for this thesis. They are **image**, **video**, **audio**, **text**, **self** and **misc**, all of them are specified below. Since it is not feasible to visit close to two million domains manually to estimate the purpose of the service behind them, a compromise was made to just assign the 100 most used domains to a content category.

This approach still covers close to 70% of all submissions (which translates to roughly 40 million submissions), as illustrated in section 4.7. The manual assignments can be found in appendix A.3. The total number of times each type of content was assigned to one of the top 100 domains is shown in table 5.1.

Note that there are some domains for which the purpose or typical use-case (from the view point of reddit) is not clearly distinguishable.

5 `http://youtube.com`

Table 5.1.: The number of domains out of the top 100 that were assigned to each type of content are shown.

Content type	Domains assigned
text	64
image	17
misc	9
video	6
audio	3
self	1

Considering the example of Youtube, some videos on the platform are just uploaded to share music, as evident by the black or motionless video feed accompanying these clips. Likewise, some subreddits concerned with music or specific music genres sometimes link to Youtube when posting songs or music. Reliably differentiating between these two use-cases of *video* and *audio* is hardly possible though. Subsequently, it was decided to assign submissions linking ambivalent domains to the content category of their main purpose, in the case of Youtube: *video*.

The domain *reddit.com* poses a similar problem. Part of the submissions under this domain link to blog posts of reddit's staff, others to comments made by users directly on reddit. The content of these submissions can technically not be considered to be external, but the way they were posted is still distinctively different from an actual *self* post (also regarding karma accumulation). Combined with the divergent interaction that is needed to read these submissions, they are presumable perceived differently from ordinary self posts. Submissions with this domain are therefore considered to be of the content type *text*.

So while not a perfect classification of the content submitted to reddit, to the author's best knowledge, it is one of the best possible faithful and accurate estimations regarding the limited information available.

Both the construction as well as the assignment process was done

in collaboration with Clemens Meinhart (see section 1.5 for more details).

The types of content are now described in more detail, throughout this thesis they are also sometimes referenced as *content tags*, for example *image tag* signifying the type of content *image*. Table 5.2 summarizes the descriptions briefly with exemplary domains for each tag.

5.3.1. image

Under this content type every domain is collected that serves images, photographs and pictures in any form. Some domains point to dedicated image hosting services, *imgur*[6] is one of those. This hosting service is particular interesting as it has been created specifically for reddit by the user MrGrim (2009). Others, mostly social networks like Facebook and Flickr, have dedicated servers for image hosting and these domains were assigned accordingly.

Media in the *Graphic Interchange Format* (GIF) is considered to be part of the image category. Memes were also chosen to be represented as content type image, although the message of a meme is often transported via text printed on a repeatedly used background picture. This has two reasons, first, not every meme is hosted on dedicated services like *quickmeme*, but often on regular image hosting services. And secondly, memes are considered by the author to transport information in the same way as ordinary images, their intent or message is grasped quickly and does not need much reading effort to understand, unlike a news article or blog post.

5.3.2. video

The *video* tag is applied to any domain with the main purpose of serving or sharing videos. This includes, but is not limited to,

6 `http://imgur.com/`

popular portals like Youtube and *Vimeo*[7] but also streaming services like *Twitch*[8].

5.3.3. audio

Every domain providing means to discover, share or listen to music or any audio tracks in general. These are often social networks like *Soundcloud*[9] and *Bandcamp*[10].

5.3.4. text

The content type *text* includes every service that transports information mainly by written text. Online portals of news papers like the *Huffington Post*[11], scientific articles or papers, blogs on *Wordpress*[12] and encyclopedia entries on *Wikipedia*[13], just to name a few. Domainwise this is the largest category of the six, 64 of the top 100 domains are assigned to this type of content.

5.3.5. self

The *self* tag covers every self post. This distinction from external content is strictly defined by reddit's design and has not been classified manually. In terms of perception and attention behavior self posts could on one hand be regarded similarly to text submissions, considering the comparable effort in consumption, but on the other hand no karma can be accumulated from them.

---

7 http://vimeo.com/
8 http://twitch.tv/
9 http://soundcloud.com/
10 http://bandcamp.com/
11 http://huffingtonpost.com/
12 http://wordpress.org/
13 http://wikipedia.org/

5.3.6. misc

The category for *miscellaneous* domains spans everything that has not been covered by one of the distinct types of content above, this includes mostly link shortening services or file hosting and web publishing for which the final type of content was impossible to derive in retrospect.

Table 5.2.: The types of content arecorrect shown with a short description and a two exemplary domains. Note that domains of content type *self* are not actual domain names but rather supplied in that form by reddit's API.

Type	Content	Exemplary Domains
image	Photographs, pictures, GIFs, memes	imgur.com quickmeme.com
video	Videos	youtube.com vimeo.com
audio	Audio tracks	soundcloud.com bandcamp.com
text	News, articles, blog posts, mainly textual sites	nytimes.com blogspot.com
self	Self posts	self.AskReddit self.funny
misc	File hosting, link shortening services, everything else	tinyurl.com amazonaws.com

5.4. Indicators of Attention on reddit

Many online services measure the received attention of resources by *pageviews* (how often a certain page or content was viewed by users) but also by the amount of *interaction* with said content. The presumption is that the more interesting a resource is, the more likely users are to view it or even get involved and interact with it.

Reddit offers no data about actual views by users besides the sum of unique visitors per months, which allows no inference towards the pageviews of an individual submissions. In the related work concerning reddit in section 3.2, Gilbert (2013) proposed an approach to estimate pageviews by relaying the problem to a popular image hosting service, which provides this metric. However, this method is not applicable to reddit as whole, but just to limited parts, since only images provided by this hosting service can be compared. Additionally, the hosting service in question has long developed into a full-fledged social network in its own, skewing any data retrieved for this purpose.

It has therefore been decided to focus on attention generated from interaction in this thesis. This decision excludes *lurkers*, as defined by Nonnecke and Preece (2000), from the analysis but instead only considers users that are involved enough to draw on the various ways of interaction offered on reddit. Owing to reddit's design, this also reveals the users appreciation or disapproval and their general perception of content. As already mentioned in chapter 2, users can, amongst other things, up- or downvote and comment on any submission.

Before diving into the different indicators of attention, it is important to reiterate that reddit's API only reveals a snapshot of these metrics, in the case of the available data, it is the final value after the submissions have been archived, preventing any further interaction.

5.4.1. Score

Although up- and downvotes build one of the cornerstones of user interaction with reddit, only their combination (precisely the subtraction of downvotes from upvotes), also known as *score*, is utilized for the default ranking mechanism. The *freshness* of content is another factor influencing the ranking and is detailed in section 2.1 Simply put, the newer a submission and the higher the score, the higher the ranking. the attention and perception of submissions

This admittedly plain ranking measure bears the pitfall that a submission with similarly large amounts of upvotes and downvotes will never be ranked very high, despite the immense attention it receives. Still, score is the most important attention indicator on the portal, since a high ranking ensures that a submission is placed on the first page of its subreddit or even on the frontpage. This increases the submissions visibility drastically, facilitating even more votes.

5.4.2. Upvotes and Downvotes

Upvotes and *downvotes* directly represent the interaction process of logged-in users with reddit. They also directly embody the appreciation or disapproval of content. Users can only vote once per submission, but they are able to retract their vote or change their vote within a given time-frame.

These votes are the only way users can actively influence the score of a submission, all other (arguably less important) factors are out of their control. Furthermore, the voting process does not require much effort, it is merely clicking on one of two buttons. It is very rare for an online social network or platform to implement a mechanic that allows to show disapproval with the same power as the opposing appreciation. Hence, exploring the relation of those two metrics is very intriguing.

5.4.3. Votes

Votes is an attention indicator constructed for this thesis, it is logically derived by adding the downvotes of a submission to its upvotes. This measurement could also be described as the total *engagement* a submission receives in terms of voting interaction. Following this reasoning, the metric does not evaluate which kind of attention a post receives, but *how much* attention it receives overall. Subsequently, this approach avoids the problem of *embezzling* high attention submission with equal amounts of upvotes and downvotes, as score does, but

on the downside loses the ability to display if this engagement is of positive or negative nature.

5.4.4. Comments

The comments themselves are not part of the accessible data and collecting them on a similarly large scale was not possible within the time scope of this work due to rigorous request limits on reddit's API. Still, the information how many comments a submission received is available, but there is no knowledge about the sentiment of these comments.

Writing a comment is assumed to require more effort than simply voting on a submission. Presumably, a user has thought about the given submission, maybe read the previous comments prior to composing her or his own, or has to come up with some sort of idea for her or his writing, as opposed to just clicking on a button. Comments, therefore, illustrate an entirely different form of engagement than the indicators described above. It will be interesting to learn if a divergence of this attention indicator from the others is observable in the following examinations and experiments.

5.5. Characteristics of Attention on reddit

For a better understanding of how the identified attention indicators are characterized and how they relate with each other, an empirical analysis has been conducted. The observations were done mainly regarding score, number of comments and number of votes a submission receives and these indicators are examined in the three dimensions described in section 5.1, namely subreddits, domains and type of content.

5.5.1. Attention per Subreddits, Domains and Type of Content

To get an initial idea of the most voted, most discussed and most appreciated parts of reddit, some preliminary numbers have been extracted from the available data. The total and average score, number of votes and number of comments per submission have been calculated differentiated for the top 20 subreddits, top 20 domains and each type of content. The attention indicators are calculated twofold, as average per submission and as total values. This facilitates two view points at the data. Total values enable to judge the distribution of interaction and attention over the whole portal, while the submission average allows a better comparison between the specific segments (be it subreddits, domains or type of content).

The results are presented in section 6.2.

5.5.2. Relation of Score and Number of Comments

As explained in section 5.4, score is crucial for reddit's ranking of submissions and directly influences the submissions position and visibility on the site. Comments express a separate kind of attention towards submissions that, despite being a detrimental part for discussions, is not utilized for ranking. It is therefore interesting to inspect the relation of those two indicators.

Section 6.2.4 illustrates the outcomes of this investigation.

5.5.3. Relation of Upvotes and Downvotes

The relation between the upvotes and downvotes of a single submission receives is not highly informative, as it just represents a single voting instance and is influenced by a variety of factors, most importantly by the actual content itself. But observing all submissions regarding their up- and downvotes together, allows a look on the

bigger picture of reddit's voting behavior and thus on the general *sentiment* of reddit users.

In section 6.2.5 the nature of this sentiment can be inspected.

5.6. Correlation of Attention Indicators

The examinations above are just based on raw numbers extracted from the dataset, correlation coefficients on the other hand are a widely used and tried and tested scientific metric for determining the relationship between two variables. The assumption is that attention indicators exhibiting a strong correlation and are influenced either by the same factors or influence each to certain degree relating to the correlation coefficient.

For example, if the score and number of comments correlate strongly over all submissions in a discussion focussed subreddit, one could presume that an active discussion is essential for getting a high score in this subreddit. As mentioned in section 4.4 however, the interpretation of this results is entirely up the observer, as the variables are considered independent. Still, identifying such correlations between the attention indicators enables a much more informed discussion of the possible synergies of the indicators coupled with a comparison of existing relations or the lack thereof in specific parts of reddit.

Both the Pearson Product Moment Correlation and the Spearman Rank Order Correlation were calculated combined with significance testing of the resulting coefficients. This approach was taken to facilitate the distinction between linear and non-linear correlation and possibly detect the rate of outliers present in the sample.

Aside from the attention indicators already mentioned, another metric was introduced in these calculations. As discussed in section 5.2, the title of a submission is one possible implicit factor influencing the perception of a submission. In adaptation to the scientific method at hand, it was decided to simplify the title to its length, based on the reasoning in section 4.6.1.

The threshold for significance testing is set to 0.05.

Additionally, the correlations were calculated in several separate setups:

- all submissions
- only self submissions
- only submissions of content type text
- only submissions of content type image
- submissions in *r/AskReddit*
- submissions in *r/worldnews*
- submissions in *r/funny*

The caluclation of the correlations for submissions in content type *self*, *text* and *image* were included as these are the categories which receive most of the attention. The distinction between the three subreddits was made to exemplary represent a discussion based subreddit with only self submissions (*r/AskReddit)* and contrastingly an image based subreddit (submissions in *r/funny*) and a typical text based subreddit (*r/worldnews*).

The results of this analysis can be found in section 6.3.

5.7. Development over Time

Section 5.7.1 discusses the computation of the evolution of the attention indicators from the perspective of subreddits, domains and type of content, while section 5.7.2 introduces how the entropy of said attention indicators was tracked over time.

5.7.1. Development per Subreddit, Domain and Type of Content

To study the development of the attention indicators, an approach similar to section 5.5 was taken and extended to be calculated for ev-

ery month. The results were combined to form a continuous view of the evolution of users perception and attention towards submissions and content on reddit.

This was done again for the score, number of comments and number of votes for subreddits, domains and types of content and likewise both total and average values were calculated.

The development of these indicators is described in section 6.4.

5.7.2. Entropy of Attention Indicators

Furthermore, the normalized entropy of all scores, upvotes, downvotes, votes and number of comments was determined on a monthly basis. A graph depicting these values has been constructed to illustrate the development of the distribution equality or inequality of attention on reddit towards all submissions, towards subreddits and towards domains. A higher entropy value for votes signifies a more impartial dissemination of voting interaction over all submissions, while lower values illustrate a more focussed grouping on particular segments of the portal.

Derived from the formula (for non-normalized entropy) presented in section 4.5

$$H(X) = -\sum_{i=1}^{n} p_i \log p_i \tag{5.1}$$

the probabilities p_i for each measured attention indicator need to be established. The specific probabilities are described in the following sections.

The outcomes and are depicted in section 6.4.4.

Score Entropy

Since all indicators can be mapped down to a fixed number per submission, the probabilities are defined as follows. The probability for a submission with score s_i is

$$p_i = \frac{s_i}{\sum_{j=1}^{n} s_j} \tag{5.2}$$

where the denominator represents the total score of all submissions observed. All other indicators are calculated accordingly.

A high score entropy indicates that the appreciation of resources on reddit is fairly equally distributed, while lower entropy values suggest that most of the appreciation is only declared towards a limited part of the system.

Upvotes and Downvotes Entropy

The entropy of the up- and downvote distributions are indicative of a similar proposition as the score entropy. The difference here is that explicitly the positive or negative sentiment is examined (contrasting to the general sentiment). Again, high values suggest balanced dissemination of either positive or negative voting, while low values give evidence to a fragmentation that is heavily dominated by just a few groups.

Votes Entropy

Following the definitions from above, voting entropy catches the parity of voting interactions throughout all submissions. Subsequently, high entropy values demonstrate an even allocation of all votes, low values consequently imply a lack thereof.

Comments Entropy

The entropy of the number of comments the submissions receive reveals whether all submissions are discussed in similar exhaustiveness (high entropy values) or if certain elements of reddit are debated more lively than others (low entropy values).

5.8. Influence of Submission Time

One possible factor explicitly influencing the perception and attention towards submission is assumed to be the submission time. These times are recorded on reddit in UTC (Temps Universel Coordonné or Coordinated Universal Time), any examination results are therefore presented aligned to this time zone.

Although a survey by Duggan and Smith (2013) in section 3.2 found that around 6% of adult internet users in the United States are visting reddit, it is not clear how many of them are actively participating in the voting or commenting process and which proportion they occupy in reddit's user base.

Additionally, reddit (2013) itself states that the site's visitors live in about 200 different countries. These reasons led to the decisions to choose UTC as the time zone for representation purposes. Still, some comparisons to Pacific Daylight Time (PDT) and Eastern Daylight Time (EDT) will be made in an attempt to explain the observed phenomena, as these are the most populated time zones in the United States.

Similar examinations have been conducted in blog posts by Singer (2013) and Olson (2013a). This work elaborates these ideas on a much larger scale with more detailed investigations of the attention indicators. Furthermore, the analysis is differentiated by selected subreddits and types of content.

Score, votes and number of comments are broken down per hour and analyzed in three different levels of abstraction:

- over the course of an average day
- over the course of an average week
- in an combined *heatmap* inspired by a post in *r/DataIsBeautiful* by /u/minimaxir (2013)

To introduce another angle of information, the number of submissions posted at any given hour were added to all observations.

These observations are illustrated in section 6.5.

5.9. Classification of Submission Titles

Submission titles are arguable the most vital part of any submission on reddit. The title is the first point of contact between a submission and a user. The content or intent of a submission has to be expressed through the title, as it is the primary way of presentation for any submission. Looking back at the frontpage of reddit shown in figure 2.4, even image submissions only display a small thumbnail of the actual content, the title is still more prominently positioned and formatted.

The assumption of this thesis is that a user may therefore presumably deem a submission uninteresting and even skip it without looking at its content by just evaluating the title, no matter the type of content. Inversely, an subjectively interesting title could spark a user's interest in a submission. This is not a particularly daring assumption, as news papers, or any textual media, meticulously craft their headlines (or titles) as it is their primary method of provoking potential readers into consuming their content.

It can also be assumed that users browsing a specific subreddit have a certain expectations toward the topics or types of content covered by the submissions listed there. The submission titles in political subreddits should therefore presumably reflect appropriate subjects, a submission with a non-descriptive or off-topic title might be ignored, downvoted or even removed. Likewise, users of the

subreddit *r/funny*[14] might rather cherish playful and amusing titles than anything more serious.

Over the time these sub-communities probably developed very community-specific abbreviations, acronyms and contingently even their own way of expressing certain situations or matters. It stands to reason that authors want to conform to these requirements in order to successfully promote a new submission.

To verify the claim that communities with their own customary idioms may have formed in subreddits, a classification experiment has been designed. In this experiment, the classifier needs to correctly assign submission titles to the subreddit the submission was posted to. If the results are expressive enough, it would also mean that an implicit factor influencing the perception and received attention of submission is inherent in the relation or combination of submission *title* and *subreddit* the submission was posted to.

As the evolution of such sub-communities their customary language is an ongoing process and could change drastically over time, only submissions from a single year, the year 2012, are considered for the following experiments.

Classifiers used were the Binary Naive Bayes classifier (BNB) and the Multinomial Naive Bayes classifier (MNB), as described in section 4.2. The experiment is conducted in the same way with both classifiers. The reasoning behind this is that the BNB classifier is in orders of magnitude faster and moreover less resource intensive than the MNB classifier, which on the other hand promises better results. If roughly comparable performance, measured by the metrics proposed in section 4.3, can be achieved, the former is preferable for future work based on this approach.

For this experiment the top 20 subreddits of the year 2012, excluding *r/POLITIC*[15], are considered the target classes. The aforementioned subreddit uses bots to mirror several politics-related subreddit. One

14 http://reddit.com/r/funny
15 http://reddit.com/r/politic

of them, *r/politics*[16], is also part of the top 20, the excluded subreddit would therefore skew the results with its duplicate submission titles and is substituted by the next most used subreddit *r/Minecraft*.

The training features consist of the submissions titles, transformed accordingly to either word-count vectors or TF-IDF vectors for each classifier. A grid search was used to identify the best set of parameters for both classifiers. Additionally, the, commonly used, *stratified 10-fold cross-validation* was applied for training and testing to ensure the quality of the results[17]. The stratified k-fold cross-validation guarantees that any fold contains roughly the same proportion of titles from all subreddits.

Furthermore, a baseline model was constructed by destroying and randomly rearranging the title-subreddit allocations. The classification was implemented again with a grid search for the best set of parameters and with stratified 10-fold cross-validation. This whole process was repeated 100 times and the performance measures of each trial run were averaged to judge the baseline performance for each classifier.

The results of this experiment are described in section 6.6.

5.10. Trend Discovery and Analysis

For the trend and popularity analysis two different approaches where formulated. Initially, trend discovery is based on the idea of TF-IDF described in section 4.6.2, but its application has been modified to accommodate the objectives of this thesis. As described in section 1.5, the same modified variant of TF-IDF is used by Clemens Meinhart. Nevertheless, his approach is applied in entirely different settings.

The temporal trend analysis is based on the classifiers trained in section 5.9. The compositions of both concepts are outlined in section

16 `http://reddit.com/r/politics`
17 Other choices for k produced very similar results

5.10.1 and section 5.10.2 respectively. As the trending terms identified with this methods possibly relate to real world events or situations directly affecting reddit, only submissions from the most recent year, 2012, are considered in these experiments to facilitate the validation of the results.

5.10.1. Modifying TF-IDF for Trend Discovery

When considering a set of documents, a high TF-IDF weight for a specific document suggests that the associated term appears often in the specific document but does not often occur in all other documents of the set. Thus, highly weighted terms are considered to be well suited to describe the contents of the document and decisively separate it from other documents in the collection.

This line of thought has been applied to reddit submissions and their titles in a temporal analysis. To identify terms that are descriptive for submissions posted during a certain timespan and simultaneously separate this timespan from the remaining submissions, TF-IDF weighting has been modified in two different ways.

Based on the analogy of a set of documents, the submission titles of one single month can be regarded as one single document for the purpose of the TF weights calculation. So all submission titles of a month m are combined into one single document d and the TF weights are computed. The formula to calculate the TF weight of a term i is adapted to

$$tf_{i,d} = \begin{cases} 1 + log f_{i,d} & \text{if} \quad f_{i,d} > 0 \\ 0 & \text{otherwise} \end{cases} \tag{5.3}$$

where d is a document comprised of the titles of all submissions in month m and every other variable remains as defined in section 4.6.2.

Analogous to the document example, the general implementation of IDF determines the number of documents a term appears in to ultimately estimate the term's specificity. Note that every single submission title outside of the month (or timespan) specified for the TF weights is now considered a document in its own. To calculate the IDF weights, the purpose of the two aforementioned approaches has to be established.

The first approach considers all other submission titles as documents for the IDF weights computation, so

$$idf_i = log(1 + \frac{T}{t_i}) \tag{5.4}$$

where T is the number of all submission titles outside the specified timespan and t_i is the number of all of these submission titles where the term i appears in. Every other variable remains as defined in section 4.6.2.

High TF-IDF weights calculated with this implementation indicate terms that are highly descriptive and significant for a specific timespan while also separating it from the rest of all recorded submissions. Put simply, these terms stand out and possibly mark events that only occurred in the specified timespan but at no other point in time, neither past nor future.

The second approach only considers the submission titles of the previous month for the IDF weights calculation, the formula is altered to reflect this.

$$idf_i = log(1 + \frac{M}{m_i}) \tag{5.5}$$

Here M expresses the number of submission titles in the previous month and m_i the number of all of these submission titles where the

term i occurs in. Every other variable again remains as defined in section 4.6.2.

In this case, high TF-IDF weights signify terms that are descriptive and rather unique to a specific month just in comparison to all titles of the previous month. This is resembling an actual trend behavior as just the timespan immediately before the specified one is studied simultaneously, but anything even farther in the past or anything in the future is not considered.

Experiments with both approaches are conducted for each of the top 20 subreddits and conclusively over all subreddits in the dataset combined.

The identified trends can be found in section 6.7.1.

5.10.2. Exploiting Classifier Coefficients for Trend Analysis

To determine trending terms in titles of submissions throughout a year, the classifiers of section 5.9 have been trained not on the whole year 2012, but on every single month. After each training, the coefficients were extracted. These coefficients roughly mark the importance of a specific term appearing in the title of submission for that title to be assigned to particular subreddit.

Thus, the largest coefficients for each subreddit can be assumed to be the most distinctive and descriptive terms for that subreddit. The ten largest coefficients of every month are tracked over all twelve months. The development of all coefficients which rank in the highest five at any point during the year are then visualized.

The resulting visualizations are demonstrated and explained in section 6.7.2.

6. Results

In this chapter, the results derived from and the observations, made while conducting the experiments described in chapter 5, are illustrated and further explained. In the following section a very basic look at the quantitative development of reddit over four years (the scope of the dataset) is presented. This will help to fully understand some of the underlying causes that may have triggered or influenced some of the uncovered properties in the succeeding sections.

6.1. Quantitative Development

During the years 2008 to 2012 reddit has undergone enourmous growth. This growth is analyzed in more detail by Clemens Meinhart in his thesis, as mentioned in section 1.5.

In figure 6.1 the number of submissions posted each month are illustrated. A brief recession of the portal in early 2010 can be attributed to the relaunch of reddit's main competitor at the time, Digg[1], and is clearly visibly by a sudden and substantial drop in submissions in just two months time. Despite this, reddit has still managed to grow excessively in the following years.

As a matter of fact, many properties of reddit have shared this extreme growth. In figure 6.2 the development of several other aspects are depicted, the number of domains used for link submissions, the number of active subreddits and the number of unique users posting submissions.

1 http://digg.com/

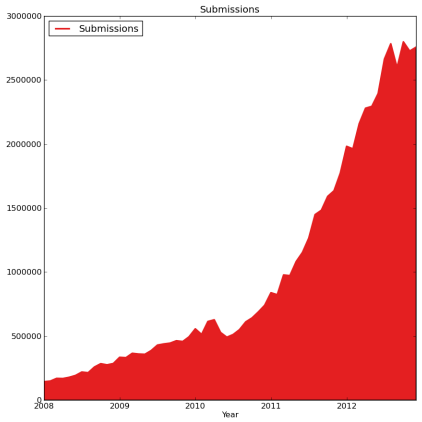


Figure 6.1.: The growth in submissions per month on reddit is shown from early 2008 to late 2012.

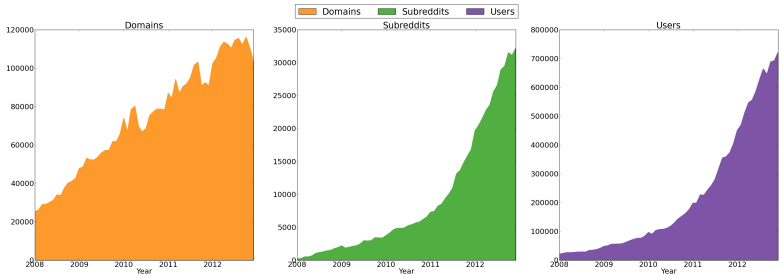


Figure 6.2.: The growth of distinct domains and subreddits in submissions is shown in the first two graphs. The third graph depicts the growth of unique users posting said submissions.

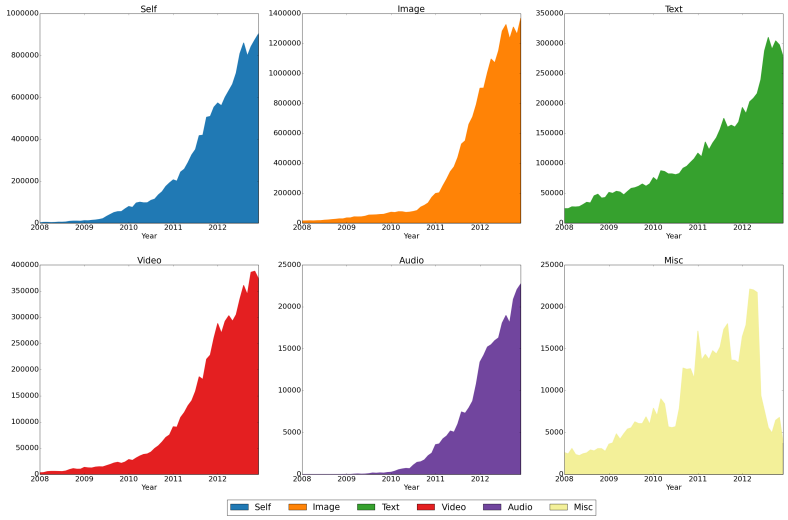


Figure 6.3.: The growth of submissions for each content category is shown.

While the subreddits and the users exhibit a growth pattern very similar to the number of submissions in figure 6.1, the number of domains increases in a rather linear way, as this is a measure also reliant on external factors, such as the availability or existence of specific services that can be linked to.

Figure 6.3 additionally shows the growth in submissions for each content type defined in section 5.3. Again, all types of content have experienced a surge analogous to overall development of submissions, although each to a different extent. The exception is *misc* with several significant drops, most of them are linked to the ban of many link shortening services over the time.

Summarizing, a massive growth of reddit in various aspects is observed and this information serves as a solid foundation for the discussion of some of the following results.

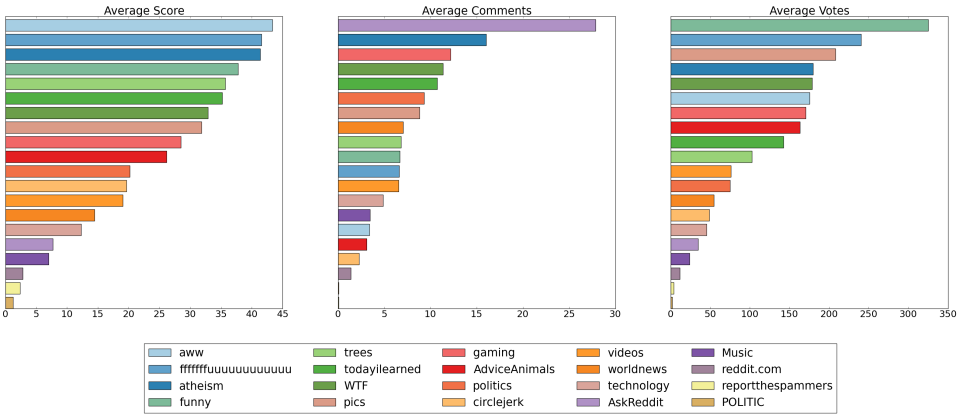


Figure 6.4.: The average score, average number of comments and average number of votes for a submission in one of the top 20 subreddits are shown.

6.2. Evaluating Attention Characteristics

Throughout the rest of the chapter, most of the examinations focus on the attention indicators *score*, *comments* and *votes*. To identify the characteristics of these indicators, their relation to the three dimensions *subreddits*, *domains*, *type of content*, as proposed in section 5.1, is investigated.

6.2.1. Subreddits

In figure 6.4 allows a look at the average attention a submission in one of the top 20 subreddits receives. Even within these top 20 subreddits, the subreddits with most submissions over the scope of the dataset, stark differences are observed.

The subreddits claiming the highest average scores are closely tied, *r/aww* and *r/fffffffuuuuuuuuuuu*, are both image based subreddits, the former encouraging submitting pictures of "Things that make you go AWW! – like puppies, and bunnies, and so on...", the latter a source for so-called "rage comics", illustrating infuriating personal stories

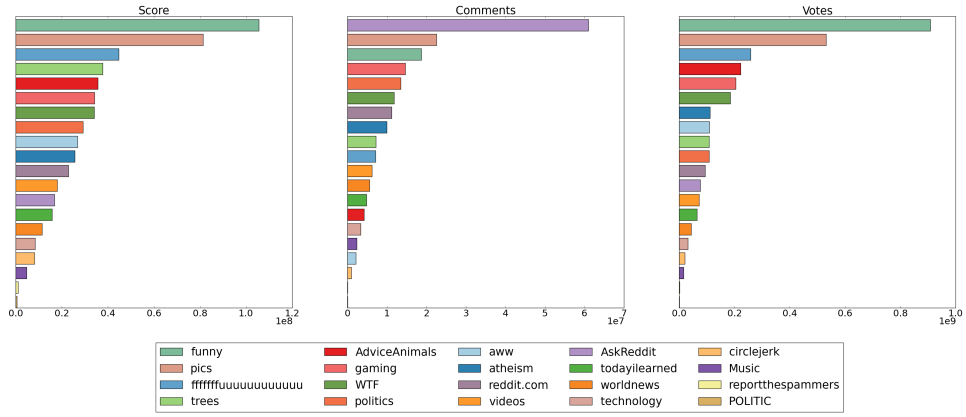


Figure 6.5.: The combined score, number of comments and number of votes received by the submissions in the top 20 subreddits are shown.

in short comic strips, mostly created by the authors themselves. The third contender is *r/atheism*, a diverse subreddit consisting mainly of discussions and news posts related to atheism and agnosticism.

On the other hand, *r/AskReddit*, a subreddit created exclusively for discourse via self-posts between *redditors* (users of reddit), boasts nearly double the amount in average comments per submission than the next one. The effect is not as pronounced when looking at the average number of votes, but still, a submission in *r/funny*, a place where users can submit anything that they consider funny, receives nearly a third more votes than in the follow-up subreddit.

The overall score, comments and votes per subreddit in figure 6.5 show that in fact, for all three observed attention indicators, a handful of subreddits receives most of the attention and user interaction. While the major recipients of votes and score, *r/funny* and *r/pics*, consist of mostly image submissions, the subreddit with the highest comment activity is again *r/AskReddit*.

Comparing the number of comments and number of votes, both average and total, it is already evident that submissions on reddit generally receive far more up- or downvotes than they receive comments.

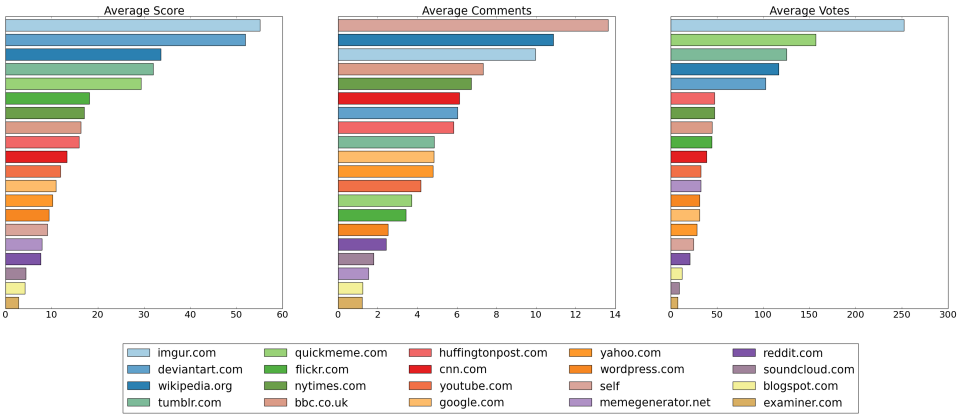


Figure 6.6.: The average score, average number of comments and average number of votes for a submission in one of the top 20 domains are shown.

6.2.2. Domains

The domination by a few subreddits translates over the values found for the top 20 domains on reddit. Average score in figure 6.6 is lead by *imgur.com* and *deviantart.com*, both image hosting services. Remember that the subreddits with the highest average scores feature mostly image submissions.

Interestingly though, while *imgur.com* is used for a variety of images, *deviantart.com*[2] is solely hosting (digital) pieces of art, from portraits and photographs to various paintings and even wallpapers or interface designs.

The average number of comments is domineered by the fictive domain *self*, containing all self-posts on reddit, but trailed by *wikipedia.org* and again *imgur.com*. This means that, on average, a wikipedia article or image is discussed with nearly the same intensity as a dedicated self-post. The average number of votes are again dominated by *imgur.com* and followed by *quickmeme.com*, a meme creation and hosting service.

2 `http://deviantart.com/`

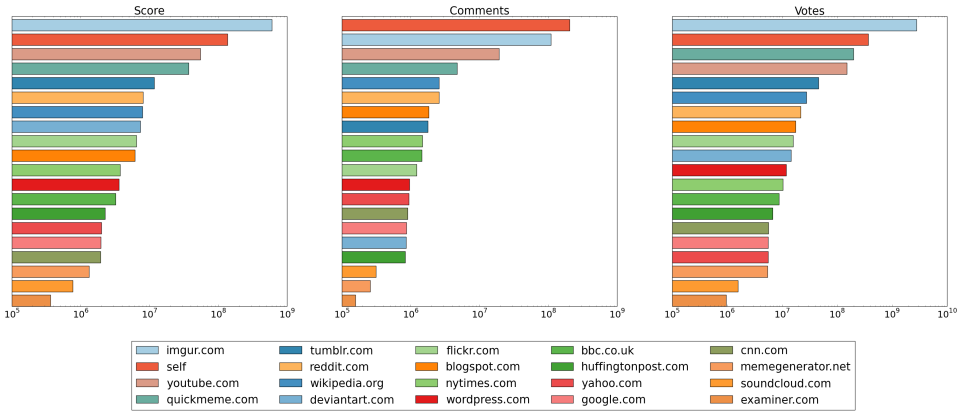


Figure 6.7.: The combined score, number of comments and number of votes received by the submissions in the top 20 domains are shown. Note, that for better visibility the bars are displayed in log-scale.

Figure 6.7 further emphasizes the role of *imgur.com* on reddit. Submissions linking to the image hosting service earn substantially more score and votes than all self-posts combined. Self posts still receive the most comments overall though.

6.2.3. Types of Content

Finally, the observations in subreddits and domains are elaborated by studying the attention received by each type of content defined in section 5.3. Figure 6.8 displays the average score, average number of comments and average number of votes for a submission in each category.

As already expected by the nature of the subreddits and domains dominating the attention indicators, an average *image* submission gets as much score as all other content types combined and receives even more votes than all other categories together. *Self* submissions get commented on slightly more than *image* submissions though.

Again, in comparison to votes, the average commenting activity is

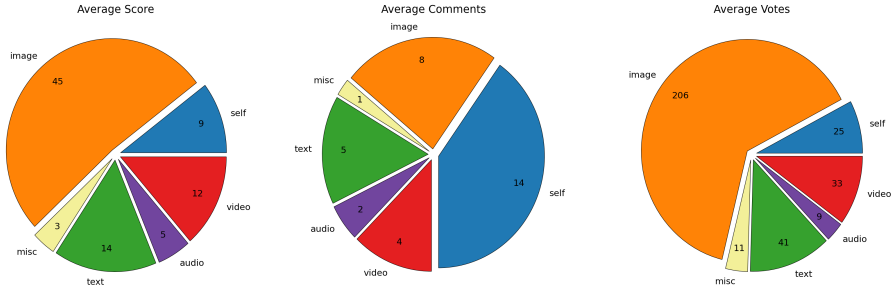


Figure 6.8.: The average score, average number of comments and average number of votes for a submission in one of the six content categories is shown.

quite low, *image* submissions, for example, receive 25 times more votes than comments, only *self* submissions seem rather balanced in this matter.

The lions share of *image* submissions grows even bigger when considering the overall amount of score, comments and votes for each content type in figure 6.9. *Self* submissions gather over half of all comments, but aside from that *image* submissions dwarf all other content categories in total attention received, especially *audio* and *misc* submissions collect merely thousandths of the total attention.

This is actually not utterly surprising, a look back at figure 6.3 reveals that these are also the content categories with the lowest numbers of submissions by far. Simply speaking, more submissions are able to collect more attention from the users overall.

Nevertheless, the average score and average votes per submission are still substantially higher for *image* submissions than any other type of content, despite having the largest proportion of reddit's posts.

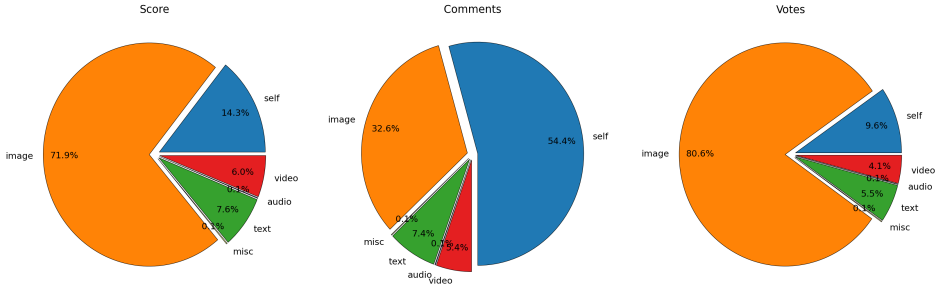


Figure 6.9.: The share of each type of content of the combined score, number of comments and number of votes over all observed submissions is shown.

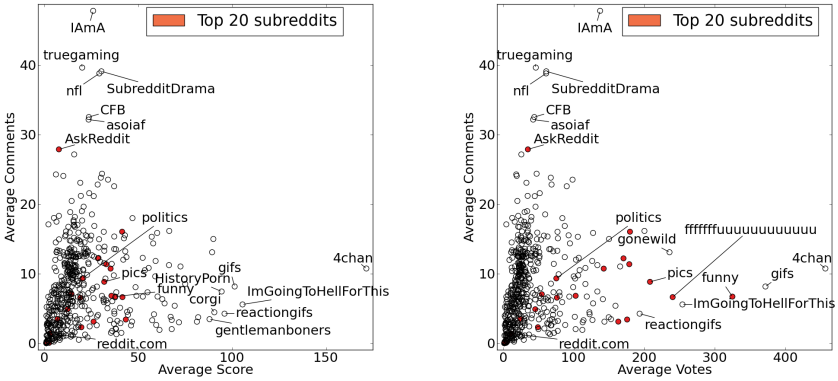
6.2.4. Differences in Perception and Attention Generation

The observations above only regarded one indicator of attention at the time to enable a more detailed comprehension of each indicator over all dimensions. The following examinations combine the attention indicators to facilitate characterising the attention generation and perception of submissions within the dimensions of subreddits and domains.

Figure 6.10a illustrates subreddits by the average score and the average number of comments a submission in one of the subreddits receives. The ten subreddits with the most submissions from 2008 to 2012 are labelled in red. This markedly shows that submissions on the most frequented subreddits accumulate just slightly higher scores than the plurality of subreddits. Regarding the average amount of comments per submission, only one of the larger subreddits, *r/AskReddit*, which was specifically created for user discussions, stands out.

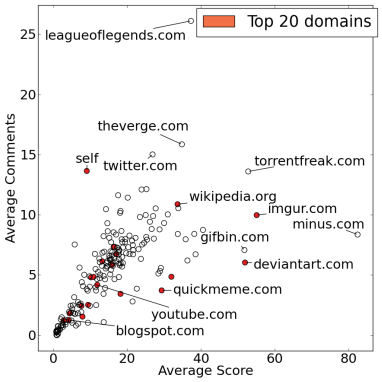
It is also interesting to find that, under the viewpoint of attention generation, there are effectively three types of subreddits. The majority of subreddits exhibit a comparably low amount of comments as

well as score. Some are leaning towards higher average scores, but still not many comments. Inversely, some subreddits seem to instil more discussion, manifested in more comments per submission, but, again, score remains relatively low.



(a)Score vs. Comments per Subreddit

(b)Votes vs. Comments per Subreddit



(c)Score vs. Comments per Domain

Figure 6.10.: Average score, average number of comments and average number of comments calculated over all submissions in each subreddit in figure 6.10a and figure 6.10b and average score and average number of comments for each domain in figure 6.10c. The ten subreddits and respectively domains with the most submissions are marked red to emphasize how most of the content on reddit is rated and commented on average. Subreddits and domains with extreme values are additionally labelled with their name.

Many of the subreddits with higher average scores were created to share pictures, GIFs or even adult content. The lone leader in average score is *r/4chan* by a considerable margin. This subreddit acts as a "best-of"-aggregator of stories on the highly anonymous community board *4chan*, which was already briefly described in section 3.1. No direct links to the portal are allowed, so the reddit community mostly resorts to submitting screenshots of noteworthy entries. The top subreddits are not far from the remaining subreddits in terms of score, yet they display average scores that are just high enough to distinctly separate them from the mass of low-comments, low-score subreddits.

The subreddits displaying a higher average amount of comments are often explicitly enforcing discussion. In *r/IAmA*, people can talk about their extraordinary jobs or experiences, questions and interaction with other users is already encouraged by the prefixes "IAmA" or "AMA" typically used in the submission titles there, denoting "I am a ..., Ask Me Anything" and simply "Ask Me Anything". The popularity of this subreddit has already been illustrated in section 1.1 and over the course of writing this thesis many more people chose to engage in this *unprecedented* type of interview, among them Tim Berners-Lee (2014) and Magnus Carlsen (2014).

Some other subreddits are used to talk about American football, serve as a place have serious discussions about computer games or theorize and conspire about "The Song of Ice and Fire", a very popular series of novels (and now TV-show). What combines these subreddits though, is a, more or less, strict moderation encouraging respectful discussions.

Figure 6.10b reveals roughly the same segmentation, although average score is here substituted with the average number of votes a submission receives. Many of the subreddits with high average score are still in similar positions regarding the average number of votes their submissions collect. Furthermore, the top subreddits are now even more clearly separated from the unnamed mass of remaining subreddits. A high number of votes seems to translate into a high

score, or the other way around, this is investigated in more detail in the following sections.

The subreddit *r/reddit.com* is, unsurprisingly, found within the thickest cluster of subreddits. It is one of the earliest subreddits and comprised most of reddit's submissions until more subreddits were created, so it seems only natural that the attention it receives reflects the majority of all subreddits.

Considering figure 6.10c, domains show tendencies very unlike the ones found with subreddits in figure 6.10a. The relation of average score and average number of comments of a submission linking to a specific domain seem much more linear here: the higher the score, the more comments. Of course, it could be the other way around, dependencies between these two indicators of attention are not clear.

The domains with primarily high average scores are, again, predominantly attached to the content type images, as most of them constitute image hosting services. Self posts are the principal method of discussion on reddit, so as expected, submissions under the domain *self* show a slightly higher number of comments then most other domains. The domain *leagueoflegends.com* seems to induce particularly long discussions. The *eSports* finals of the massively successful F2P (Free to Play) game *League of Legends* were watched by over eight million concurrent viewers[3]. The game also occupies a large community scrutinizing every detail of related eSports events and new updates to the game in dedicated subreddits.

Submissions pointing to the technology news site *theverge.com* and *torrentfreak.com*, a news platform "dedicated to bringing the latest news about copyright, privacy, and everything related to filesharing"[4], seem to be appreciated, indicated by the above average scores and more intensive commenting. This further emphasizes the interest of reddit's community in technology and copyright management, as already briefly noted in section 1.1 of this thesis.

3 http://theverge.com/2013/11/19/5123724/league-of-legends-world-championship-32-million-v
4 http://torrentfreak.com/about

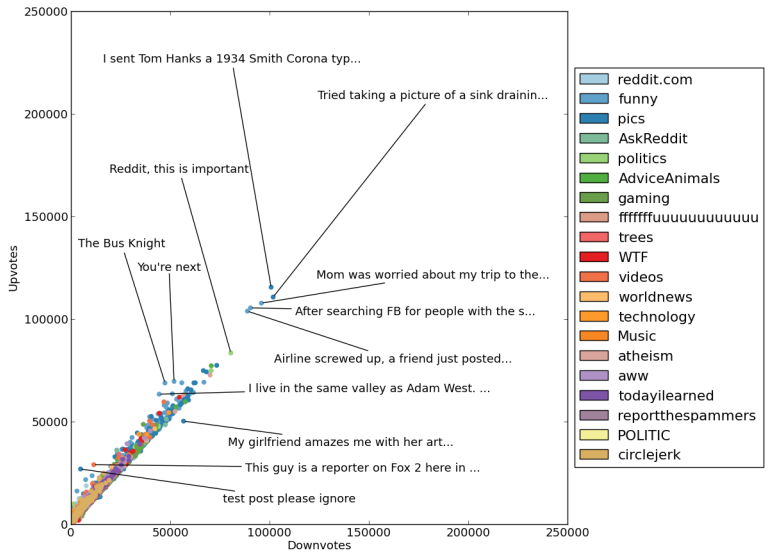


Figure 6.11.: Upvotes and downvotes for every submission in the top 20 subreddits are shown. Each submission is colored by the subreddit it was posted in. The highest rated submissions in the dataset and one fraudulent post are labelled with their title.

6.2.5. General Sentiment towards Submissions

Each and every upvote or downvote a submission receives, represents the appreciation or aversion of a logged-in redditor towards the presented content. In section 5.5.3, it was already discussed that votes of a single submission are largely dependent on the content itself, something that can not be reliable judged from the available data. Instead, by combining all submissions and their votes in a single observation, the content of each and every submission itself becomes negligible and, on the other hand, the voting tendencies become more visible.

Such an investigation has been conducted in figure 6.11, where all submissions from the top 20 subreddits are illustrated in relation to their upvotes and downvotes. In figure 6.12, the remaining submissions are displayed,

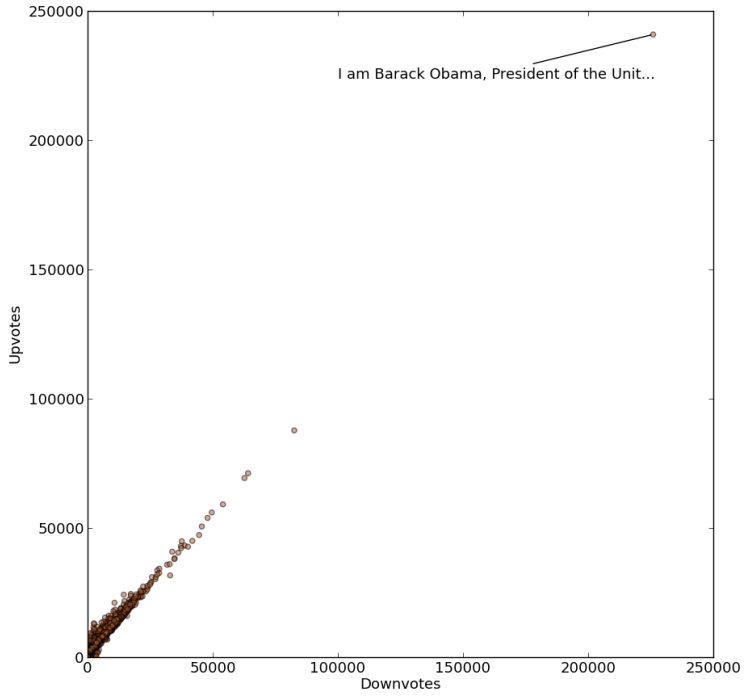


Figure 6.12.: Upvotes and downvotes for each submission in all but the top 20 subreddits are shown. The highest rated submission in the dataset is labelled with its title.

In both figures a clear pattern is emerging, submissions generally are upvoted roughly as many times as they are downvoted, with a slim tendency for slightly more upvotes. As small as this tendency is, with tens of thousands or even hundreds of thousands of votes only a fractional edge in upvotes already evokes a phenomenal score, as reddit's calculation of score is simply the difference between the positive and negative votes.

Just a handful of outliers break the observed constellation. while there are quite a few outliers above the pattern, all of them top rated submissions, there is one notable exception below. The post titled "My girlfriend amazes me with her art"[5] got downvoted heavily after it was found that the author had just submitted the work of another artist. The incident was so popular that it sparked the creation of a new meme, to this date redditors regularly submit pictures of world-famous paintings, claiming it to be artistic work of their girlfriends.

All of the other labelled submissions are submissions with the highest scores found in the dataset. The domination of *images* is also evident among these posts, but a leaning towards humorous and *funny* content is similarly conspicuous.

For the sake of completeness, figure 6.12 displays the remaining submissions, albeit lacking denotions for their associated subreddits. These submissions also follow the already observed pattern. The notable exception is the AMA by Obama (2012) that was already mentioned in section 1.1. It still is one of the most voted and highest scoring posts on reddit.

6.3. Assessing the Correlation of Attention Indicators

Following the voting and commenting tendencies observed in section 6.2.4 and section 6.2.5, the Pearson and Spearman correlation was

5 `http://redd.it/14cypy`

calculated to shed further light on the relation between the attention indicators. For easier reading, "Pearson" and "Spearman" are used synonymously with "Pearson Product Moment Correlation" and "Spearman Rank Order Correlation" for the remainder of this chapter.

As already described in section 4.4 and section 5.6, Pearson and Spearman are two different approaches for assessing the correlation between two variables, which are assumed independent. Pearson is able to discover *linear* relationships, while Spearman can also detect *non-linear monotonic* relationships. Spearman's calculation is additionally based on ranks instead of raw values, this way, the possibility that outliers and disparities in the data skew the results is mitigated.

Table 6.1.: The Pearson correlation coefficient and significance test p for the distributions of score, comments, upvotes, downvotes, total votes and title length to each other over all submissions in the dataset is found in table 6.2a, the results for the Spearman correlation are found in table 6.2b accordingly.

(a)Pearson

All - Pearson	**coef.**	p
Score to Comments	0.36	0.00
Comments to Votes	0.35	0.00
Score to Votes	0.76	0.00
Score to Ups	0.80	0.00
Score to Downs	0.69	0.00
Score to Title	0.01	0.00
Comments to Title	0.03	0.00
Votes to Title	0.00	0.00

(b)Spearman

All - Spearman	**coef.**	p
Score to Comments	0.44	0.00
Comments to Votes	0.60	0.00
Score to Votes	0.52	0.00
Score to Ups	0.69	0.00
Score to Downs	0.21	0.00
Score to Title	0.04	0.00
Comments to Title	0.11	0.00
Votes to Title	0.02	0.00

Table 6.1 shows the results for both correlations over every submission in the dataset. Both correlation methods arrive at proportionally similar coefficients for each calculation. The results for these calculations, and the results of every other calculation proposed in section 5.6 for that matter, exhibit p values of 0.00 for *every* correlation, thus the null hypothesis can be rejected for every correlation experiment.

Nevertheless, a striking problem with Pearson for this data is apparent in table 6.2a. The correlation of score and upvotes is at 0.80, meaning that submissions with higher numbers of upvotes also have higher scores. This tendency was of course expected, as every received upvote increases the score of a submission. Likewise, any downvote decreases the score of a submission. Yet, a similarly positive correlation between score and downvotes is detected by Pearson.

This can most likely be traced back to a fact evident in figure 6.11 and already discussed in section 6.2.5. Due to the slightly positive voting sentiment of reddit's community and the implementation of the score calculation, submissions are more likely to get higher scores the more votes they have. Thus Person calculates such a high positive correlation even for score and downvotes, since it only relies on the raw values for each submission. Still, the nature of this result is counter-intuitive and the problem behind it could potentially carry over to other results, where it is probably not that easily recognized.

Spearman, although still denoting a positive correlation between score and downvotes, shows significantly better results in this regard. By using ranks, the distinction between upvotes and downvotes and their correlation to score is more decisively and offers a more logical interpretation. For this reason, the following experiments are just discussed considering the results obtained from Spearman. Nonetheless, all results for Pearson and Spearman that are not included here can be found in appendix B.

Back to table 6.2b, the Spearman coefficient shows that score and upvotes, as expected, exhibit a strong positive correlation, the correlation between score and downvotes is much weaker. The dependence between upvotes, downvotes and score is well defined, which makes interpretation of these results rather easy.

Score and comments also share a positive correlation. This could mean that higher scores cause more comments. Plausible explanations for this would be that either high scoring submissions are more likely to offer discussion-worthy content or that the additional

exposure from high ratings causes more people to comment on a submission. Alternatively, one could also discern that only submissions that are discussed more lively are able to get higher scores. Looking at the even stronger correlation between comments and votes, it can be argued that voting and commenting is often happening together, although this cannot be proven, because unlike with comments, the users who cast votes can not be traced.

The strong correlation between score and votes confirms the earlier reasoning that more votes increase the chance of a submission to receive a higher score. It is also noteworthy that the title length seems to share nearly no correlation with the attention indicators, yet the correlation between comments and titles significantly higher. A longer and potentially more descriptive title could be more inviting to start a discussion.

Since this experiment encompassed all submissions from the dataset, many different types of contents and intentions behind the submissions are mixed here. To facilitate interpretation of the results, these types are now split up and noteworthy differences to the previous results are displayed and further explained. Again, complete results can be found in appendix B.

Table 6.3.: The Spearman rank correlation coefficient and significance test p for the distributions of score, comments, upvotes, downvotes and votes to each other over *self* submissions.

Self - Spearman	**coef.**	p
Score to Comments	0.46	0.00
Comments to Votes	0.66	0.00
Score to Ups	0.72	0.00
Score to Downs	0.09	0.00

The correlation coefficients for *self* submissions in table 6.3 show that score, comments and votes share similar correlations as when considering all submissions. The relation between downvotes and score is much lesser pronounced though, suggesting that negative votes are more carefully cast on self posts. Additionally this reveals that high scores for self posts are achieved rather by having proportionally

much more upvotes than downvotes, as opposed to having manis y votes with a proportionally modest edge towards upvotes.

Table 6.4.: The Spearman rank correlation coefficient and significance test p for the distribution of score to comments, votes and downvotes over submissions in *r/AskReddit*.

r/AskReddit - Spearman	**coef.**	p
Score to Comments	0.34	0.00
Score to Votes	0.20	0.00
Score to Downs	-0.13	0.00

The results for the submissions in *r/AskReddit*, exemplary for a self based subreddit, are provided in table 6.4. Downvotes actually show a negative correlation to score, something that was expected all along. This emphasizes the idea that votes on self posts are decided more carefully and not used as inflationary as on other submissions, explaining why score and votes are not as strongly correlated as previously known. Surprisingly the amount of comments also has less influence on the score, or vice versa.

Table 6.5.: The Spearman rank correlation coefficient and significance test p for the distributions of score, comments and votes to title length over *text* submissions.

Text - Spearman	**coef.**	p
Score to Title	0.10	0.00
Comments to Title	0.21	0.00
Votes to Title	0.19	0.00

For text submissions, which are often news or politics related, a longer title seems important to induce attention, probably because a longer title can potentially transport more information. The positive correlations between title length and score and especially votes and comments found in table 6.5 are much more pronounced than in other types of content.

The Spearman coefficients in table 6.6 expressly underline the previous observations regarding title length in text submissions, as they

were calculated for the subreddit *r/worldnews*, a subreddit primarily consisting of news related *text* submissions.

Image submissions are a prime example for inflationary voting behavior. The correlation of score and the number of votes in table 6.7 is significantly stronger than for other content types. Concurrently, the relation of score and downvotes is also stronger than usual, a definite sign that these submissions gain their high scores by experiencing much more attention through votes than other types of content. Interestingly, the correlation between title length and votes is even slightly negative. This could mean that redditors looking at image posts are primarily interested in the image itself, titles however should rather be short and precise and submissions with longer titles are possibly even skipped.

The subreddit *r/funny* is generally filled with *image* submissions, but anything considered *funny* is allowed to be posted there. So it comes as a little surprise that score, votes and comments are not similarly strongly correlated as was the case with image posts in general. Additionally, the correlation of score and downvotes comparably weak as for self posts. A possible explanation is, again, that votes are cast more carefully in this subreddit. On the other hand, most of the content in the subreddit could be equally appreciated by all voters, thus making the total number of votes less important for reaching high scores. Shorter titles are again inducing marginally more voting interaction than longer ones.

Table 6.6.: The Spearman rank correlation coefficient and significance test p for the distributions of score, comments and votes to title length over submissions in *r/worldnews*.

r/worldnews - Spearman	**coef.**	p
Score to Title	0.19	0.00
Comments to Title	0.18	0.00
Votes to Title	0.31	0.00

6.4. Analyzing the Development over Time

This section looks the evolution of the attention indicators over a timespan of 60 months, from 2008 to 2012. In each month, statistics for each indicator have been collected in regarding their distribution over subreddits, domains and type of content. Additionally, the entropy of these distributions has been calculated to evaluate the equality or inequality of attention allocation over all submissions, over all active subreddits and over all posted domains.

6.4.1. Subreddits

Figure 6.13 was constructed by calculating the share of, for example, the score accumulated in *r/funny* in January 2010 in comparison to all score that was amassed on reddit in this month. The height of the stack of any subreddit reflects its share of the total score at each month.

The proportion of score and comments allocated in the top 20 subreddits slowly decreases over the years. The shutdown of one of the oldest subreddits, *r/reddit.com*, at the end of 2011[6] only accelerates the fragmentation of attention to other subreddits. The content that was previously posted in subreddit open to any kind of submissions has then presumably spread to more specific and topically diverse subreddits.

Table 6.7.: The Spearman rank correlation coefficient and significance test p for the distributions of score, votes, downvotes and title-lenghts to each other over *image* submissions.

Image - Spearman	**coef.**	p
Score to Votes	0.67	0.00
Score to Downs	0.36	0.00
Votes to Title	-0.04	0.00

6 `http://redditblog.com/2011/10/saying-goodbye-to-old-friend-and.html`

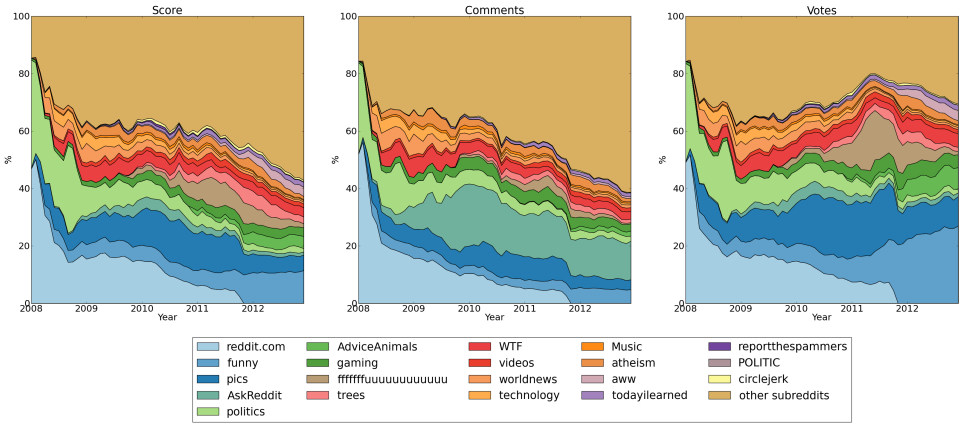


Figure 6.13.: The evolution of attention is shown for the top 20 subreddits individually, the remaining subreddits are combined into one variable. The score, number of comments and number of votes are shown as a relative proportion of the overall accumulation of the respective attention indicators at each month.

In figure 6.2 it was already shown that the number of active subreddits has grown excessively. Now it also clear that both the appreciation (score) and the discussion (comments) has simultaneously spread to this new subreddits.

Only the voting activity in the top 20 subreddits keeps roughly the same share over the whole timespan, especially through booming subreddits like *r/fffffffuuuuuuuuuuuu* in 2011, or the ever growing *r/funny*. Nevertheless, following the shutdown of *r/reddit.com*, a slight

Table 6.8.: The Spearman rank correlation coefficient and significance test p for the distributions of score, comments and votes to title length over submissions in *r/funny*.

r/funny - Spearman	**coef.**	p
Score to Comments	0.34	0.00
Score to Votes	0.34	0.00
Score to Downs	0.08	0.00
Votes to Title	-0.04	0.00

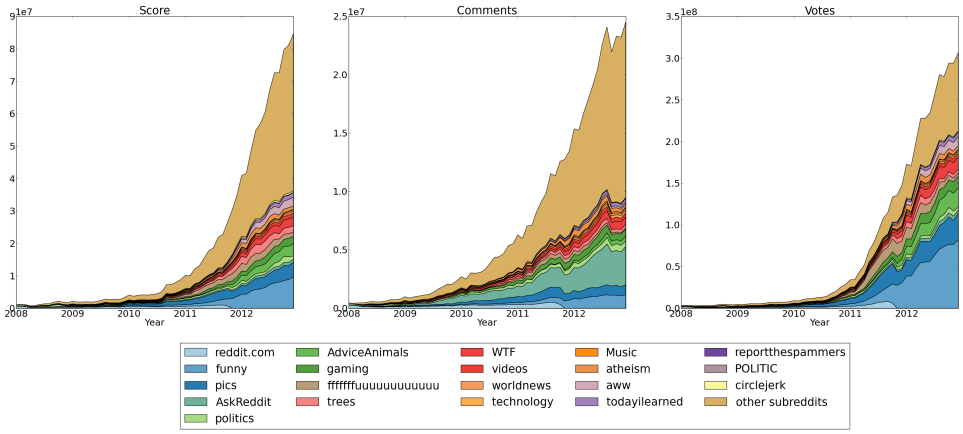


Figure 6.14.: The evolution of attention is shown for the top 20 subreddits individually, the remaining subreddits are combined into one variable. The score, number of comments and number of votes are shown as total values accumulated at each month.

rise in the share of all other subreddits is visible.

Figure 6.14 simply displays the total amount of, for example, votes that were cast in each month in each subreddit. The development displayed there further emphasizes the observation of a diversification of attention. Particularly the positive appreciation through score and discussion through comments fragment away from the top 20 into a rising number of other subreddits.

6.4.2. Domains

While attention was steadily draining from the top 20 subreddits to other subreddits, domains show an entirely reversed development regarding attention allocation, as can be inspected in figure 6.15.

At the beginning of 2008 the top 100 domains accumulated about 40% of all score, comments and votes on reddit, with the top 20 domains occupying roughly half that attention. At the end of 2012 the top 100 domains capture closer to 90% of all attention.

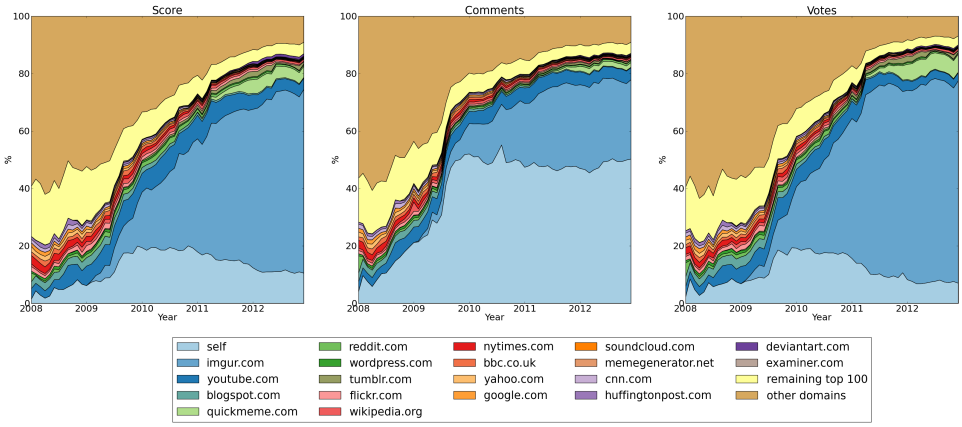


Figure 6.15.: The evolution of attention is shown for the top 20 domains individually, the remaining domains within the top 100 used for content categorization and all other domains are combined into two separate variables. The score, number of comments and number of votes are shown as a relative proportion of the overall accumulation of the respective attention indicators at each month.

Even more surprising, just one domain, *imgur.com*, manages to accumulate over 50% of all score and close to 60% of all votes at this point in time. Self posts can contend this domination by being responsible for about 50% of all comments written on reddit since late 2009.

Figure 6.16 puts this expansion of just a few domains into perspective to the overall increase of comments, votes and resulting score on reddit. Furthermore this illustrates that, parallel to reddit's users growth and diversification into more and more subreddits, the same users focused on submissions pointing to less and less domains.

Drastically speaking, around four fifth of the redditor's discussions, positive sentiment and general voting interaction is revolving around self posts and submissions pointing to an image hosting service that was specifically created to give redditors a convenient way to host their own pictures.

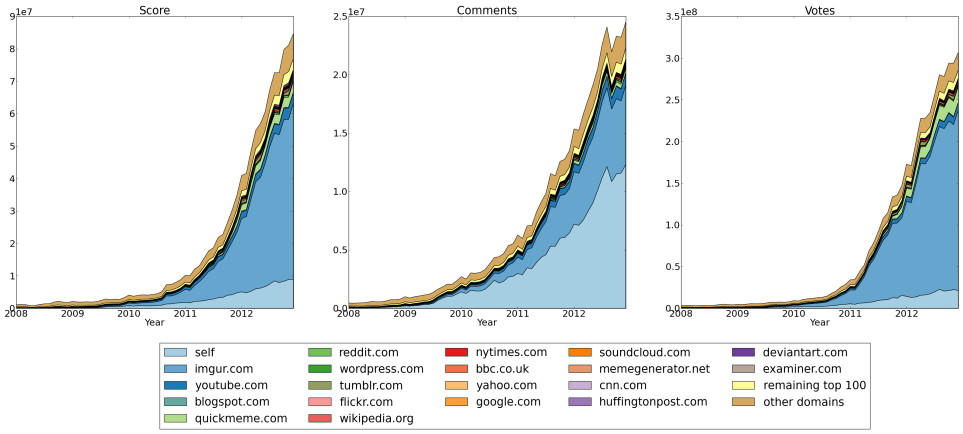


Figure 6.16.: The evolution of attention is shown for the top 20 domains individually, the remaining domains within the top 100 used for content categorization and all other domains are combined into two separate variables. The score, number of comments and number of votes are shown as total values accumulated at each month.

6.4.3. Type of Content

Figure 6.17 depicts the development of attention allocation per type of content. From this relative viewpoint the shares of *audio* and *misc* submissions are barely visible. *Video* submissions ability to generate attention in any form diminishes somewhat over the years.

The content type *text* undergoes more drastic changes. Starting off as the primarily voted and commented category on reddit, it is hardly able to hold on to as much comments, votes and subsequently score as *video* submissions in the end. So despite a growing diversification of topics and the appreciation of this diversification, observed in section 6.4.1, the attention allocation per type of content is rather lopsided.

Image submissions exhibit an enormous growth in attention received and are responsible for about 85% of all votes and a similarly high fraction of all score at the end of 2012. They also generate nearly one third of all comments on reddit since mid 2011. Contrastingly,

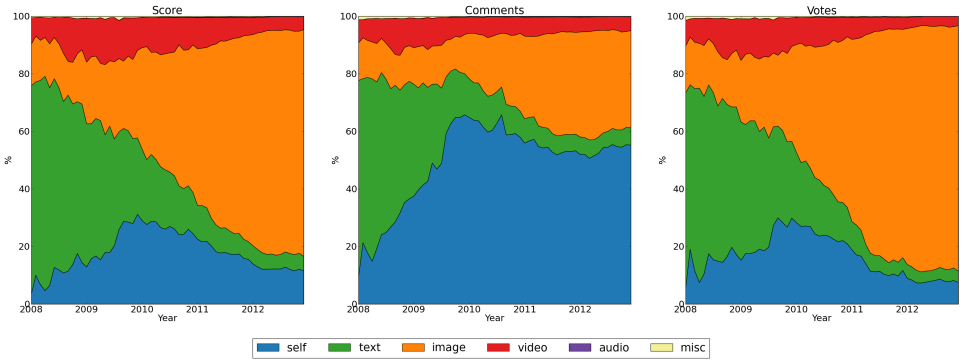


Figure 6.17.: The evolution of attention is shown for type of content individually individually. The score, number of comments and number of votes are shown as a relative proportion of the overall accumulation of the respective attention indicators at each month.

self posts get voted less and less but are still the major source of comments on the platform.

Figure 6.18 puts these drastic development of appreciation towards image submissions into context to the growth of attention generated by each content type. While user numbers and submissions were growing, *image* submissions seem to have profited the most, their overall scores and votes rise nearly perfectly analogous to the growth of reddit. This observations give way to often voiced concerns that, by becoming more mainstream, reddit has quickly turned from the "front-page of the internet" into an image-board.

Moreover, looking back at figure 6.3 tells that the numbers of submissions in all types of content have experienced gigantic growth, both *image* submissions and *self* just substantially more so than the others. Therefore, to examine if the perception of different types of content has actually changed on reddit regardless of the number of submissions, figure 6.19 illustrates the development of the average score, average number of comments and average number of votes a submission in one of the content categories receives.

This investigation reveals that with the exception of *image*, all types of

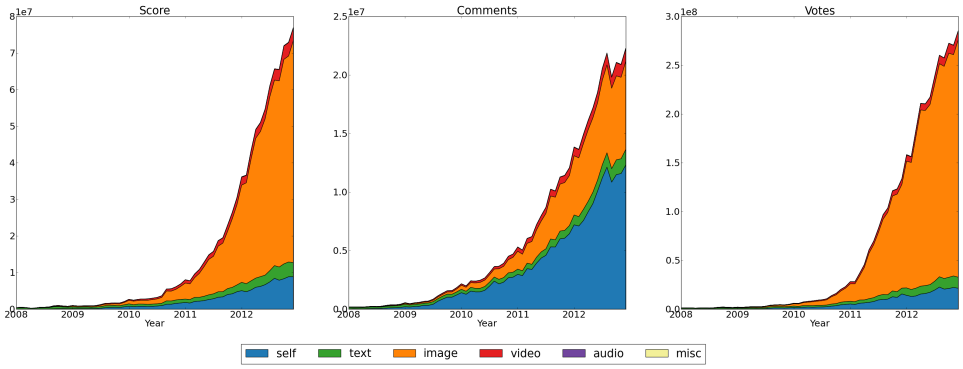


Figure 6.18.: The evolution of attention is shown for type of content individually individually. The score, number of comments and number of votes are shown as the total values accumulated at each month.

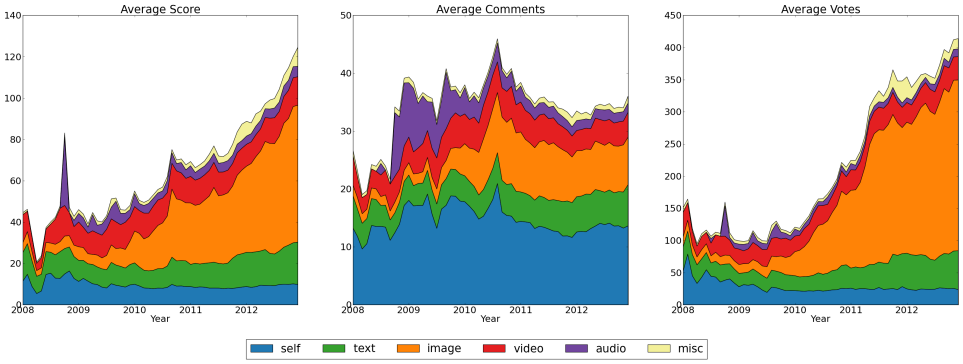


Figure 6.19.: The evolution of attention is shown for type of content individually individually. The score, number of comments and number of votes are shown as the average values a submission in the respective category received at each month.

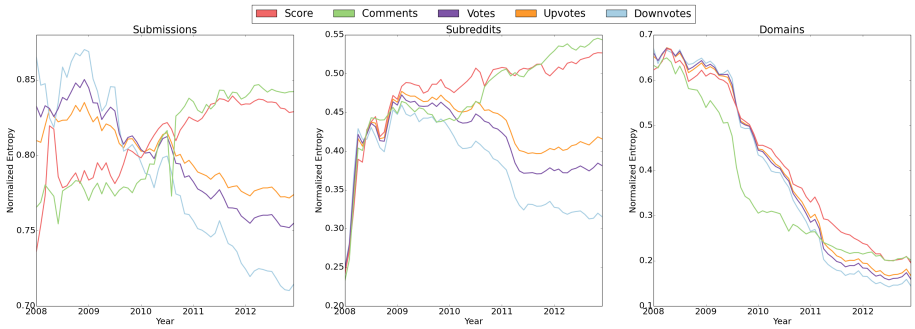


Figure 6.20.: Development of normalized entropy of the distributions of all attention indicators over time. At each month the normalized Shannon entropies for the distributions of score, number of comments, votes, upvotes and downvotes during this month are calculated. The figures depict the distribution of said attention indicators over all submissions, all subreddits and all domains from left to right.

content largely maintained the same power to draw specific attention. An average *image* submission, however, receives nearly 35 times as much votes in December 2012 than it did in 2008. Consequently, the average score also got drastically higher during this time. Likewise, the average amounts of comments has grown for *image* submissions, albeit not the same extent, suggesting that different types of content are also more affine to specific types of attention.

Notable is also the spike in average score, comments and votes for *music* submissions in late 2008, following the sudden popularity of the online audio sharing platform *Soundcloud*[7]. Average comments for self submissions spiking in late 2010 can be solely attributed to a post breaking the news of the marriage of two reddit founding members[8], nearly 360.000 redditors commented to congratulate.

6.4.4. Entropy of Attention Indicators

In figure 6.20 the evolution of equality or inequality of attention distribution over all submissions, all subreddits and all domains in each month can be inspected. Higher entropy values signify a more equal dissemination of, for example, all votes over all submissions, a lower value henceforth represents a more unequal allocation, where few submissions receive most of the votes. A normalized entropy value of 1 indicates perfect equality, every submission received exactly the same amount of votes, an normalized entropy value of 0 the opposite, perfect inequality, a single submissions received all votes.

For submissions, the spread of attention over all of them is rather balanced, all entropy values reside in a very high range. Nevertheless, the enormous growth of reddit changed their allocation behavior in an interesting way. The equality of score and comment distribution over all submissions has become more even over the time, meaning the positive appreciation towards submissions and the willingness to discuss them stretches over nearly everything submitted to reddit. On the other hand, upvotes, downvotes and general voting behavior has become slightly more focussed on less submissions over time. This could mean that a large part of all voting is done on fewer but more popular submissions. Looking closely, a stark downward spike in the otherwise rising entropy of comments can be seen, this is, again, caused by the marriage post mentioned earlier.

Considering subreddits, a similar development is revealed. In early 2008 the addition of user created subreddits[9] caused attention to spread from the few default subreddits to all of those newly created subreddits. More subreddits were created over the years and their content was perceived positively and discussed lively, evident by the rising normalized entropy of score and comments. This confirms the observations that were made in section 6.4.1, a thriving diversification

7 http://soundcloud.com/
8 http://redd.it/d14xg
9 http://redditblog.com/2008/03/make-your-own-reddit.html

of positive perception and discussions into the growing number of subreddits. On the contrary, most of the voting, including upvotes and downvotes, was concentrated on fewer subreddits and stays this way, as indicated by the decreasing entropy values.

While the dissemination of attention over all domains started out fairly balanced, the dominance of the two major players *imgur.com* and *self* witnessed in section 6.4.2 is also unmistakeable visible here. The steady decrease of the entropy of all attention indicators signals an utter inequality of their allocation, most of all voting, appreciation and discussion is directed at submissions of just a few selected domains.

6.5. On the Influence of Submission Time

In this section, it is studied what influence the time, at which a submission was posted, has on its perception and the attention it generates. The figures here represent this attention and perception based on all submissions between 2008 and 2012. More detailed results for the individual content types and subreddits can be found in appendix C.

Figure 6.22 depicts the total amount of score that was received, the total number of comments that were written and the total number of votes that were cast during an average week. For a more exhaustive examination, the number of submissions per day were added to figure, illustrated by the blue line. Note that these are not the actual submission numbers but a scaled representation.

Users tend to vote and comment much more during the workdays, nearing the weekend, on Friday, user attention towards content on reddit drops significantly until it reaches its lowest point on Saturday and rises again on Sunday. This could mean that redditors prefer to keep up-to-date or use reddit for relaxation during workdays, but many stay away from the portal on the weekends. These attention

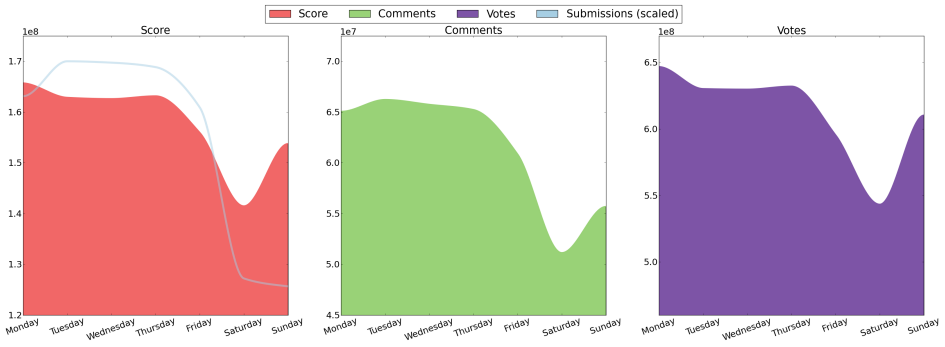


Figure 6.21.: The total amount of score, comments and votes for submissions posted on each weekday is shown. The blue line symbolizes the number of submissions for each day in a scaled manner, these are not the actual submission values.

levels are also rather analogous to the number of submissions, suggesting that the day of the week has influence on the general activity of the portal.

Since submission activity proportionally drops much lower than attention activity on weekends, the average score, commenting and voting activity per submission is higher on weekends, but slightly lower on workdays, as shown in figure 6.22.

A typical day on reddit is characterized by stark fluctuations in commenting and voting activity (and therefore in resulting score). Figure 6.23 shows the attention development on a hourly basis. Again, the amount of submissions per hour is added to the illustration. Considering time zones of the United States instead of UTC, the pattern becomes much easier to interpret.

Attention activity drops heavily around midnight and reaches its lowest point just a few hours later, which makes sense, as it is the middle of the night in these time zones. During the early morning hours the activity levels rise again drastically and peak somewhere around midday and fall again after that.

All in all, the attention levels on reddit clearly mirror an arbitrary day

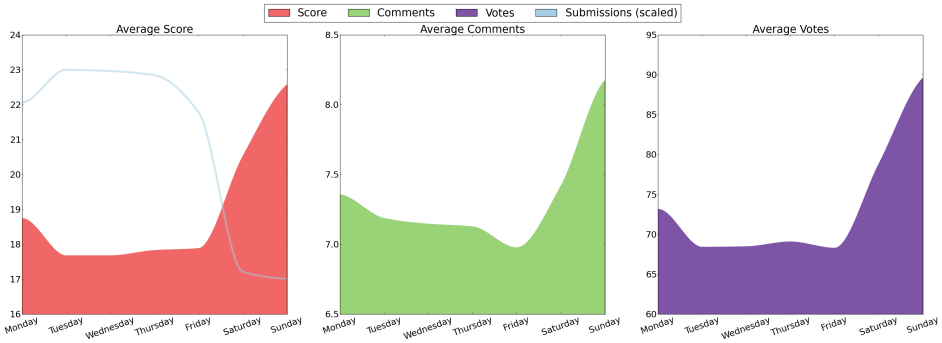


Figure 6.22.: The average score, average number of comments and average number of votes per submission for submissions posted on each weekday is shown. The blue line symbolizes the number of submissions for each day in a scaled manner, these are not the actual submission values.

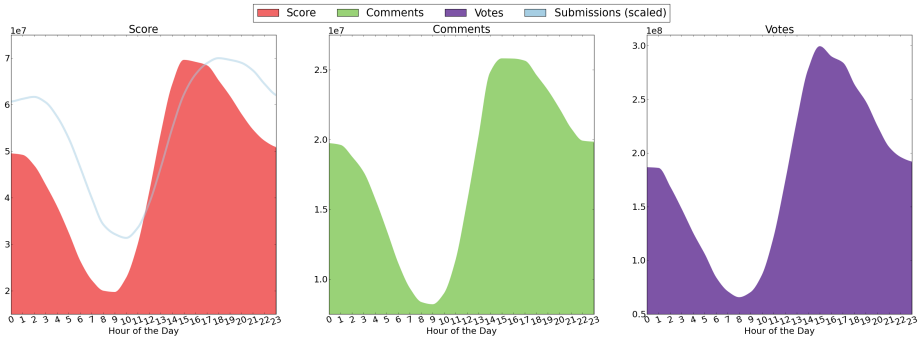
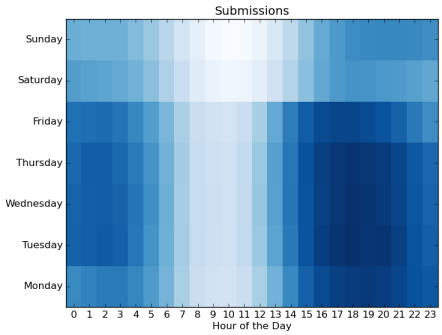
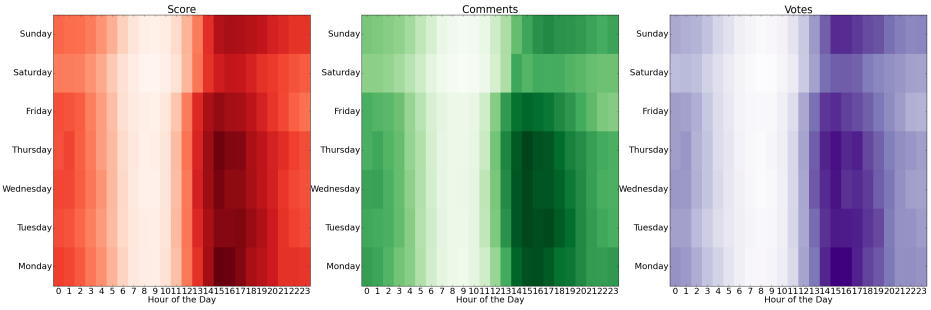


Figure 6.23.: The total amount of score, comments and votes for submissions posted on each hour of the day is shown. The blue line symbolizes the number of submissions for each hour in a scaled manner, these are not the actual submission values.

in Eastern or Pacific Daylight Time, implying that a large proportion of redditors actually live in these time zones (or at least manage their time like they would live there).



(a)Submissions



(b)Attention

Figure 6.24.: Heatmaps combining the weekly and hourly data are shown. Each box symbolizes one hour on a specific day. The darker the box, the higher the value at this hour and day, subsequently lower values are signified by lighter boxes. Figure 6.24a displays the number of submissions posted each hour and day. Figure 6.24b illustrates the total score, number of comments and number of votes received by submissions posted at each hour and day.

Figure6.24 combines the daily and weekly activity levels into heatmaps denoting both the number of posted submissions and the received attention at each hour of each weekday.

6.6. Performance of Classification of Submission Titles

In table 6.10 the precision, recall and resulting f1-score of the binary naive Bayes classification experiment can be found. Comparing these results to the ones from the multinomial naive Bayes classification experiment in table 6.11 reveals that both classifiers performed fairly even in assigning submission titles to their respective subreddits.

Surprisingly, the much simpler approach of the BNB classifier performs slightly better overall with an average f1-score of 0.51 compared to the 0.48 of the MNB classifier. The MNB classifier shows marginally better precision values but lacks the superior recall performance of the BNB classifier.

For comparison, the performance of the baseline models of the classifiers, constructed as described in section 5.9, is presented in table 6.9.

Table 6.9.: The performance of the BNB and the MNB classifiers in comparison to their average baseline (randomly shuffled) over 100 iterations are shown via f1-score.

	optimal	**baseline**
BNB	0.51	0.054
MNB	0.48	0.061

Inspecting the values at class-level additionally shows that the classification performance varies wildly per subreddit. Topically more specific subreddits, like *r/tf2trade*, a subreddit acting as a marketplace for the trading of in-game items of the F2P computer game *Team Fortress 2*[10], and *r/leagueoflegends*, a subreddit concerned with the F2P game *League of Legends* (already briefly mentioned in section 6.2.4), exhibit astonishing f1-scores of 0.95 and 0.82 respectively.

Subreddits covering broader topics perform significantly worse. For example, *r/funny*, with f1-scores around 0.40 in both classifiers, ac-

10 http://teamfortress.com/

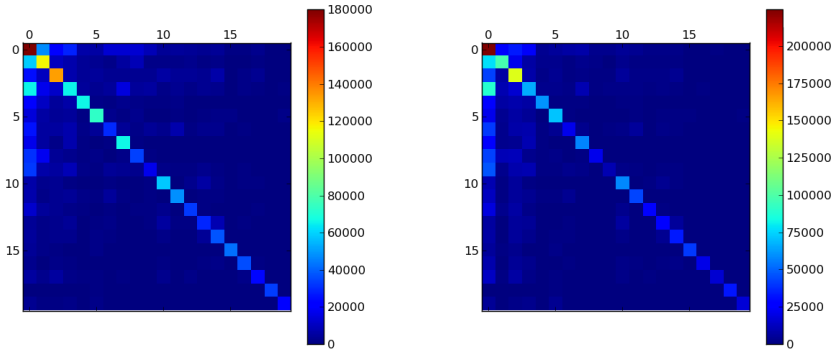
cepts all kind of submissions whose content is deemed funny (most of them are *images*). Similarly, *r/pics*, having even lower f1-scores, allows pictures and photographs of any kind, only limited by a rather forgiving rule-set. Likewise, the f1-score for submissions to *r/videos* is well below the average.

Table 6.10.: Precision, recall and f1-score for the binary naive Bayes classifier on all submissions of the 20 most active subreddits in 2012.

	precision	**recall**	**f1-score**
funny	0.34	0.50	0.41
AdviceAnimals	0.45	0.51	0.48
AskReddit	0.59	0.59	0.59
pics	0.40	0.31	0.35
trees	0.68	0.54	0.60
gaming	0.67	0.61	0.64
videos	0.41	0.28	0.33
aww	0.55	0.63	0.59
ffffffuuuuuuuuuuuu	0.44	0.35	0.39
WTF	0.36	0.19	0.25
politics	0.65	0.67	0.66
Music	0.62	0.62	0.62
atheism	0.60	0.50	0.54
worldnews	0.49	0.47	0.48
technology	0.52	0.61	0.57
leagueoflegends	0.86	0.79	0.82
todayilearned	0.73	0.73	0.73
circlejerk	0.64	0.49	0.55
tf2trade	0.95	0.95	0.95
Minecraft	0.76	0.63	0.69
avg / total	0.52	0.51	0.51

Table 6.11.: Precision, recall and f1-score for the multinomial naive Bayes classifier on all submissions of the 20 most active subreddits in 2012.

	precision	**recall**	**f1-score**
funny	0.31	0.62	0.41
AdviceAnimals	0.53	0.43	0.47
AskReddit	0.45	0.61	0.52
pics	0.38	0.31	0.34
trees	0.75	0.50	0.60
gaming	0.66	0.59	0.62
videos	0.43	0.20	0.28
aww	0.62	0.54	0.58
ffffffuuuuuuuuuuuu	0.68	0.22	0.33
WTF	0.45	0.11	0.17
politics	0.60	0.67	0.63
Music	0.64	0.56	0.60
atheism	0.65	0.40	0.50
worldnews	0.51	0.43	0.47
technology	0.57	0.55	0.56
leagueoflegends	0.90	0.75	0.81
todayilearned	0.73	0.41	0.52
circlejerk	0.73	0.34	0.46
tf2trade	0.95	0.93	0.94
Minecraft	0.84	0.51	0.63
avg / total	0.53	0.48	0.48



(a)Binary naive Bayes (b)Multinomial naive Bayes

Figure 6.25.: The confusion matrices for the binary naive Bayes classification experiment (figure 6.25a) and the multinomial naive Bayes classification experiment (figure 6.25b) are shown. The classes are listed from 0 to 19 according to the ordering of table 6.10 from top to bottom. The Y-axis portrays the actual classes and the X-axis the classes assigned by the classifier. The color of each matrix field depicts the number of assignments. The diagonal of the matrix represents the correctly assigned samples.

In figure 6.25 the confusion matrices for both classifiers are displayed. According to these values, submissions from *r/funny* are more often misclassified than others, they are actually mostly mixed up with other *image*-based subreddits.

Generally it seems that the titles of more topically focussed subreddits are also differentiated better. This suggests that the communities discussing more specific topics have also more strongly developed their own customary idioms and expressions, heavily influencing the style of their submission titles. In some communities, especially the aforementioned computer game related ones, expressions may have been indoctrinated by the terminology used in the games themselves. On the other hand, subreddits like *r/AskReddit* and *r/todayilearned*, a place, where redditors post interesting facts or stories they just "learned today", have no such external influencers. Still, their classification performance is way above the average.

So, while it is not always clear how or why these customary idioms or separate terminologies have formed, it can atleast be observed that most of the top 20 subreddits indeed have them to a degree, where a differentiation is more or less reliably possible.

The associated terms of the largest coefficients of both classifiers for each of the top20 subreddits are listed in appendix D.

6.7. Performing Trend Discovery and Analysis

Exemplary results for trend discovery via a modified variant of TFIDF are displayed in section 6.7.1. Likewise, selected results from the analysis via classification coefficients are found in section 6.7.2. As these experiments were performed over twelve months and 20 subreddits, presentation and discussion of all of them is not feasible within the scope of this thesis, however, some of the raw results are attached in appendix E.

6.7.1. Discovery by Means of Modified TF-IDF

As described in section 5.10.1, two methods were chosen to discover terms that are trending in a specific month. In table 6.13, the results of the TFIDF modification just considering the prior month are presented, this approach will be related to as *monthly* from now on. Table 6.12 shows the trending terms obtained by considering the whole year for the calculation, this approach will be referred to as *overall*. The trending terms illustrated here were calculated over all submissions, regardless of subreddits.

Both approaches discover "cinsere" to be the most trending term in January 2012. This is related to reddit-internal scandal, involving the

alleged embezzlement of many redditors donations by the moderator of the subreddit *r/trees*[11].

Interestingly, only the monthly approach considers the term "concordia", relating the cruiser "Costa Concordia", which shipwrecked in January 2012, a trending term. This is probably ignored by the overall variant, because news about the incident continued for several months, thus a significant enough trend just for January was not detected.

On the other hand, only the overall method identifies "pipa" (PROTECT Intellectual Property Act) and "sopa" (Stop Online Piracy Act) as important terms in January. Although protests against this bills were heavily supported by redditors, leading to a "blackout" of reddit.com together with other popular internet services[12], both terms do not appear in the monthly approach. This can most likely be attributed to the fact that these topics were already discussed on reddit for months and therefore not considered a new trend by the monthly variant.

Looking at the trending terms for October, both approaches deliver the same results, although ranked slightly different. This is not surprising, as all of the related events just happened within that month and had not that much repercussion prior or afterwards. They are as follows:

Table 6.12.: Overall trending words for submissions in the top 20 subreddits of 2012, calculated at each month in comparison to the whole year. Only submissions from 2012 are valid for this experiment.

January	cinsere, pipa, sopa, blackout, ...
...	...
October	baumgartner, binders, stratos, frankenstorm, felix, ..., violentacrez, ...

11 http://redd.it/ojeom

12 http://redditblog.com/2012/01/stopped-they-must-be-on-this-all.html

- the supersonic freefall of "Felix" "Baumgartner" during project "Stratos"[13]
- hurricane "Sandy" merging with a bad weather front forming "Frankenstorm"[14]
- an incident with Mitt Rommney using the phrase "binders full of women"[15], spawning much controversy and even its own dedicated meme
- the revelation of the identity of one of reddit's most notorious trolls, "violentacrez"[16]

Even after this brief inspection of the results, it is obvious that both approaches have their own drawbacks in detecting trends, depending on the temporal development of the topic. Future work could try to optimize these methods, possible even by combining them with an elaborate weighting scheme.

6.7.2. Analysis by Means of Classification Coefficients

Figure 6.26 illustrates the trained classifier coefficients and their associated terms for each month, ranked by their importance for

Table 6.13.: Monthly trending words for submissions in the top 20 subreddits of 2012, calculated at each month in comparison to month before. Only submissions from 2012 and December 2011 are valid for this experiment.

January	cinsere, concordia, acta, ..., blackout, ...
...	...
October	sandy, binders, baumgartner, frankenstorm, ..., stratos, violentacrez,..., felix, ...

13 http://redbullstratos.com/
14 http://science.time.com/2012/10/29/frankenstorm-why-hurricane-sandy-will-be-historic
15 http://youtu.be/wfXgpem78kQ
16 http://gawker.com/5950981/unmasking-reddits-violentacrez-the-biggest-troll-on-the-web/all

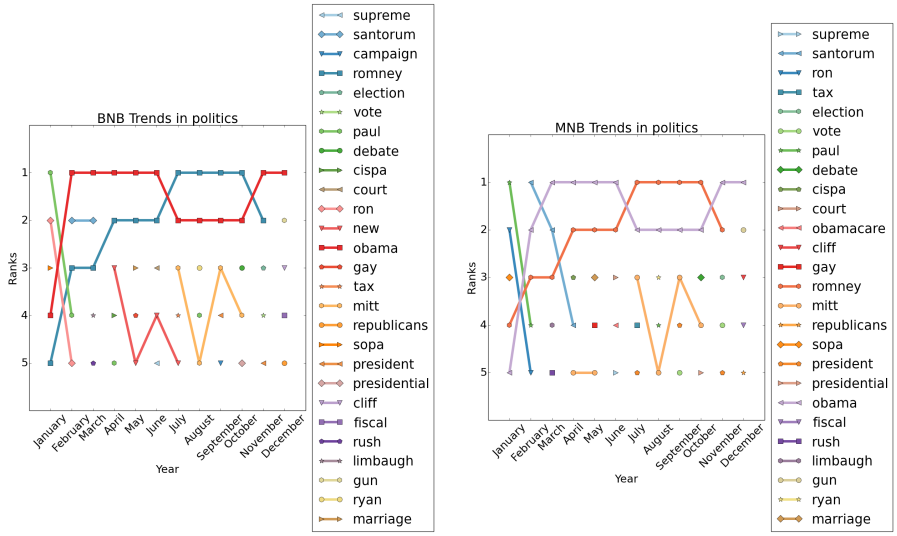
the respective subreddit, in this case *r/politics*. Terms that appear for consecutive months are visually connected to demonstrate their ongoing popularity in the subreddit.

While there are many new terms in the top five of this ranking in each month, some remain popular for longer time spans. The terms and their development appear to be rather similar for both classifiers.

Important events in the political landscape of the United States are clearly visible, for example, the discussion over the Stop Online Piracy Act (SOPA) in January or a related proposal, the Cyber Intelligence Sharing and Protection Act (CISPA) in April, as well as the presidential debate in October and the following election in November.

The year is dominated by the race for presidency between Mitt Romney and Barack Obama. The former dips in popularity in November after the debate and completely drops off the grid after losing the election against Obama, while the new president remains the most important topic for the last month of the observed year.

Being able to make this observations simply by visualizing classification coefficients and their development over time deems the proposed approach viable for trend analysis. Future work could possibly improve this technique by additionally considering the raw values of the coefficients as opposed to just their rankings.



(a)Binary naive Bayes (b)Multinomial naive Bayes

Figure 6.26.: The top ranked classifier coefficients trained each month in 2012 on *r/politics* are shown. Trending terms staying in the top 5 ranks over consecutive months are connected for a more convenient viewing experience. The results extracted from the binary naive Bayes classifier are displayed in figure 6.26a and the results from the multinomial naive Bayes classifier in figure 6.26b accordingly.

7. Discussion of Results

This chapter briefly summarizes the insights gained from the results obtained in chapter 6.

The indicators of attention can differ greatly for each observed dimension, suggesting that specific subreddits, domains or types of content encourage different interactions and behavioral patterns. The general sentiment of reddit's community, indicated by its voting behavior, is positive and appreciative.

Voting is the prevalent way of interacting with the portal. Although a logged-in user can only vote once per submission, but is able to comment the same submission multiple times, an average submission receives much more votes than comments.

Correlations between the attention indicators have been detected. Their intensity and peculiarity differs considerably per type of content and subreddit, reinforcing the assumption that specific topics or content categories elicit distinct patterns of user interaction and behavior.

The enormous growth in regard to posted submissions, creation of subreddits, usage of distinct domains and active users greatly influenced the evolution of attention allocation on reddit. The diversification of submissions into more and more subreddits was well received, as evident by the dispersion of appreciation and discussion.

Contrastingly, fewer and fewer domains, mainly *imgur.com* and *self*, accumulate most of the attention, challenging reddit's declaration to be the "front page of the internet". These observations are confirmed

by the examination of the development of attention entropy over time.

Image and *self* submissions are the major recipients of attention content-wise, further undermining reddit's self-proclaimed purpose, as users tend to appreciate self-created content and pictures more than other external sources. While more and more votes are cast, only the average number of votes and average score for *image* submissions has considerably grown, suggesting that new users primarily join reddit for this type of content.

The consumption of *image* submissions requires arguably less cognitive effort than the consumption of the textual *self* submissions. Conspicuously, *image* submissions attract the similarly effortless voting attention, while *self* posts provoke more comments, which obviously require more effort than simply clicking the upvote or downvote buttons.

Submission time has been identified as an external factor with substantial impact on the collective attention and additionally implies that reddit's community is most active during daytime in North American time zones.

The results of the classification experiment indicate the utilization of customary idioms and specific terminology in different subcommunities. This effect is more pronounced in more topically specific subreddits.

Trends can successfully be extracted from submission titles with various approaches, their visualization proves suitable for temporal analysis. However, all methods exhibit certain drawbacks, which need to be improved in future work.

8. Conclusion

To conclude this thesis, the research questions raised in section 1.2 are revisited and answered.

i. Attention indicators other than score could be leveraged for better understanding of user attention on reddit. Besides the positive appreciation, upvotes and downvotes allow a better differentiation of user perception. The overall votes of a submission were identified as a simple engagement metric, similarly, the number of comments were utilized to assess the liveliness of discussions.
ii. A detailed characterization of these indicators is provided, revealing stark discrepancies in their utilization and bias toward specific subreddits, domains and types of content. Considerable correlations between these attention indicators are uncovered. The shape of these relationships is dissimilar for distinct types of content and subreddits.
iii. The evolution of attention indicators and users' perception is investigated over several years and reveals an interesting development of appreciation of the emerging topically diversity contrasting a growing focus on fewer types of content.
iv. Consequently, only some of the observations and their implications are globally applicable and are more often limited to certain parts of reddit, attributed to a growing number of diverse subreddits and sub-communities with various expectations and interests. Furthermore, it was uncovered that the characteristics and generation of attention can differ greatly among these sub-communities.

Additionally, implicit and explicit factors, influencing attention on reddit, have been identified in submission titles and submission time. The former induces certain types of attention and acts as a distinguishing feature between subreddits. The latter greatly influences how much attention is received. Submission titles have been successfully exploited for trend discovery and analysis, however, further improvement is needed for the constructed techniques.

8.1. Limitations and Threats to Validity

Several points are to note, generally potentially threatening the validity of this thesis. First, he choice of statistical and scientific methods utilized and their application to arrive at the results presented here. The methods and constructed experiments are well suited to provide meaningful insights and answers to the posed problems. Nevertheless, better approaches providing superior performance may exist. All applications of such methods were tested with multiple variants and implementations, ultimately, the best performing models were chosen.

Additionally, the process of content categorization is based on a manual classification. This bears a potential threat to the validity of content-related analysis in this work. However, multiple people were partook in this process and several adaptations of the categorization scheme upon their feedback was performed to mitigate this threat.

Furthermore, some internal and external factors could also have skewed the results. Technical changes to reddit and utilization of browser-plugins that transform the user interface are largely disregarded, but could have influenced the results to a greater extent than expected. Other threats are posed by bots and alleged voting manipulation on reddit, both are not discussed in this thesis.

Regarding the conclusions discussed in section 8, the following specific threats to their validity arise:

i. The identified indicators of attention are largely dictated by the information available in the dataset, a finite snapshot of publicly available data, and therefore by no means constitute a complete view on attention on reddit. Other indicators could possibly be identified by studying click-data or the chronological development of voting and comment threads, both of which are suggested in section 8.2 for subsequent work. Click data is not publicly available so far and the collection of comment threads proved difficult to accomplish to a satisfying extent within the scope of this thesis.
ii. The uncovered characteristics of attention indicators are analyzed in more detail for just a few selected subreddits and types of content. A more fine-grained approach could reveal even more details about their nature, but was left to future work. The correlation analysis was conducted with two different approaches to expose skewing of results due to outliers or effects imposed by the peculiarity of the attention distributions. Furthermore, statistical significance testing was performed for these results to refute any claims that they were a mere product of happenstance.
iii. Events in the development of attention indicators over time were described to the best knowledge of the author, but there is no definite proof that the observed events were caused by or imply the proposed interpretations. Therefore, to further confirm these conclusions, the evolution of the entropy of attention indicators was investigated and indicated a strikingly similar development.
iv. Again, only a selected subset of all sub-communities and otherwise distinguishing factors were considered for the experiments, covering a greater subset or all of reddit could potentially lead to different results. This effect was reduced by choosing the most active subreddits, which comprise a large part of reddit. The classification experiment is subject to some general threats imposed on supervised learning methods, such as overfitting, which was reduced by applying 10-fold cross validation.

8.2. Future Work

This thesis allows an extensive look at attention patterns on reddit. The results presented here offer a foundation for more detailed analysis of attention behavior and its influencing factors, both externally and internally. As the examinations provide an exhaustive overview of the most popular and largest parts of the platform, many of the experiments could be extended to capture an even greater scope.

Additionally, most of the techniques and approaches utilized here are easily applicable to similar social news and content aggregators, like *Hacker News*[1], *Digg* or even *Imgur*'s social platform, as all of them implement similar interaction and ranking structures. It would be interesting to validate if the observations made in this thesis also apply to other communities and can be valued as general observations for such collaborative filtering platforms or if they are exclusive to reddit.

The trend discovery and analysis presented in this thesis, albeit providing valuable results, need to be further improved. Combinations of the applied methods could potentially limit their specific disadvantages, future work could investigate possible optimizations of these approaches.

The observations of this work are based on a data set of finite snapshots of submissions, subsequent work could be founded on temporal data. To be able to analyze the chronological development of attention on submission level would provide new and interesting insights, especially regarding the correlation of attention indicators, or even reveal new attention indicators.

Apart from considering their number per submission, comments themselves were largely disregarded in this thesis, due to the technical constraints posed by reddit's API. Nevertheless, inspecting comment threads is a compelling idea. Comments can be interacted with the same way as submissions, they can be upvoted, downvoted, are ranked accordingly and users can comment them. They could

1 http://news.ycombinator.com/

therefore potentially reveal much about the communities discussion behavior and sentiment.

Another interesting opportunity for research regarding attention and user behavior would be provided with the acquisition of click-data. A way to do this could perhaps involve the *Reddit Enhancement Suite* (RES)[2], a browser-plugin designed to improve the user experience on reddit, as there is a possibility that RES-users would consent to have their click-data anonymously collected through the plugin for scientific purposes.

The choice of the default subreddits as well as the specifications for karma generation certainly have an impact on the attention behavior of redditors and need to be assessed in future work.

Furthermore, it could prove valuable to expand the inclusion of reddit's community into the research process by tapping resources like the subreddit *r/theoryofreddit*, a sub-community focussed on reddit related research and meta-knowledge about the portal, or by conducting surveys similar to the ones that already proved fruitful in Singer et al. (2014).

2 http://redditenhancementsuite.com/

Appendix

Appendix A.

General Information

A.1. Sources for User Numbers

Table A.1 shows the sources used to retrieve the user numbers for the social networks discussed in section 3.1. While some of them could be found on the platforms themselves, others do not publish such information and the numbers are therefore based on press releases or recent reports in renowned media.

A.2. Examples for Domain Consolidations

Examples for domain consolidations are shown in table A.2. The consolidations include merging domains pointing to the same service with different top level domains, different country codes, different domains for content hosting and domains with the sole purpose of internal link shortening.

Table A.1.: List of recent reports used to arrive at the approximate user numbers for the various social networks mentioned in table 3.1 and table 3.2.

Network	**Source**
Facebook	http://newsroom.fb.com/company-info
LinkedIn	http://press.linkedin.com/about
Twitter	https://about.twitter.com/company
Digg	http://en.wikipedia.org/wiki/Digg#History
reddit	http://reddit.com/about
Orkut	http://onforb.es/r9SwhW
Sina Weibo	http://techinasia.com/ sina-weibo-60m-daily-active-users-q3-2013
Myspace	http://myspace.com/pressroom/pressreleases
4chan	http://4chan.org/advertise
Hi5	http://about.tagged.com/hi5
Slashdot	http://diceholdingsinc.com/phoenix.zhtml?c= 211152&p=irol-newsArticle&ID=1735911
Mendeley	http://blog.mendeley.com/start-up-life/ mendeley-has-2-5-million-users
Boards.ie	http://boards.ie/content/about-us

Table A.2.: Exemplary domain consolidations are shown.

Original Domain	**Consolidated Domain**
.blogspot.	blogspot.com
youtu.be	youtube.com
staticflickr.com	flickr.com
deviantart.net	deviantart.com
qkme.me	quickmeme.com

A.3. Domain Assignments for Type of Content

Table A.3 and table A.4 show the top 100 domains and their respective type of content that has been manually assigned. The reasoning behind certain assignments is described in more detail in various sections throughout the thesis.

Table A.3.: The top 40 domains and the manually assigned content category is shown. The domains are in descending order of number of submissions to the respective domain.

Domain	Type
self	self
imgur.com	image
youtube.com	video
blogspot.com	image
quickmeme.com	image
reddit.com	text
wordpress.com	text
tumblr.com	image
flickr.com	image
wikipedia.org	text
nytimes.com	text
bbc.co.uk	text
yahoo.com	text
google.com	text
soundcloud.com	audio
memegenerator.net	image
cnn.com	text
huffingtonpost.com	text
deviantart.com	image
examiner.com	text

Domain	Type
facebook.com	text
guardian.co.uk	text
wp.me	misc
vimeo.com	video
twitter.com	text
go.com	text
washingtonpost.com	text
amazon.com	text
reuters.com	text
hubpages.com	misc
msn.com	text
telegraph.co.uk	text
fbcdn.net	image
squidoo.com	misc
dailymail.co.uk	text
photobucket.com	image
latimes.com	text
tinypic.com	image
bit.ly	misc
minus.com	image

Table A.4.: The top 41 through 100 domains and the manually assigned content category is shown. The domains are in descending order of number of submissions to the respective domain.

Domain	Type	Domain	Type
wired.com	text	cnet.com	text
imageshack.us	image	livejournal.com	text
wsj.com	text	time.com	text
alternet.org	text	talkingpointsmemo.com	text
ezinearticles.com	text	etsy.com	text
npr.org	text	gizmodo.com	text
salon.com	text	goarticles.com	text
rawstory.com	text	steampowered.com	text
bandcamp.com	audio	weebly.com	misc
foxnews.com	text	articlesbase.com	text
wikimedia.org	image	slate.com	text
thinkprogress.org	text	omegle.com	text
allvoices.com	text	politico.com	text
liveleak.com	video	cheezburger.com	image
arstechnica.com	text	theglobeandmail.com	text
twitch.tv	video	tinyurl.com	misc
buzzfeed.com	text	amazonaws.com	misc
onlywire.com	misc	associatedcontent.com	misc
cbc.ca	text	tubemonsoon.com	video
bloomberg.com	text	gawker.com	text
imdb.com	text	kickstarter.com	text
theatlantic.com	text	gifsound.com	image
usatoday.com	text	boston.com	text
cbsnews.com	text	sfgate.com	text
forbes.com	text	engadget.com	text
craigslist.org	text	abc.net.au	text
dailykos.com	text	nasa.gov	image
businessinsider.com	text	usspost.com	text
typepad.com	text	theonion.com	text
independent.co.uk	text	economist.com	text

Appendix B.

Complete Correlation Results

The following tables present the complete correlation results both for *Pearson* correlation in section B.1 and *Spearman* correlation in section B.2.

B.1. Pearson Correlation

Table B.1.: The Pearson correlation coefficient and significance test p for the distributions of score, comments, upvotes, downvotes, total votes and title length to each other over all submissions is shown.

All - Pearson	**coef.**	p
Score to Comments:	0.36	0.00
Score to Votes:	0.76	0.00
Score to Ups:	0.80	0.00
Score to Downs:	0.69	0.00
Score to Title:	0.01	0.00
Comments to Votes:	0.35	0.00
Comments to Ups:	0.36	0.00
Comments to Downs:	0.34	0.00
Comments to Title:	0.03	0.00
Votes to Ups:	1.00	0.00
Votes to Downs:	1.00	0.00
Votes to Title:	-0.00	0.00

Table B.2.: The Pearson correlation coefficient and significance test p for the distributions of score, comments, upvotes, downvotes, total votes and title length to each other over all self submissions is shown.

self - Pearson	**coef.**	p
Score to Comments:	0.41	0.00
Score to Votes:	0.72	0.00
Score to Ups:	0.78	0.00
Score to Downs:	0.63	0.00
Score to Title:	0.04	0.00
Comments to Votes:	0.43	0.00
Comments to Ups:	0.44	0.00
Comments to Downs:	0.41	0.00
Comments to Title:	0.02	0.00
Votes to Ups:	1.00	0.00
Votes to Downs:	0.99	0.00
Votes to Title:	0.03	0.00

Table B.3.: The Pearson correlation coefficient and significance test p for the distributions of score, comments, upvotes, downvotes, total votes and title length to each other over all image submissions is shown.

image - Pearson	**coef.**	p
Score to Comments:	0.62	0.00
Score to Votes:	0.76	0.00
Score to Ups:	0.81	0.00
Score to Downs:	0.71	0.00
Score to Title:	0.01	0.00
Comments to Votes:	0.72	0.00
Comments to Ups:	0.72	0.00
Comments to Downs:	0.71	0.00
Comments to Title:	0.04	0.00
Votes to Ups:	1.00	0.00
Votes to Downs:	1.00	0.00
Votes to Title:	0.00	0.00

Table B.4.: The Pearson correlation coefficient and significance test p for the distributions of score, comments, upvotes, downvotes, total votes and title length to each other over all text submissions is shown.

text - Pearson	**coef.**	p
Score to Comments	0.73	0.00
Score to Votes	0.80	0.00
Score to Ups	0.85	0.00
Score to Downs	0.72	0.00
Score to Title	0.09	0.00
Comments to Votes	0.74	0.00
Comments to Ups	0.76	0.00
Comments to Downs	0.70	0.00
Comments to Title	0.07	0.00
Votes to Ups	1.00	0.00
Votes to Downs	0.99	0.00
Votes to Title	0.05	0.00

Table B.5.: The Pearson correlation coefficient and significance test p for the distributions of score, comments, upvotes, downvotes, total votes and title length to each other over all submissions in r/AskReddit is shown.

r/AskReddit - Pearson	**coef.**	p
Score to Comments:	0.77	0.00
Score to Votes:	0.81	0.00
Score to Ups:	0.85	0.00
Score to Downs:	0.76	0.00
Score to Title:	0.05	0.00
Comments to Votes:	0.70	0.00
Comments to Ups:	0.72	0.00
Comments to Downs:	0.67	0.00
Comments to Title:	0.02	0.00
Votes to Ups:	1.00	0.00
Votes to Downs:	1.00	0.00
Votes to Title:	0.04	0.00

Table B.6.: The Pearson correlation coefficient and significance test p for the distributions of score, comments, upvotes, downvotes, total votes and title length to each other over all submissions in r/funny is shown.

r/funny - Pearson	**coef.**	p
Score to Comments:	0.64	0.00
Score to Votes:	0.84	0.00
Score to Ups:	0.86	0.00
Score to Downs:	0.82	0.00
Score to Title:	-0.01	0.00
Comments to Votes:	0.74	0.00
Comments to Ups:	0.74	0.00
Comments to Downs:	0.73	0.00
Comments to Title:	0.01	0.00
Votes to Ups:	1.00	0.00
Votes to Downs:	1.00	0.00
Votes to Title:	-0.01	0.00

Table B.7.: The Pearson correlation coefficient and significance test p for the distributions of score, comments, upvotes, downvotes, total votes and title length to each other over all submissions in r/worldnews is shown.

r/worldnews - Pearson	**coef.**	p
Score to Comments	0.80	0.00
Score to Votes	0.87	0.00
Score to Ups	0.90	0.00
Score to Downs	0.81	0.00
Score to Title	0.11	0.00
Comments to Votes	0.79	0.00
Comments to Ups	0.81	0.00
Comments to Downs	0.76	0.00
Comments to Title	0.08	0.00
Votes to Ups	1.00	0.00
Votes to Downs	1.00	0.00
Votes to Title	0.07	0.00

B.2. Spearman Correlation

Table B.8.: The Spearman correlation coefficient and significance test p for the distributions of score, comments, upvotes, downvotes, total votes and title length to each other over all submissions is shown.

All - Spearman	**coef.**	p
Score to Comments:	0.44	0.00
Score to Votes:	0.52	0.00
Score to Ups:	0.69	0.00
Score to Downs:	0.21	0.00
Score to Title:	0.04	0.00
Comments to Votes:	0.60	0.00
Comments to Ups:	0.61	0.00
Comments to Downs:	0.50	0.00
Comments to Title:	0.11	0.00
Votes to Ups:	0.96	0.00
Votes to Downs:	0.90	0.00
Votes to Title:	0.02	0.00

Table B.9.: The Spearman correlation coefficient and significance test p for the distributions of score, comments, upvotes, downvotes, total votes and title length to each other over all self submissions is shown.

self - Spearman	**coef.**	p
Score to Comments:	0.46	0.00
Score to Votes:	0.51	0.00
Score to Ups:	0.72	0.00
Score to Downs:	0.09	0.00
Score to Title:	0.01	0.00
Comments to Votes:	0.66	0.00
Comments to Ups:	0.67	0.00
Comments to Downs:	0.51	0.00
Comments to Title:	0.05	0.00
Votes to Ups:	0.94	0.00
Votes to Downs:	0.85	0.00
Votes to Title:	0.07	0.00

Table B.10.: The Spearman correlation coefficient and significance test p for the distributions of score, comments, upvotes, downvotes, total votes and title length to each other over all image submissions is shown.

image - Spearman	**coef.**	p
Score to Comments:	0.54	0.00
Score to Votes:	0.67	0.00
Score to Ups:	0.80	0.00
Score to Downs:	0.36	0.00
Score to Title:	0.02	0.00
Comments to Votes:	0.69	0.00
Comments to Ups:	0.68	0.00
Comments to Downs:	0.60	0.00
Comments to Title:	0.11	0.00
Votes to Ups:	0.97	0.00
Votes to Downs:	0.89	0.00
Votes to Title:	-0.04	0.00

Table B.11.: The Spearman correlation coefficient and significance test p for the distributions of score, comments, upvotes, downvotes, total votes and title length to each other over all text submissions is shown.

text - Spearman	**coef.**	p
Score to Comments	0.43	0.00
Score to Votes	0.56	0.00
Score to Ups	0.72	0.00
Score to Downs	0.27	0.00
Score to Title	0.10	0.00
Comments to Votes	0.61	0.00
Comments to Ups	0.60	0.00
Comments to Downs	0.56	0.00
Comments to Title	0.21	0.00
Votes to Ups	0.96	0.00
Votes to Downs	0.91	0.00
Votes to Title	0.19	0.00

Table B.12.: The Spearman correlation coefficient and significance test p for the distributions of score, comments, upvotes, downvotes, total votes and title length to each other over all submissions in r/AskReddit is shown.

r/AskReddit - Spearman	**coef.**	p
Score to Comments:	0.34	0.00
Score to Votes:	0.20	0.00
Score to Ups:	0.52	0.00
Score to Downs:	-0.13	0.00
Score to Title:	0.07	0.00
Comments to Votes:	0.71	0.00
Comments to Ups:	0.73	0.00
Comments to Downs:	0.57	0.00
Comments to Title:	0.08	0.00
Votes to Ups:	0.89	0.00
Votes to Downs:	0.91	0.00
Votes to Title:	0.10	0.00

Table B.13.: The Spearman correlation coefficient and significance test p for the distributions of score, comments, upvotes, downvotes, total votes and title length to each other over all submissions in r/funny is shown.

r/funny - Spearman	**coef.**	p
Score to Comments:	0.34	0.00
Score to Votes:	0.34	0.00
Score to Ups:	0.55	0.00
Score to Downs:	0.08	0.00
Score to Title:	0.00	0.00
Comments to Votes:	0.59	0.00
Comments to Ups:	0.57	0.00
Comments to Downs:	0.54	0.00
Comments to Title:	0.03	0.00
Votes to Ups:	0.94	0.00
Votes to Downs:	0.93	0.00
Votes to Title:	-0.04	0.00

Table B.14.: The Spearman correlation coefficient and significance test p for the distributions of score, comments, upvotes, downvotes, total votes and title length to each other over all submissions in r/worldnews is shown.

r/worldnews - Spearman	**coef.**	p
Score to Comments	0.37	0.00
Score to Votes	0.40	0.00
Score to Ups	0.54	0.00
Score to Downs	0.27	0.00
Score to Title	0.19	0.00
Comments to Votes	0.50	0.00
Comments to Ups	0.50	0.00
Comments to Downs	0.49	0.00
Comments to Title	0.18	0.00
Votes to Ups	0.96	0.00
Votes to Downs	0.97	0.00
Votes to Title	0.31	0.00

Appendix C.

Additional Results on Submission Time

In the following, additional results regarding submission time and the impact on attention are shown. The figures show attention over all submissions as well as for specific types of content and exemplary subreddits.

C.1. Total and Average Attention per Weekday and per Hour

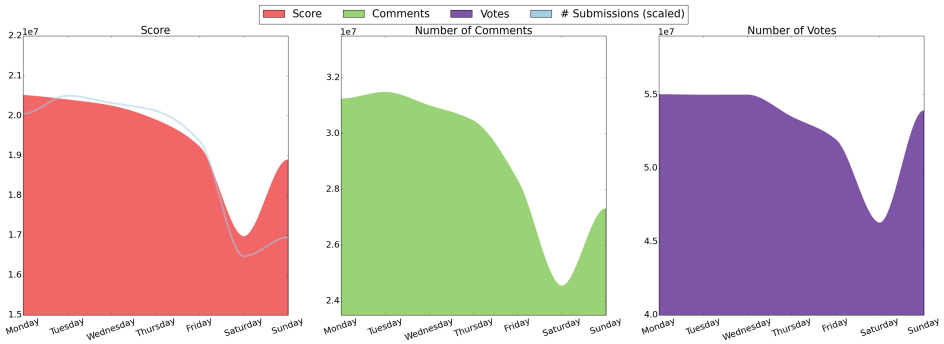


Figure C.1.: The total amount of score, comments and votes for self submissions posted on each weekday is shown. The blue line symbolizes the number of submissions for each day in a scaled manner, these are not the actual submission values.

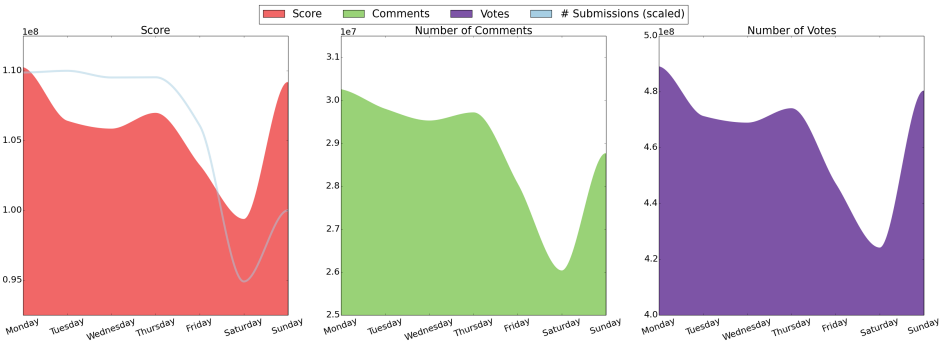


Figure C.2.: The total amount of score, comments and votes for image submissions posted on each weekday is shown. The blue line symbolizes the number of submissions for each day in a scaled manner, these are not the actual submission values.

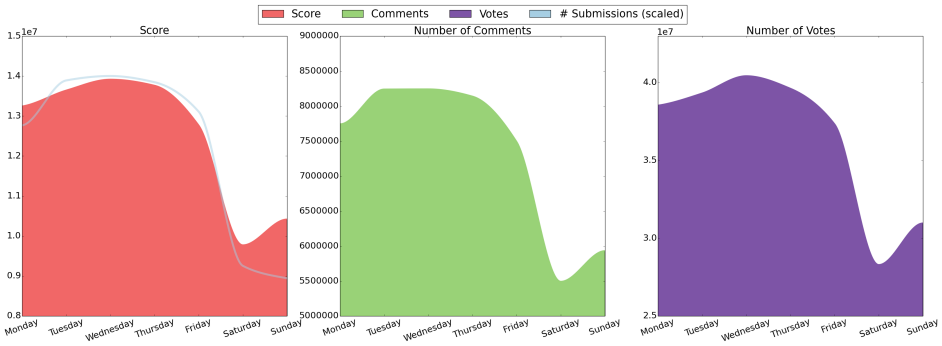


Figure C.3.: The total amount of score, comments and votes for text submissions posted on each weekday is shown. The blue line symbolizes the number of submissions for each day in a scaled manner, these are not the actual submission values.

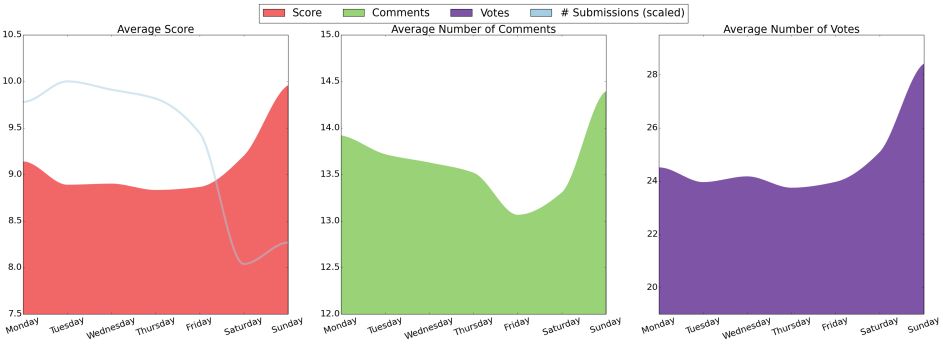


Figure C.4.: The average amount of score, comments and votes for self submissions posted on each weekday is shown. The blue line symbolizes the number of submissions for each day in a scaled manner, these are not the actual submission values.

H

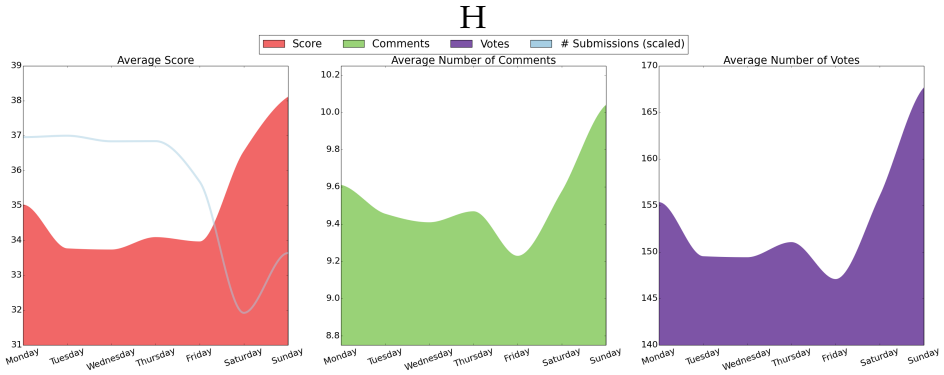


Figure C.5.: The average amount of score, comments and votes for image submissions posted on each weekday is shown. The blue line symbolizes the number of submissions for each day in a scaled manner, these are not the actual submission values.

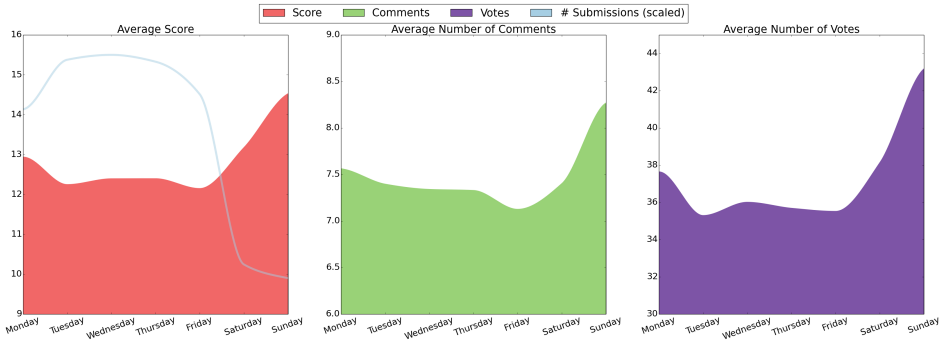


Figure C.6.: The average amount of score, comments and votes for text submissions posted on each weekday is shown. The blue line symbolizes the number of submissions for each day in a scaled manner, these are not the actual submission values.

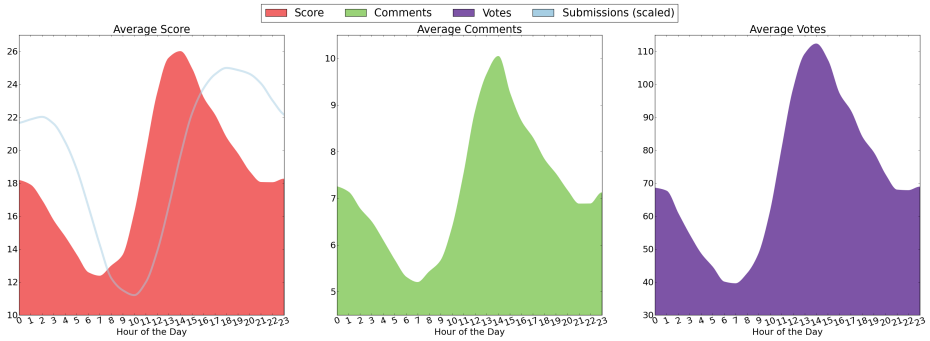


Figure C.7.: The average amount of score, comments and votes for submissions posted in each hour of the day is shown. The blue line symbolizes the number of submissions for each hour in a scaled manner, these are not the actual submission values.

C.2. Total and Average Attention Heatmaps

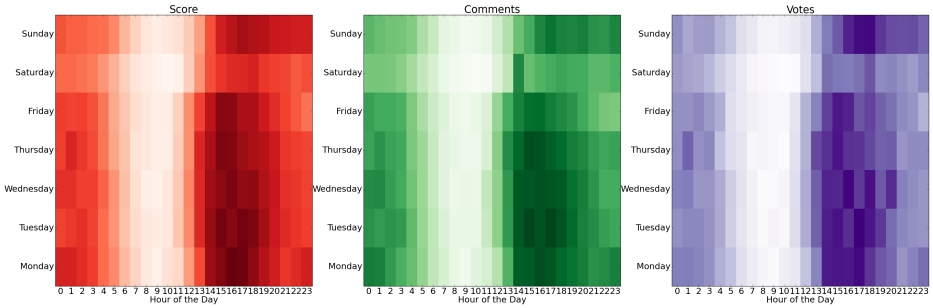


Figure C.8.: Heatmaps combining the weekly and hourly total attention for self submissions are shown. Each box symbolizes one hour on a specific day. The darker the box, the higher the value at this hour and day, subsequently lower values are signified by lighter boxes.

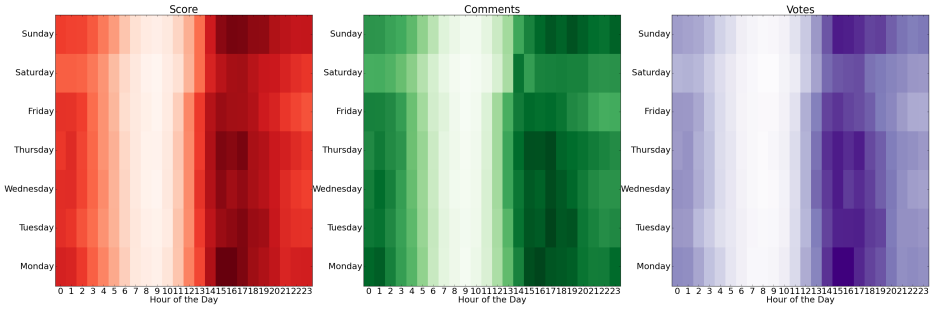


Figure C.9.: Heatmaps combining the weekly and hourly total attention for image submissions are shown. Each box symbolizes one hour on a specific day. The darker the box, the higher the value at this hour and day, subsequently lower values are signified by lighter boxes.

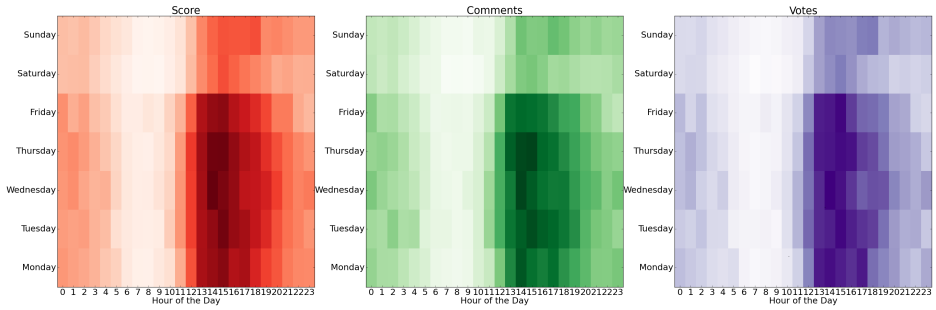


Figure C.10.: Heatmaps combining the weekly and hourly total attention for text submissions are shown. Each box symbolizes one hour on a specific day. The darker the box, the higher the value at this hour and day, subsequently lower values are signified by lighter boxes.

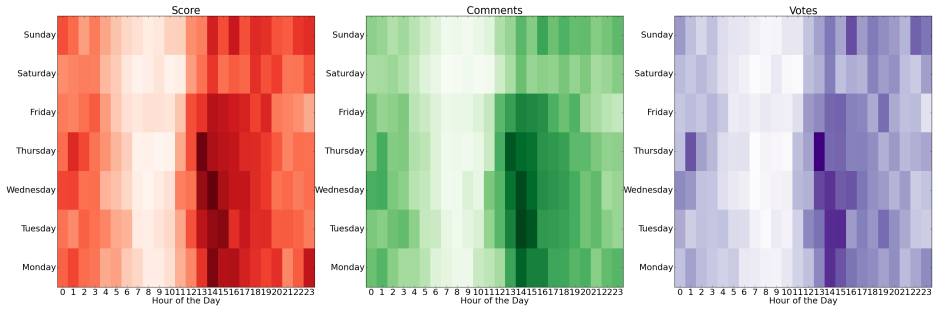


Figure C.11.: Heatmaps combining the weekly and hourly total attention in in *r/AskReddit* are shown. Each box symbolizes one hour on a specific day. The darker the box, the higher the value at this hour and day, subsequently lower values are signified by lighter boxes.

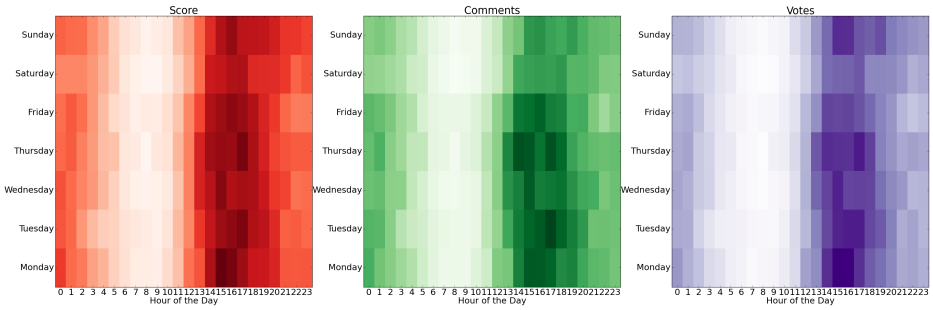


Figure C.12.: Heatmaps combining the weekly and hourly total attention in *r/funny* are shown. Each box symbolizes one hour on a specific day. The darker the box, the higher the value at this hour and day, subsequently lower values are signified by lighter boxes.

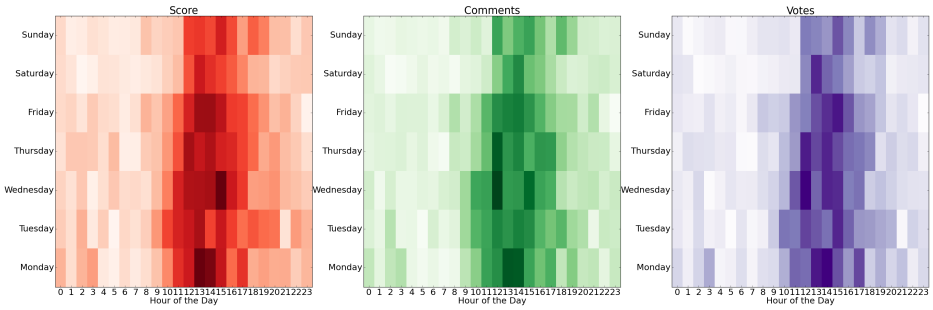


Figure C.13.: Heatmaps combining the weekly and hourly total attention in *r/worldnews* are shown. Each box symbolizes one hour on a specific day. The darker the box, the higher the value at this hour and day, subsequently lower values are signified by lighter boxes.

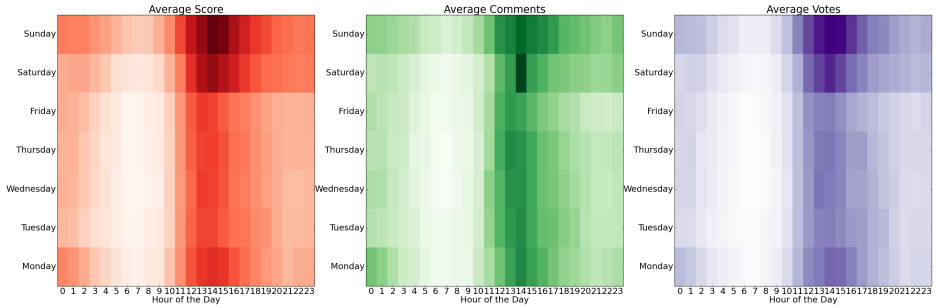


Figure C.14.: Heatmaps combining the weekly and hourly average attention are shown. Each box symbolizes one hour on a specific day. The darker the box, the higher the value at this hour and day, subsequently lower values are signified by lighter boxes.

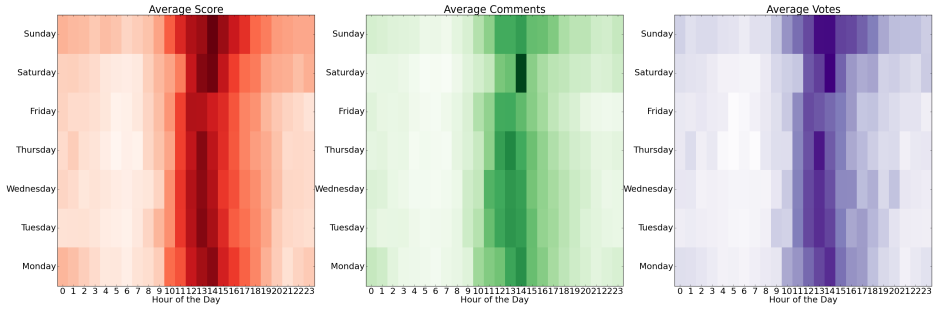


Figure C.15.: Heatmaps combining the weekly and hourly average attention for self submissions are shown. Each box symbolizes one hour on a specific day. The darker the box, the higher the value at this hour and day, subsequently lower values are signified by lighter boxes.

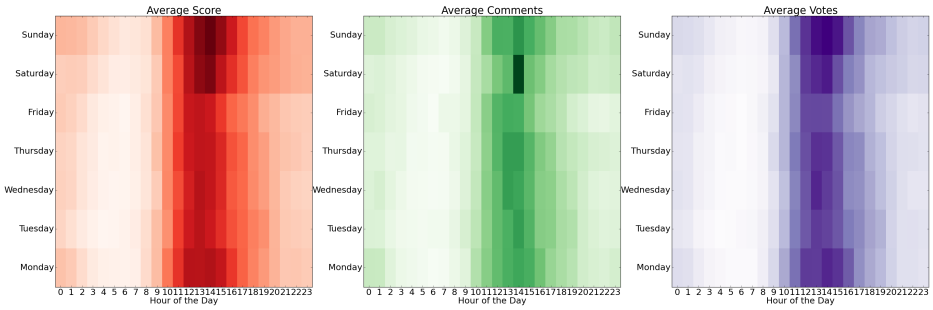


Figure C.16.: Heatmaps combining the weekly and hourly average attention for image submissions are shown. Each box symbolizes one hour on a specific day. The darker the box, the higher the value at this hour and day, subsequently lower values are signified by lighter boxes.

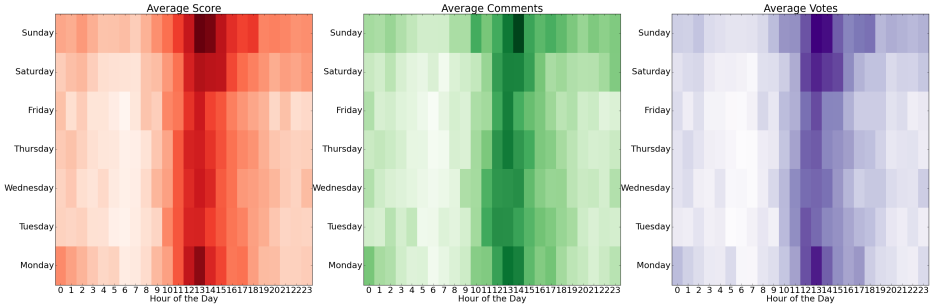


Figure C.17.: Heatmaps combining the weekly and hourly average attention for text submissions are shown. Each box symbolizes one hour on a specific day. The darker the box, the higher the value at this hour and day, subsequently lower values are signified by lighter boxes.

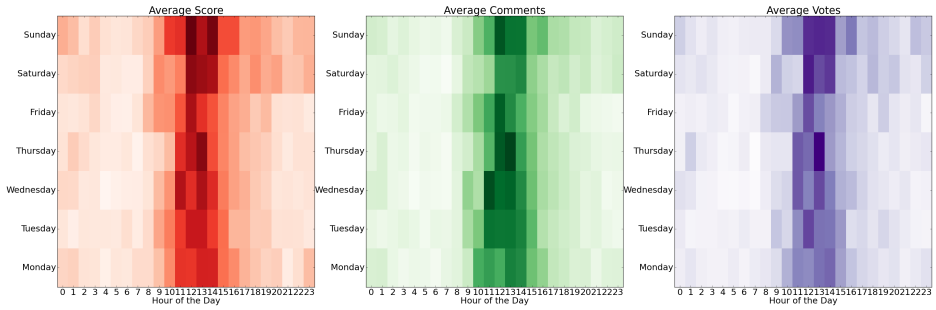


Figure C.18.: Heatmaps combining the weekly and hourly average attention in in *r/AskReddit* are shown. Each box symbolizes one hour on a specific day. The darker the box, the higher the value at this hour and day, subsequently lower values are signified by lighter boxes.

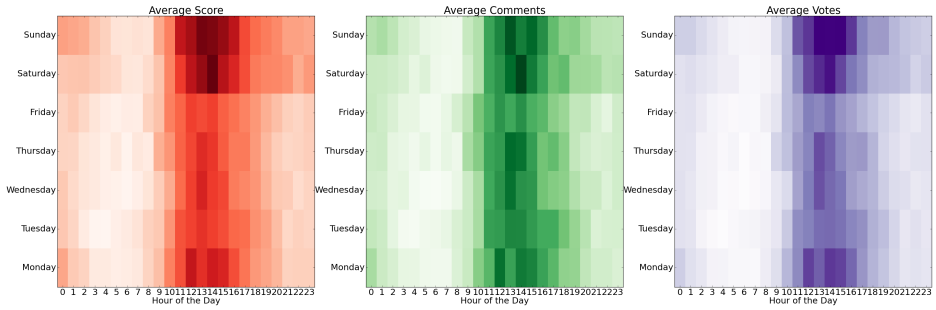


Figure C.19.: Heatmaps combining the weekly and hourly average attention in *r/funny* are shown. Each box symbolizes one hour on a specific day. The darker the box, the higher the value at this hour and day, subsequently lower values are signified by lighter boxes.

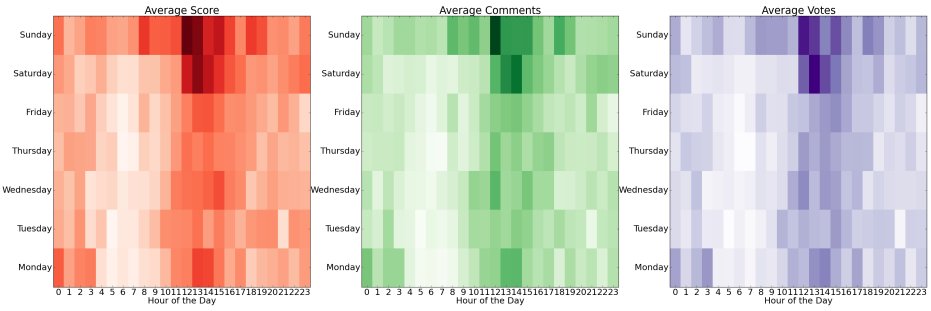
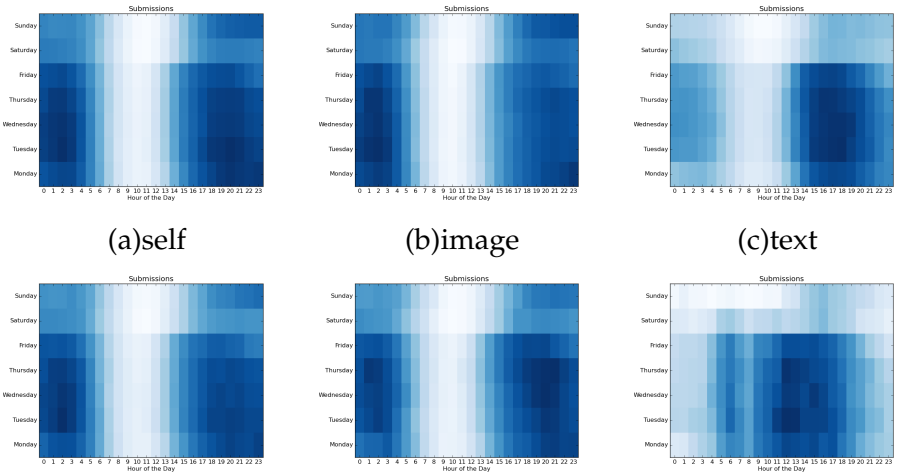


Figure C.20.: Heatmaps combining the weekly and hourly average attention in *r/worldnews* are shown. Each box symbolizes one hour on a specific day. The darker the box, the higher the value at this hour and day, subsequently lower values are signified by lighter boxes.

C.3. Submissions Heatmaps



(a)self (b)image (c)text

(d)r/AskReddit (e)r/funny (f)r/worldnews

Figure C.21.: Heatmaps combining the weekly and hourly submission numbers are shown. Each box symbolizes one hour on a specific day. The darker the box, the higher the value at this hour and day, subsequently lower values are signified by lighter boxes.

Appendix D.

Additional Classification Results

The following tables present the top keywords extracted from the *binary naive Bayes* and the *multinomial naive Bayes* classification experiments.

Table D.1.: Top 10 keywords of the binary naive Bayes classifier for the top 20 subreddits of 2012.

funny	got, today, new, day, facebook, think, feel
videos	watch, music, time, reddit, best, man, new
gaming	help, best, time, video, new, play, steam
aww	baby, friend, day, old, meet, new, puppy
technology	facebook, online, best, free, web, com, iphone
circlejerk	ron, paul, post, upvotes, dae, karma, upvote
fffffffuuuuuuuuuuu	happens, know, happened, fixed, troll, day, comic
trees	today, guys, got, ent, time, high, smoke
pics	best, time, got, new, think, picture, day
atheism	think, people, jesus, religious, atheist, christians, religion
todayilearned	new, year, actually, man, used, called, years
AskReddit	did, need, best, think, thing, know, does
Music	amp, love, live, cover, youtube, band, video
AdviceAnimals	brian, today, good, guy, feel, bad, world
WTF	today, fuck, guy, friend, facebook, know, think
worldnews	syria, china, amp, iran, world, com, online
politics	news, election, people, gop, president, romney, obama
Minecraft	survival, play, think, build, world, suggestion, help
tf2trade	key, hats, pc, buds, store, offers, metal
leagueoflegends	stream, skin, ranked, legends, riot, elo, new

Table D.2.: Top 10 keywords of the multinomial naive Bayes classifier for the top 20 subreddits from 2012.

funny	day, know, fixed, friend, oh, think, time
videos	new, time, amazing, watch, music, man, best
gaming	best, help, playing, video, new, play, steam
aww	new, baby, kitty, cute, kitten, meet, little
technology	ipad, development, mobile, android, windows, web, new
circlejerk	paul, ron, circlejerk, upvotes, til, dae, karma
ffffffuuuuuuuuuuu	know, hate, day, happens, troll, happened, comic
trees	smoking, weed, guys, time, 10, ent, high
pics	time, new, cat, got, think, day, friend
atheism	christians, bible, religious, jesus, religion, atheist, god
todayilearned	make, man, actually, used, world, called, til
AskReddit	think, did, need, know, best, people, does
Music	album, love, live, youtube, video, cover, new
AdviceAnimals	happened, guy, brian, reddit, feel, problems, bad
WTF	guy, really, saw, think, know, man, facebook
worldnews	china, india, world, syria, com, iran, new
politics	america, vote, republican, election, gop, romney, obama
Minecraft	play, world, mod, survival, build, suggestion, help
tf2trade	key, pc, hats, buds, store, offers, metal
leagueoflegends	stream, champion, new, ranked, legends, game, riot

Appendix E.

Additional Results for Trend Discovery and Analysis

Section E.1 displays additional exemplary results from trend discovery approach via modified TFIDF, in section E.2 the complete results from the trend analysis via classifier coefficients can be found.

E.1. Additional Results from Trend Discovery via Modified TFIDF

Table E.1.: Overall trending words for submissions in the top 20 subreddits of 2012, calculated at each month in comparison to the whole year. Only submissions from 2012 are valid for this experiment.

January	cinsere, pipa, sopa, blackout, sotu, ces, cundiff
February	valentine, whitney, 12w08a, superbowl, lin, harrelson, grammys
March	farugo, kony, limbaugh, kony2012, weedception, patricks, me3
April	cispa, ipl4, hologram, coachella, easter, zeddie, april
May	yauch, mca, cinco, mht, seau, supermoon, eclipse
June	e3, karmanaut, bradbury, venus, salts, khil, prometheus
July	boson, higgs, canabbis, july, aurora, tdkr, particle
August	doomba, gamescom, doombas, mckayla, trapwire, 12w32a, curiosity
September	dnc, mesa, supermanv2, purrer, 12w38a, bacile, 12w37a
October	baumgartner, binders, stratos, frankenstorm, felix, bayonets, romnesia
November	thanksgiving, petraeus, broadwell, morals, hostess, twinkies, gaza
December	newtown, mallard, lanza, knettel37, aristocat, merry, xmas

Table E.2.: Monthly trending words for submissions in the top 20 subreddits of 2012, calculated at each month in comparison to month before. Only submissions from 2012 are valid for this experiment.

January	cinsere, concordia, acta, resentful, ahlquist, eel, sejuani
February	whitney, rampart, lin, harrelson, komen, breadfriend, fiora
March	kony, trayvon, kony2012, zimmerman, berks, farugo, limbaugh
April	cispa, hecarim, photogenic, hologram, varus, ridiculously, zeddie
May	yauch, mca, sendak, eclipse, darius, cinco, beastie
June	rainblower, scorched, lollichop, bradbury, sandusky, overly, attached
July	relatable, zyra, higgs, boson, ouya, fil, slender
August	botkiller, doomba, mckayla, akin, rengar, maroney, sikh
September	zix, kha, benghazi, pyrotechnic, endeavour, ios6, 47%
October	sandy, binders, baumgartner, frankenstorm, carbonado, stratos, violentacrez
November	petraeus, twinkies, hostess, powerball, secede, twinkie, secession
December	newtown, mallard, lanza, aristocat, nra, shootings, inouye

Table E.3.: Overall trending keywords in 2012 over all submissions in r/AskReddit calculated at every month in comparison to the whole year. Only submissions from 2012 are valid for this experiment. The subreddit r/funny is part of the 20 most active subreddits in the noted timespan.

r/AskReddit	
January	sopa, pipa, blackout, megaupload, acta, ndaa, 18th
February	valentine, superbowl, woody, harrelson, whitney, oscars, halftime
March	kony, megamillions, trayvon, mega, limbaugh, patricks, april
April	cispa, april, fools, easter, photogenic, prom, timeline
May	avengers, diablo, mothers, graduation, memorial, slabs, cinco
June	karmanaut, prometheus, karen, circlejerk, salts, klein, fathers
July	canabbis, aurora, higgs, boson, rises, tdkr, knight
August	rover, mars, fil, curiosity, olympics, armstrong, assange
September	homecoming, embassies, dnc, muslims, 11, september, 9
October	felix, halloween, baumgartner, stratos, hurricane, amanda, violentacrez
November	thanksgiving, gaza, powerball, petraeus, hostess, skyfall, twinkies
December	newtown, lanza, connecticut, christmas, resolutions, shootings, merry

Table E.4.: Overall trending keywords in 2012 over all submissions in r/funny calculated at every month in comparison to the whole year. Only submissions from 2012 are valid for this experiment. The subreddit r/funny is part of the 20 most active subreddits in the noted timespan.

r/funny	
January	sopa, blackout, pipa, cundiff, megaupload, mlk, january
February	valentine, whitney, superbowl, lin, grammys, harrelson, february
March	kony, referential, kony2012, patricks, coney, uganda, limbaugh
April	april, easter, fools, hologram, maize, rpg, ridiculously
May	cinco, mayo, diablo, avengers, eclipse, mca, haaaaaa
June	venus, prometheus, e3, salts, answersfull, #039, humanized
July	higgs, boson, july, kodiak, rises, 4th, uniforms
August	doomba, curiosity, rover, mars, mckayla, doombas, isaac
September	refs, dnc, byo, vowels, september, ios6, nfl
October	binders, baumgartner, felix, sandy, stratos, hurricane, debate
November	thanksgiving, hostess, petraeus, twinkies, november, movember, election
December	merry, christmas, xmas, santa, december, wrapping, knettel37

Table E.5.: Overall trending keywords in 2012 over all submissions in r/worldnews calculated at every month in comparison to the whole year. Only submissions from 2012 are valid for this experiment. The subreddit r/worldnews is part of the 20 most active subreddits in the noted timespan.

	r/worldnews
January	pipa, sopa, paterno, concordia, megaupload, urinating, january
February	whitney, feb, colvin, vostok, homs, kashgari, february
March	kony, toulouse, invisible, masturbating, joseph, emo, bales
April	cispa, tele, compensatory, zimmerman, renters, april, dick
May	sendak, strong, maurice, fotograf, seau, profi, chen
June	rodney, bradbury, dingo, magnotta, khil, eduard, lauren
July	boson, higgs, aurora, knight, batman, holmes, sally
August	sikh, armstrong, neil, behel, boomsy, preseason, pussy
September	duncan, clarke, bacile, consulate, innocence, nakoula, griselda
October	lucasfilm, baumgartner, felix, stratos, uggs, malala, todd
November	petraeus, jabari, koli, gaza, hamas, hostess, twinkies
December	newtown, lanza, elementary, connecticut, hook, gangrape, desember

Table E.6.: Overall trending keywords in 2012 over all submissions in r/politics calculated at every month in comparison to the whole year. Only submissions from 2012 are valid for this experiment. The subreddit r/politics is part of the 20 most active subreddits in the noted timespan.

	r/politics
January	sotu, pipa, blackout, huntsman, megaupload, cordray, sopa
February	komen, cpac, handel, nutshell, bowl, hoekstra, toews
March	kony, advertiser, hoodie, slut, carbonite, breitbart, fluke
April	cispa, rosen, oaksterdam, gsa, nugent, derbyshire, zimmerman
May	saverin, nato, nc, memorial, lugar, worley, ipo
June	contempt, upheld, survives, bilderberg, scotus, munro, leonhart
July	aurora, libor, saxon, batman, abedin, huma, holmes
August	sikh, isaac, jameson, camerawoman, trapwire, raub, jenna
September	nakoula, dnc, bacile, embassies, 47%, refs, univision
October	bayonets, binders, romnesia, hofstra, bird, rumble, mourdock
November	broadwell, petraeus, secession, secede, gaza, hostess, thanksgiving
December	newtown, lanza, costas, hook, hagel, lapierre, inouye

Table E.7.: Overall trending keywords in 2012 over all submissions in r/leagueoflegends calculated at every month in comparison to the whole year. Only submissions from 2012 are valid for this experiment. The subreddit r/funny is part of the 20 most active subreddits in the noted timespan.

	r/leagueoflegends
January	kiev, sopa, sleepshotgg, sejuani, kings, mtl, jared
February	valentine, february, revel, lunar, revelry, nautilus, heroic
March	hanover, hannover, march, weo, cypher, lulu, fatal1ty
April	ipl4, april, bubble, gatekeeper, pop, hecarim, assembly
May	spectator, diablo, varus, darius, mode, spectate, featured
June	anaheim, gesl, xhazzard, june, pulsefire, summoned, rocks
July	july, thorns, ecc, zyra, poland, jayce, cappy
August	gamescom, raleigh, faceoff, august, pridestalker, retrospective, battlecast
September	voidreaver, september, zix, kha, oktoberfest, syndra, forellenlord
October	wc, isles, azf, playoffs, af, cheating, worlds
November	lone, zed, clash, dallas, lonestar, borders, jree
December	vi, enforcer, gifting, belle, cleavers, sightstone, christmas

Table E.8.: Monthly trending keywords in 2012 over all submissions in r/AskReddit calculated at every month in comparison to the month before. Only submissions from 2012 and December 2011 are valid for this experiment. The subreddit r/AskReddit is part of the 20 most active subreddits in the noted timespan.

	r/AskReddit
January	acta, blackout, megaupload, pipa, valentine, 18th, romney
February	harrelson, woody, valentine, whitney, rampart, lent, chris
March	kony, trayvon, fools, zimmerman, mega, april, limbaugh
April	cispa, photogenic, brick, 13th, easter, titanic, timeline
May	diablo, avengers, memorial, mothers, eclipse, carolina, graduation
June	obamacare, karen, prometheus, circlejerk, karmanaut, ufo, salts
July	aurora, higgs, boson, fil, tdkr, rises, olympics
August	rover, armstrong, mars, nasa, assange, curiosity, julian
September	islamic, muslims, halloween, 11, libya, homecoming, pirate
October	sandy, hurricane, felix, baumgartner, gawker, violentacrez, amanda
November	thanksgiving, powerball, gaza, secede, thankful, colorado, twinkies
December	shootings, newtown, connecticut, lanza, resolutions, shooting, wbc

Table E.9.: Monthly trending keywords in 2012 over all submissions in r/funny calculated at every month in comparison to the month before. Only submissions from 2012 and December 2011 are valid for this experiment. The subreddit r/funny is part of the 20 most active subreddits in the noted timespan.

r/funny	
January	blackout, santorum, pipa, eel, mlk, catfacts, megaupload
February	whitney, lin, harrelson, valentine, grammys, woody, houstone
March	kony, kony2012, limbaugh, uganda, daylight, referential, hunger
April	photogenic, ridiculously, hologram, rpg, easter, maize, 13th
May	cinco, mayo, diablo, eclipse, avengers, mammoth, mothers
June	attached, overly, funnyjunk, obamacare, oag, salts, venus
July	higgs, boson, tdkr, fil, olympics, uniforms, rises
August	doomba, mckayla, curiosity, maroney, rnc, mars, rover
September	refs, september, ios6, ios, nfl, dnc, vowels
October	sandy, binders, baumgartner, felix, hurricane, stratos, lehrer
November	hostess, twinkies, petraeus, thanksgiving, loophole, powerball, movember
December	finals, merry, wrapping, doomsday, forecast, amazes, xmas

Table E.10.: Monthly trending keywords in 2012 over all submissions in r/worldnews calculated at every month in comparison to the month before. Only submissions from 2012 and December 2011 are valid for this experiment. The subreddit r/worldnews is part of the 20 most active subreddits in the noted timespan.

r/worldnews	
January	concordia, megaupload, acta, urinating, pipa, paterno, marines
February	whitney, koran, bölüm, 48, shelling, colvin, homs
March	kony, toulouse, trayvon, joseph, masturbating, invisible, lulzsec
April	cispa, dre, breivik, prada, beats, couture, chanel
May	fotograf, sendak, houla, münchen, pitching, bark, strong
June	sandusky, rodney, bradbury, dingo, morsi, paraguay, asylum
July	transplantation, libor, fil, aleppo, aurora, chick, knight
August	armstrong, neil, sikh, isaac, julian, lance, curiosity
September	innocence, clarke, consulate, benghazi, filmmaker, duncan, film
October	sandy, malala, lucasfilm, yousafzai, nobel, baumgartner, felix
November	mcafee, petraeus, hostess, statehood, twinkies, tel, hamas
December	newtown, elementary, connecticut, hook, westboro, baptist, lanza

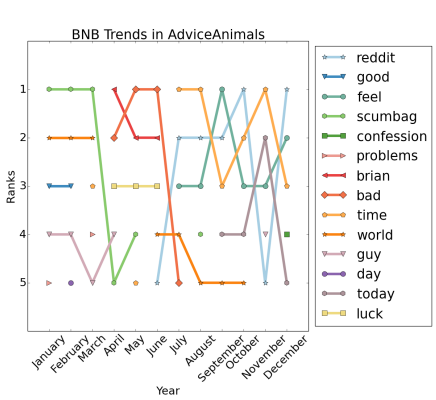
Table E.11.: Monthly trending keywords in 2012 over all submissions in r/politics calculated at every month in comparison to the month before. Only submissions from 2012 and December 2011 are valid for this experiment. The subreddit r/politics is part of the 20 most active subreddits in the noted timespan.

r/politics	
January	acta, sotu, 18th, megaupload, giffords, vermin, urinating
February	komen, cpac, hoekstra, parenthood, handel, eastwood, heartland
March	trayvon, kony, zimmerman, hoodie, fluke, rush, limbaugh
April	cispa, nugent, gsa, rosen, fallon, mystics, colombia
May	jp, jpmorgan, memorial, nato, saverin, indefinite, bain
June	scotus, scalia, furious, holder, upheld, obamacare, aca
July	fil, chick, aurora, anaheim, libor, batman, naacp
August	akin, clint, eastwood, todd, sikh, ryan, isaac
September	libya, benghazi, 47%, stevens, ambassador, consulate, embassies
October	binders, sandy, bayonets, jeep, binder, sesame, bird
November	petraeus, hostess, papa, secede, broadwell, secession, gaza
December	newtown, nra, hook, lanza, shootings, lapierre, inouye

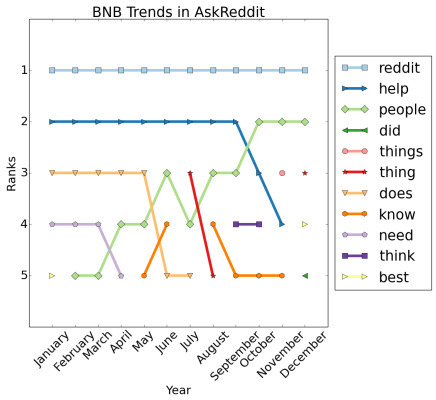
Table E.12.: Monthly trending keywords in 2012 over all submissions in r/leagueoflegends calculated at every month in comparison to the month before. Only submissions from 2012 and December 2011 are valid for this experiment. The subreddit r/leagueoflegends is part of the 20 most active subreddits in the noted timespan.

r/leagueoflegends	
January	sejuani, m5, ziggs, ipl, sopa, kings, nerfs
February	nautilus, fiora, ipl4, qualifier, shen, titan, february
March	lulu, cypher, iem, weo, 100k, hanover, sorceress
April	hecarim, varus, bubble, pop, code, sivir, ipl4
May	darius, draven, spectate, spectator, noxus, qualifier, ipl5
June	pfe, jayce, pulsefire, mlg, grounds, proving, anaheim
July	zyra, diana, thorns, july, ogn, xin, jayce
August	rengar, raleigh, regionals, syndra, arcade, gamescom, regional
September	zix, kha, voidreaver, diamond, oktoberfest, september, forellenlord
October	elise, isles, azf, shadow, tpa, honor, wc
November	nami, preseason, eternum, dreamhack, zed, borders, clash
December	vi, gifting, enforcer, cleaver, snowdown, cleavers, neon

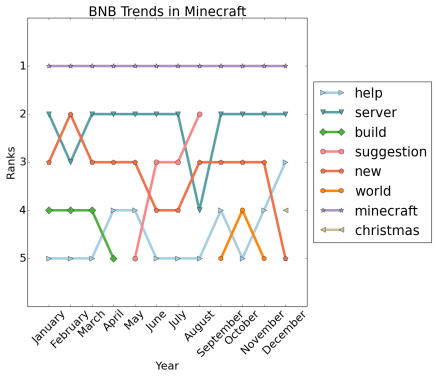
E.2. Trend Analysis via Classification Coefficients



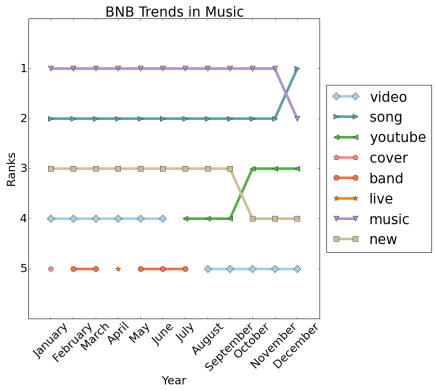
(a)AdviceAnimals



(b)AskReddit

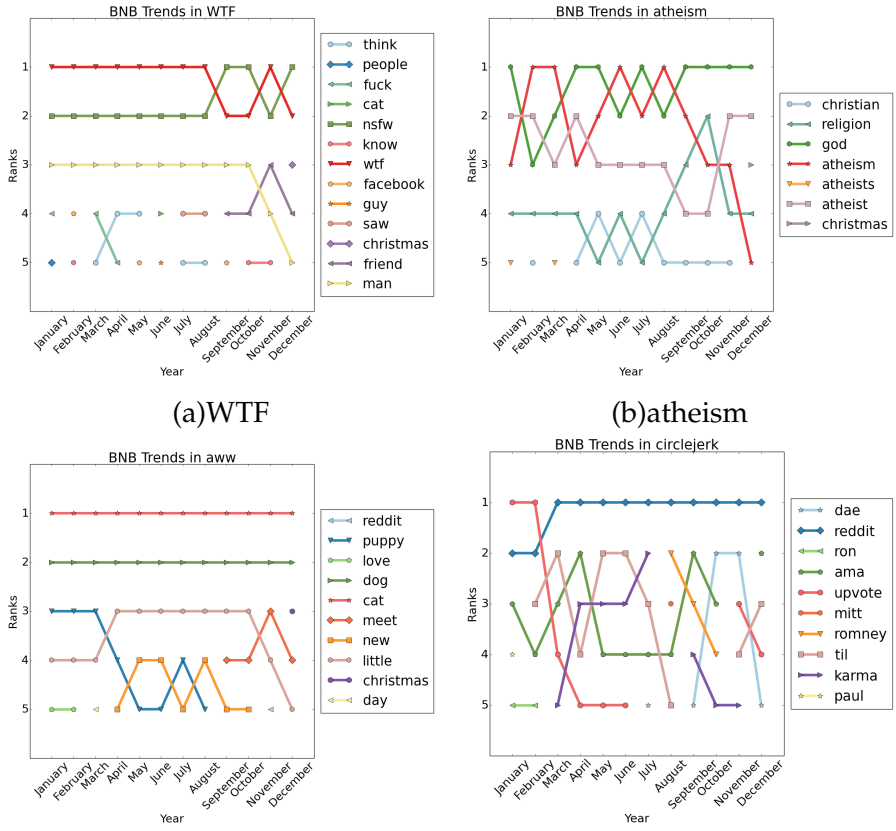


(c)Minecraft



(d)Music

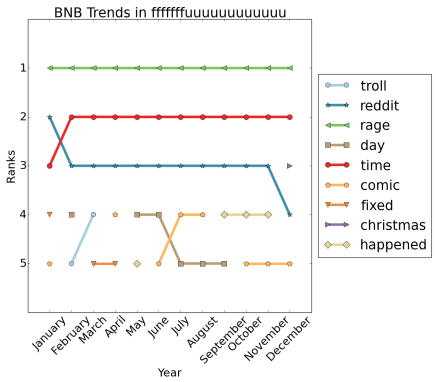
Figure E.1.: Trends by classification coefficients with binary naive Bayes for r/AdviceAnimals, r/AskReddit, r/Minecraft and r/Music



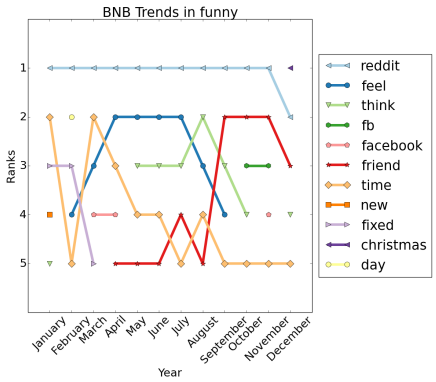
(a)WTF (b)atheism

(c)aww (d)circlejerk

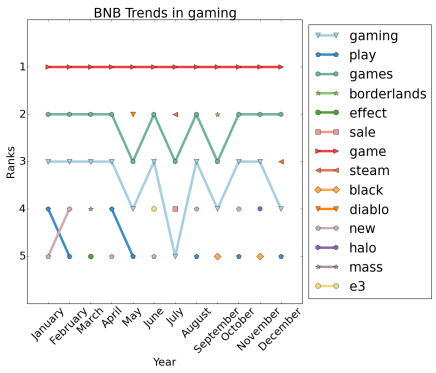
Figure E.2.: Trends by classification coefficients with binary naive Bayes for r/WTF, r/atheism, r/aww and r/circlejerk



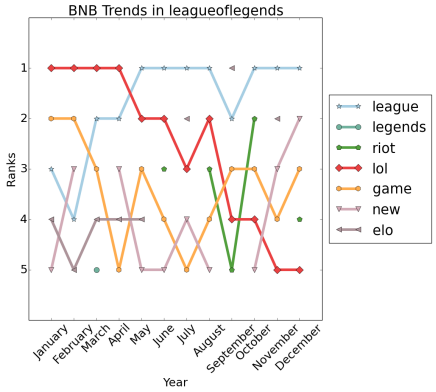
(a)fffffffuuuuuuuuuuu



(b)funny

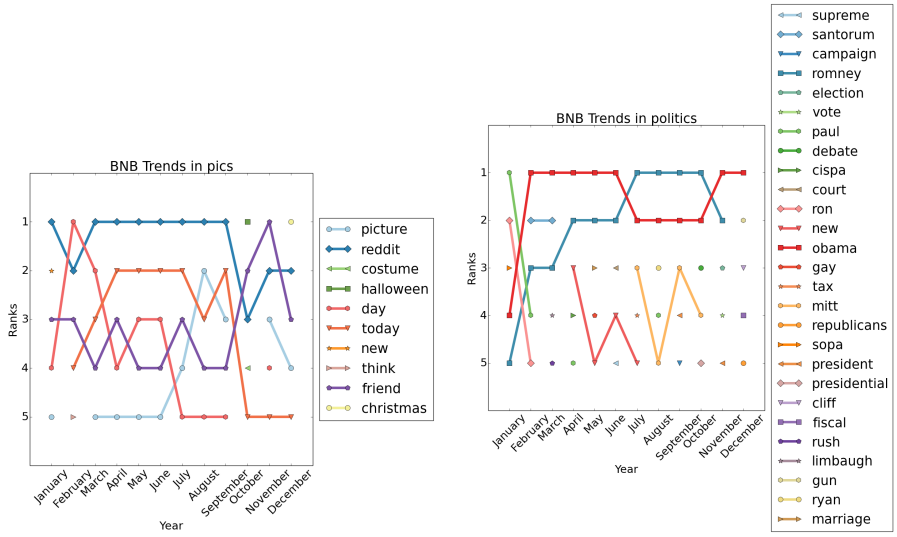


(c)gaming



(d)leagueoflegends

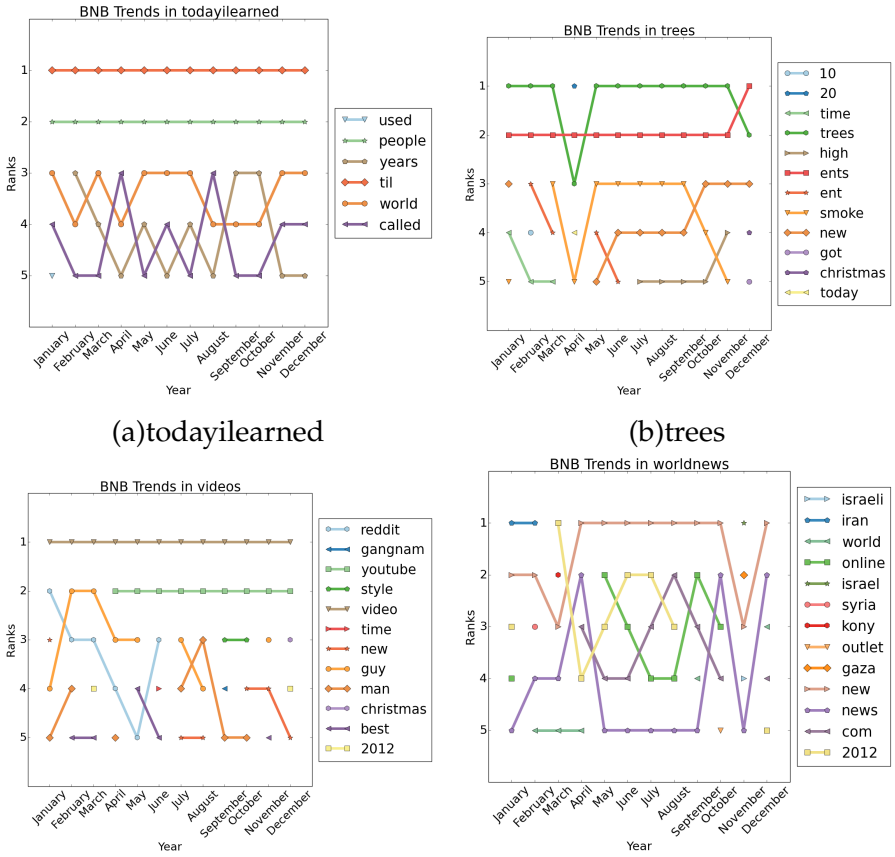
Figure E.3.: Trends by classification coefficients with binary naive Bayes for r/fffffffuuuuuuuuuuu, r/funny, r/gaming and r/leagueoflegends



(a)pics (b)politics

(c)technology (d)tf2trade

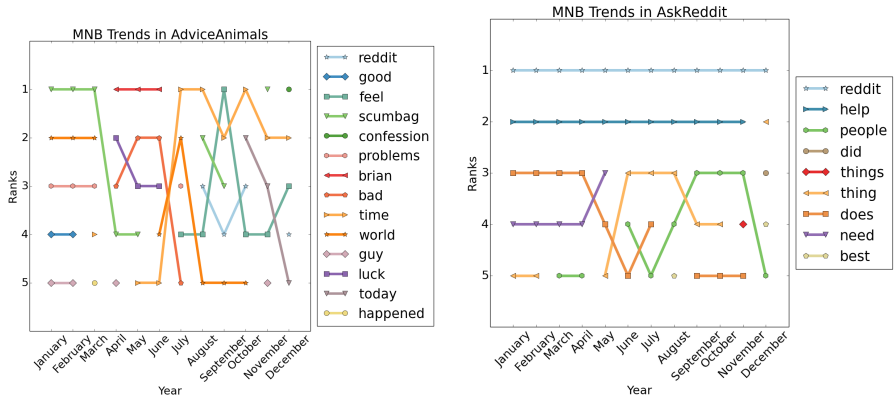
Figure E.4.: Trends by classification coefficients with binary naive Bayes for r/pics, r/politics, r/technology and r/tf2trade



(a)todayilearned (b)trees

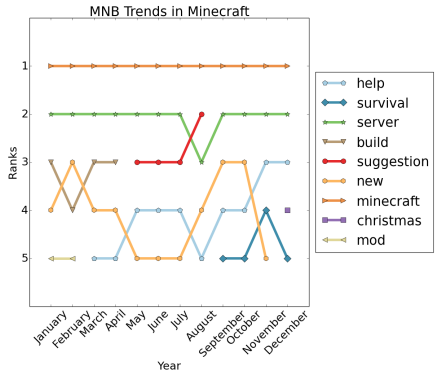
(c)videos (d)worldnews

Figure E.5.: Trends by classification coefficients with binary naive Bayes for r/todayilearned, r/trees, r/videos and r/worldnews

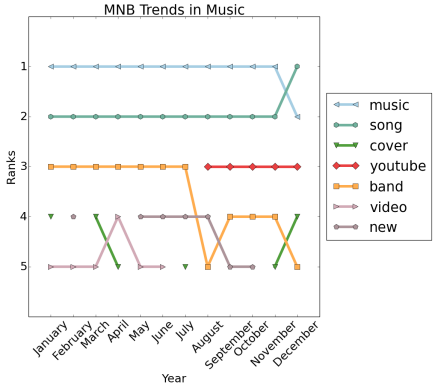


(a)AdviceAnimals

(b)AskReddit

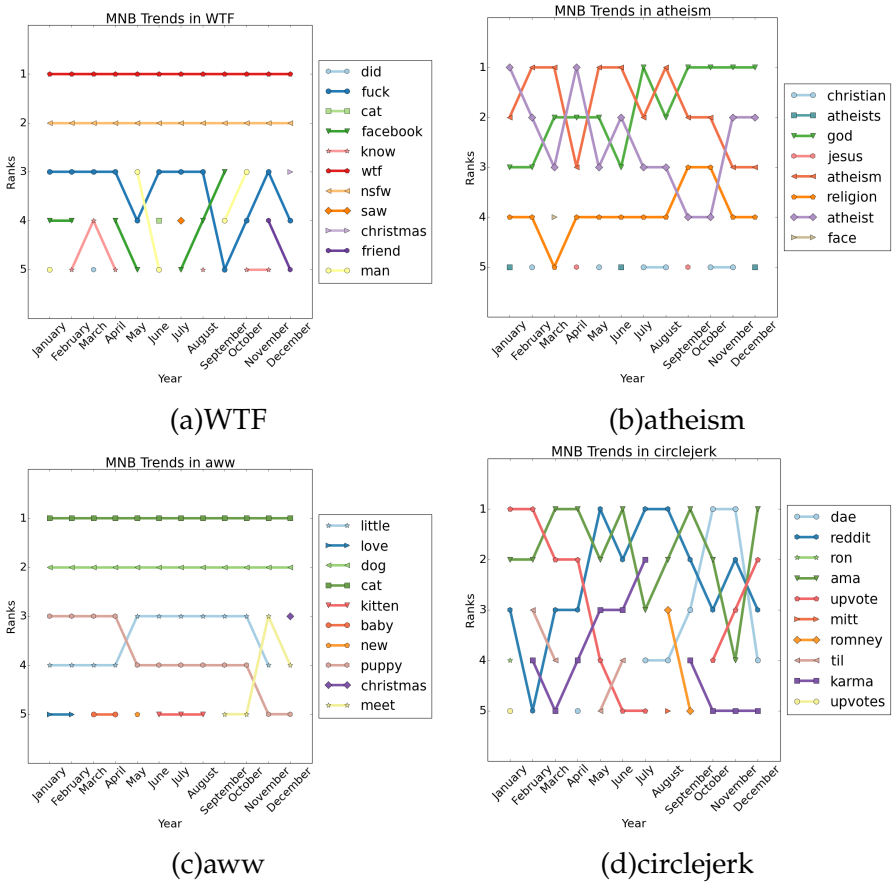


(c)Minecraft



(d)Music

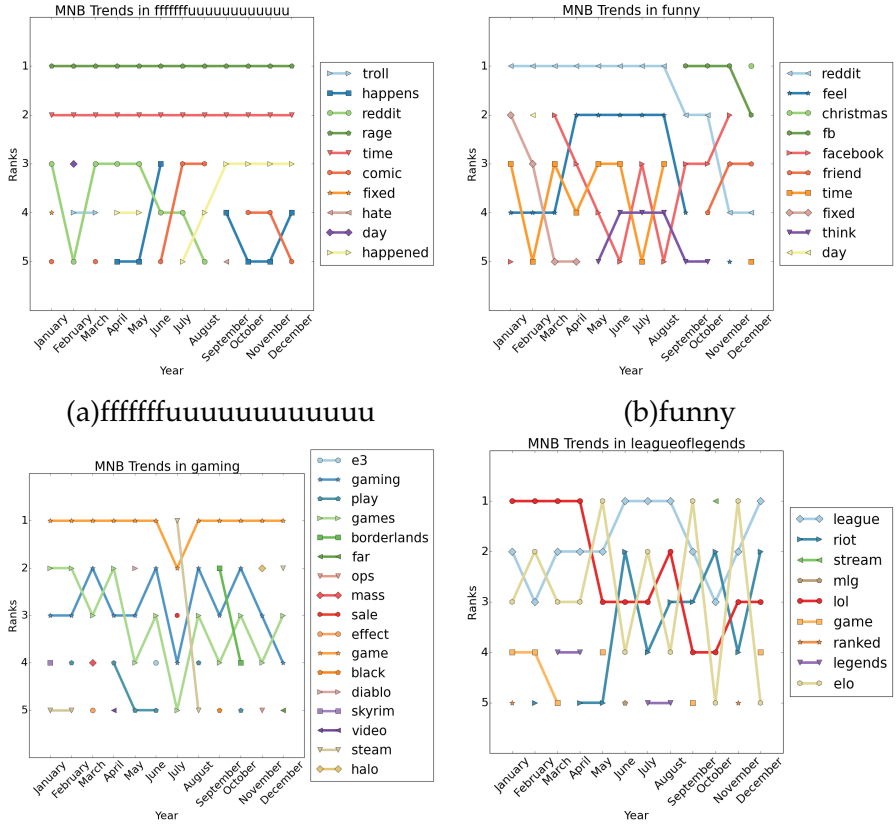
Figure E.6.: Trends by classification coefficients with multinomial naive Bayes for r/AdviceAnimals, r/AskReddit, r/Minecraft and r/Music



(a)WTF (b)atheism

(c)aww (d)circlejerk

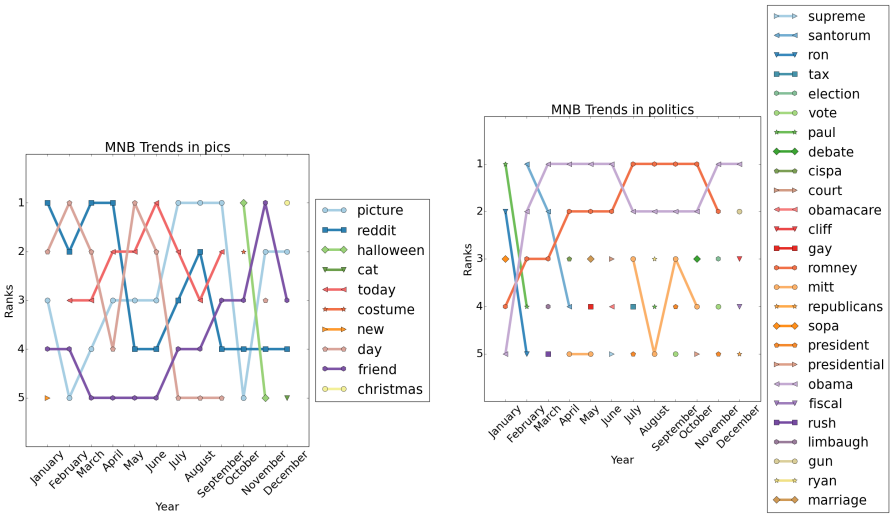
Figure E.7.: Trends by classification coefficients with multinomial naive Bayes for r/WTF, r/atheism, r/aww and r/circlejerk



(a)fffffffuuuuuuuuuuu (b)funny

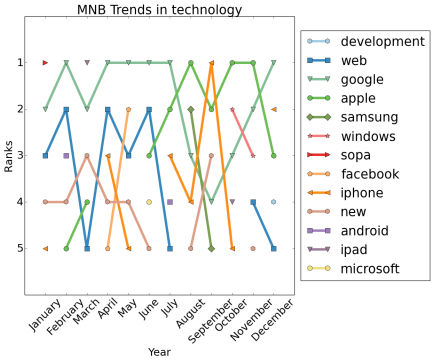
(c)gaming (d)leagueoflegends

Figure E.8.: Trends by classification coefficients with multinomial naive Bayes for r/fffffffuuuuuuuuuuu, r/funny, r/gaming and r/leagueoflegends

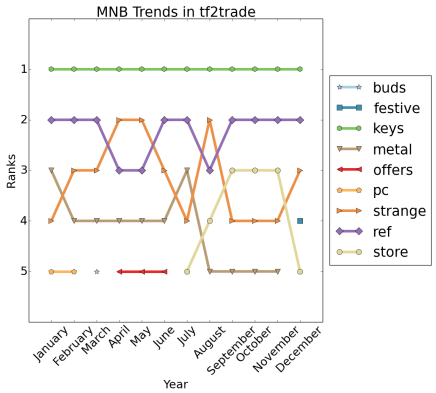


(a)pics

(b)politics

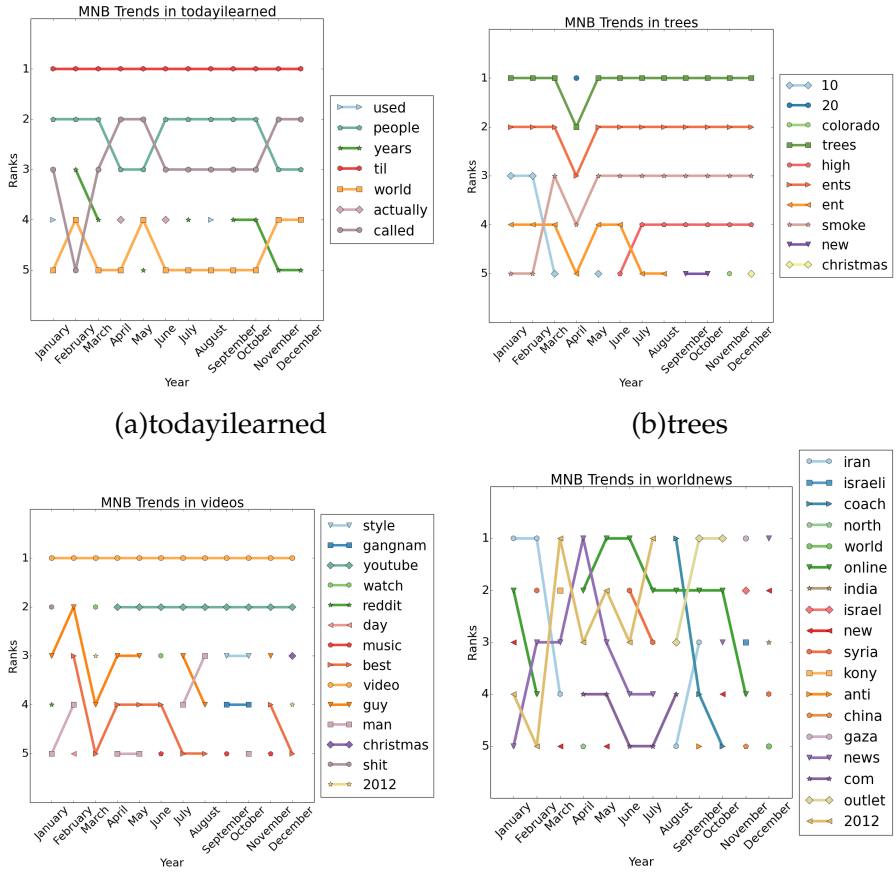


(c)technology



(d)tf2trade

Figure E.9.: Trends by classification coefficients with multinomial naive Bayes for r/pics, r/politics, r/technology and r/tf2trade



(a)todayilearned (b)trees

(c)videos (d)worldnews

Figure E.10.: Trends by classification coefficients with multinomial naive Bayes for r/todayilearned, r/trees, r/videos and r/worldnews

Bibliography

Backstrom, Lars et al. (2011). "Center of Attention: How Facebook Users Allocate Attention across Friends". In: *Fifth International AAAI Conference on Weblogs and Social Media*. Ed. by Lada A. Adamic, Ricardo A. Baeza-Yates, and Scott Counts. AAAI Publications. URL: http://www.aaai.org/ocs/index.php/ICWSM/ICWSM11/paper/view/2899 (cit. on pp. 21, 33).

Baeza-Yates, Ricardo and Berthier Ribeiro-Neto (2011). *Modern Information Retrieval. the concepts and technology behind search*. Pearson Education Limited (cit. on pp. 53, 55–57, 59, 68–71).

Bakshy, Eytan et al. (2011). In: *Proceedings of the fourth ACM international conference on Web search and data mining*. Hong Kong, China, pp. 65–74. ISBN: 978-1-4503-0493-1. DOI: 10.1145/1935826.1935845 (cit. on pp. 21, 32).

Bao, Peng et al. (2013). "Popularity Prediction in Microblogging Network: A Case Study on Sina Weibo". In: *CoRR* abs/1304.4324. URL: http://arxiv.org/abs/1304.4324 (cit. on pp. 22, 36).

Bayes, Thomas (1763). "An essay towards solving a problem in the doctrine of chances". In: *Phil. Trans. of the Royal Soc. of London* 53. DOI: 10.1098/rstl.1763.0053 (cit. on p. 55).

Benevenuto, Fabrício et al. (2009). "Characterizing User Behavior in Online Social Networks". In: *Proceedings of the 9th ACM SIGCOMM Conference on Internet Measurement Conference*. IMC '09. Chicago, Illinois, USA: ACM, pp. 49–62. ISBN: 978-1-60558-771-4. DOI: 10.1145/1644893.1644900 (cit. on pp. 2, 21, 22, 38, 39).

Berners-Lee, Tim (2014). *I am Tim Berners-Lee. I invented the WWW 25 years ago and I am concerned and excited about its future. AMA*. URL: http://redd.it/2091d4 (visited on 03/25/2014) (cit. on p. 116).

Bernstein, Michael S. et al. (2011). "4chan and /b/: An Analysis of Anonymity and Ephemerality in a Large Online Community." In: *ICWSM*. Ed. by Lada A. Adamic, Ricardo A. Baeza-Yates, and Scott Counts. The AAAI Press. URL: http://www.aaai.org/ocs/index.php/ICWSM/ICWSM11/paper/view/2873 (cit. on pp. 22, 35).

Bravais, Auguste (1846). "Analyse Mathematique sur les Probabilites des Erreurs de Situation d'un Point". In: *Memoires par divers Savans* 9, pp. 255–332 (cit. on pp. 60, 64).

Carlsen, Magnus (2014). *Hello Reddit – I'm Magnus Carlsen, the World Chess Champion and the highest rated chess player of all time. AMA.* URL: http://redd.it/20t4pv (visited on 03/25/2014) (cit. on p. 116).

Cha, Meeyoung et al. (2010). "Measuring user influence in Twitter: The million follower fallacy". In: URL: http://citeseer.ist.psu.edu/viewdoc/summary?doi=10.1.1.167.192 (cit. on pp. 21, 32).

Cheng, Justin et al. (2011). "Predicting Reciprocity in Social Networks." In: *Privacy, security, risk and trust (passat), 2011 ieee third international conference on and 2011 ieee third international conference on social computing (socialcom)*. IEEE, pp. 49–56. ISBN: 978-1-4577-1931-8. DOI: 10.1109/PASSAT/SocialCom.2011.110 (cit. on pp. 21, 32).

Cohen, Raviv and Derek Ruths (2013). "Classifying Political Orientation on Twitter: It's Not Easy!" In: *Seventh International AAAI Conference on Weblogs and Social Media*. The AAAI Press. ISBN: 978-1-57735-610-3 (cit. on pp. 21, 31).

Conover, Michael et al. (2011). "Predicting the Political Alignment of Twitter Users." In: *Privacy, security, risk and trust (passat), 2011 ieee third international conference on and 2011 ieee third international conference on social computing (socialcom)*. IEEE, pp. 192–199. ISBN: 978-1-4577-1931-8. DOI: 10.1109/PASSAT/SocialCom.2011.34 (cit. on pp. 21, 31).

Corominas-Murtra, Bernat and Ricard V. Solé (2010). "Universality of Zipf's law". In: *Phys. Rev. E* 82 (1), p. 011102. DOI: 10.1103/PhysRevE.82.011102. URL: http://link.aps.org/doi/10.1103/PhysRevE.82.011102 (cit. on p. 67).

Duan, Yajuan et al. (2010). "An Empirical Study on Learning to Rank of Tweets". In: *Proceedings of the 23rd International Conference on*

Computational Linguistics. COLING '10. Beijing, China: Association for Computational Linguistics, pp. 295–303. URL: http://dl.acm.org/citation.cfm?id=1873781.1873815 (cit. on pp. 21, 30).

Duggan, Maeve and Aaron Smith (2013). *6 % of Online Adults are reddit Users*. Social Networking Report. Pew Internet & American Life Project. URL: http://pewinternet.org/Reports/2013/reddit.aspx (visited on 12/23/2013) (cit. on pp. 21, 40, 98).

Fan, Rui et al. (2013). "Anger is More Influential Than Joy: Sentiment Correlation in Weibo." In: abs/1309.2402 (cit. on pp. 22, 36).

Gates, Bill (2014). *Hello Reddit – I'm Bill Gates, co-chair of the Bill & Melinda Gates Foundation and Microsoft founder. Ask me anything.* URL: http://redd.it/1xj56q (visited on 02/11/2014) (cit. on p. 3).

Gilbert, Eric (2013). "Widespread Underprovision on Reddit". In: *Proceedings of the 2013 Conference on Computer Supported Cooperative Work*. CSCW '13. San Antonio, Texas, USA: ACM, pp. 803–808. ISBN: 978-1-4503-1331-5. DOI: 10.1145/2441776.2441866 (cit. on pp. 21, 40–43, 90).

Gómez, Vicenç, Andreas Kaltenbrunner, and Vicente López (2008). "Statistical Analysis of the Social Network and Discussion Threads in Slashdot". In: *Proceedings of the 17th International Conference on World Wide Web*. WWW '08. Beijing, China: ACM, pp. 645–654. ISBN: 978-1-60558-085-2. DOI: 10.1145/1367497.1367585 (cit. on pp. 22, 28).

Goodman, Steven N. (1999). "Toward Evidence-Based Medical Statistics. 1: The P Value Fallacy". In: *Annals of Internal Medicine* 130.12, pp. 995–1004. DOI: 10.7326/0003-4819-130-12-199906150-00008 (cit. on p. 66).

Hanks, Tom (2013). *Hi reddit, Tom Hanks here. Ask Me Anything.* URL: http://redd.it/1ngfwy (visited on 10/27/2013) (cit. on p. 2).

Hong, Liangjie, Ovidiu Dan, and Brian D. Davison (2011). "Predicting Popular Messages in Twitter". In: *Proceedings of the 20th International Conference Companion on World Wide Web*. WWW '11. Hyderabad, India: ACM, pp. 57–58. ISBN: 978-1-4503-0637-9. DOI: 10.1145/1963192.1963222 (cit. on pp. 21, 31).

Jamali, Salman and Huzefa Rangwala (2009). "Digging Digg: Comment Mining, Popularity Prediction, and Social Network Analysis". In: *Proceedings of the 2009 International Conference on Web Information*

Systems and Mining. WISM '09. IEEE Computer Society, pp. 32–38. ISBN: 978-0-7695-3817-4. DOI: 10.1109/WISM.2009.15 (cit. on pp. 21, 26).

Jensen, Tina Blegind and Signe Dyrby (2013). "Exploring Affordances Of Facebook As A Social Media Platform In Political Campaigning". In: *ECIS 2013 Completed Research*. URL: http://aisel.aisnet.org/ecis2013_cr/40 (cit. on pp. 21, 34).

Kaltenbrunner, Andreas, Vicenc Gomez, and Vicente Lopez (2007). "Description and Prediction of Slashdot Activity". In: *Proceedings of the 2007 Latin American Web Conference*. LA-WEB '07. Washington, DC, USA: IEEE Computer Society, pp. 57–66. ISBN: 0-7695-3008-7. DOI: 10.1109/LA-WEB.2007.59 (cit. on pp. 22, 28).

Kunegis, Jérôme, Andreas Lommatzsch, and Christian Bauckhage (2009). "The Slashdot Zoo: Mining a Social Network with Negative Edges". In: *Proceedings of the 18th International Conference on World Wide Web*. WWW '09. Madrid, Spain: ACM, pp. 741–750. ISBN: 978-1-60558-487-4. DOI: 10.1145/1526709.1526809 (cit. on pp. 22, 28).

Kwak, Haewoon et al. (2010). "What is Twitter, a Social Network or a News Media?" In: *Proceedings of the 19th International Conference on World Wide Web*. WWW '10. Raleigh, North Carolina, USA: ACM, pp. 591–600. ISBN: 978-1-60558-799-8. DOI: 10.1145/1772690.1772751 (cit. on pp. 21, 29, 30).

Lakkaraju, Himabindu, Julian J. McAuley, and Jure Leskovec (2013). "What's in a Name? Understanding the Interplay between Titles, Content, and Communities in Social Media." In: *ICWSM*. Ed. by Emre Kiciman et al. The AAAI Press. ISBN: 978-1-57735-610-3. URL: http://www.aaai.org/ocs/index.php/ICWSM/ICWSM13/paper/view/6085 (cit. on pp. 21, 41).

Lerman, Kristina (2006). "Social Networks and Social Information Filtering on Digg". In: *CoRR* abs/cs/0612046. URL: http://arxiv.org/abs/cs/0612046 (cit. on pp. 21, 23–25, 27).

Lerman, Kristina (2007). "Dynamics of Collaborative Document Rating Systems". In: *Proceedings of the 9th WebKDD and 1st SNA-KDD 2007 Workshop on Web Mining and Social Network Analysis*. WebKDD/SNA-KDD '07. San Jose, California: ACM, pp. 46–55.

ISBN: 978-1-59593-848-0. DOI: 10.1145/1348549.1348555 (cit. on pp. 21, 26).

Lerman, Kristina and Aram Galstyan (2008). "Analysis of Social Voting Patterns on Digg". In: *CoRR* abs/0806.1918 (cit. on pp. 21, 25).

Leskovec, Jure, Lada A. Adamic, and Bernardo A. Huberman (2007). "The Dynamics of Viral Marketing". In: *ACM Trans. Web* 1.1. ISSN: 1559-1131. DOI: 10.1145/1232722.1232727 (cit. on p. 3).

Luhn, Hans P. (1957). "A Statistical Approach to Mechanized Encoding and Searching of Literary Information". In: *IBM Journal of Research and Development* 1.4, pp. 309–317. ISSN: 0018-8646. DOI: 10.1147/rd.14.0309 (cit. on p. 69).

Mitchell, Tom (1997). *Machine Learning*. McGraw Hill. ISBN: 0070428077 (cit. on pp. 45, 46).

MrGrim (2009). *My Gift to Reddit: I created an image hosting service that doesn't suck. What do you think?* URL: http://redd.it/7zlyd (visited on 03/10/2014) (cit. on p. 87).

Ng, Andrew (2014). *Machine Learning*. URL: https://class.coursera.org/ml-003/lecture/index (visited on 03/02/2014) (cit. on pp. 46–48).

Nonnecke, Blair and Jennifer Preece (2000). "Lurker demographics: counting the silent." In: *CHI*. Ed. by Thea Turner and Gerd Szwillus. ACM, pp. 73–80. ISBN: 1-58113-216-6 (cit. on pp. 39, 90).

Obama, Barack (2012). *I am Barack Obama, President of the United States – AMA*. URL: http://redd.it/z1c9z (visited on 10/27/2013) (cit. on pp. 2, 120).

Olson, Randal S. (2013a). *A data-driven guide to creating successful Reddit posts.* URL: http://dx.doi.org/10.6084/m9.figshare.652965 (visited on 02/12/2014) (cit. on pp. 42, 98).

Olson, Randal S. (2013b). *Retracing the evolution of Reddit through post data.* URL: http://dx.doi.org/10.6084/m9.figshare.650851 (visited on 02/12/2014) (cit. on p. 42).

Page, Lawrence et al. (1999). *The PageRank Citation Ranking: Bringing Order to the Web.* Technical Report 1999-66. Stanford InfoLab (cit. on p. 30).

Pak, Alexander and Patrick Paroubek (2010). "Twitter as a Corpus for Sentiment Analysis and Opinion Mining". In: *Proceedings of the*

Seventh International Conference on Language Resources and Evaluation (LREC'10). Ed. by Nicoletta Calzolari et al. European Language Resources Association. ISBN: 2-9517408-6-7. URL: http://www.lrec-conf.org/proceedings/lrec2010/summaries/385.html (cit. on pp. 21, 32).

Pearson, Karl (1896). "Mathematical Contributions to the Theory of Evolution. III. Regression, Heredity and Panmixia". In: *Philosophical Transactions of the Royal Society of London* 187, pp. 253–318. DOI: 10.1098/rsta.1896.0007 (cit. on pp. 60, 64).

Rao, Delip et al. (2010). "Classifying latent user attributes in Twitter". In: *In Proc. of SMUC*. URL: http://citeseerx.ist.psu.edu/viewdoc/summary?doi=10.1.1.208.5011 (cit. on pp. 21, 31).

reddit, team (2008). *reddit.com's code*. URL: https://github.com/reddit (visited on 10/27/2013) (cit. on p. 17).

reddit, team (2010). *reddit gold about page*. URL: http://reddit.com/gold/about (visited on 12/31/2013) (cit. on p. 17).

reddit, team (2013). *reddit.com: about reddit*. URL: http://reddit.com/about (visited on 10/27/2013) (cit. on pp. 1, 98).

Reeves, Keanu (2013). *Keanu Reeves. Ask me, if you want, almost anything.* URL: http://redd.it/1ouqge (visited on 10/19/2013) (cit. on p. 2).

Rowe, Matthew, Sofia Angeletou, and Harith Alani (2011). "Predicting Discussions on the Social Semantic Web". In: *The Semanic Web: Research and Applications*. Ed. by Grigoris Antoniou et al. Vol. 6644. Lecture Notes in Computer Science. Springer Berlin Heidelberg, pp. 405–420. ISBN: 978-3-642-21063-1. DOI: 10.1007/978-3-642-21064-8_28 (cit. on pp. 21, 32).

Salihefendic, Amir (2010). *How Reddit ranking algorithms work*. URL: http://amix.dk/blog/post/19588 (visited on 11/05/2013) (cit. on pp. 4, 9).

Salton, Gerard and Chu-Sing Yang (1973). "On the specification of term values in automatic indexing". In: *Journal of Documentation* 29 (cit. on p. 71).

Samuel, Arthur L. (1959). "Some Studies in Machine Learning Using the Game of Checkers". In: *IBM J. Res. Dev.* 3.3, pp. 210–229. ISSN: 0018-8646. DOI: 10.1147/rd.33.0210. URL: http://dx.doi.org/10.1147/rd.33.0210 (cit. on pp. 45, 46).

Schöfegger, Karin et al. (2012). "Learning User Characteristics from Social Tagging Behavior". In: *Proceedings of the 23rd ACM Conference on Hypertext and Social Media*. HT '12. Milwaukee, Wisconsin, USA: ACM, pp. 207–212. ISBN: 978-1-4503-1335-3. DOI: 10.1145/2309996.2310031 (cit. on pp. 22, 38).

Shannon, Claude E. (1948). "A Mathematical Theory of Communication". In: *Bell System Technical Journal* 27, pp. 379–423. DOI: 10.1002/j.1538-7305.1948.tb01338.x (cit. on p. 66).

Singer, Philipp (2013). *One year of Reddit submissions: Redditors tend to post more during the week*. URL: http://philippsinger.info/?p=161 (visited on 03/19/2014) (cit. on p. 98).

Singer, Philipp et al. (2014). "Evolution of Reddit: From the Front Page of the Internet to a Self-referential Community?" In: *WWW'14: Proceedings of the 19th International World Wide Web Conference*. Seoul, South Korea: ACM. URL: http://arxiv.org/abs/1402.1386 (cit. on pp. 7, 16, 155).

Sparck Jones, Karen (1988). "Document Retrieval Systems". In: ed. by Peter Willett, pp. 132–142. URL: http://dl.acm.org/citation.cfm?id=106765.106782 (cit. on p. 70).

Spearman, Charles (1904). "The Proof and Measurement of Association Between Two Things". In: *American Journal of Psychology* 100, pp. 441–471 (cit. on p. 64).

Spiliotopoulos, Tasos and Ian Oakley (2013). "Understanding Motivations for Facebook Use: Usage Metrics, Network Structure, and Privacy". In: CHI '13, pp. 3287–3296. DOI: 10.1145/2470654.2466449 (cit. on pp. 21, 33).

Szabó, Gábor and B. A. Huberman (2008). "Predicting the popularity of online content". In: *CoRR* abs/0811.0405. URL: http://arxiv.org/abs/0811.0405 (cit. on pp. 21, 23, 24).

Tang, Siyu et al. (2011). "Digging in the Digg Social News Website." In: *IEEE Transactions on Multimedia* 13.5, pp. 1163–1175. DOI: 10.1109/TMM.2011.2159706 (cit. on pp. 21, 25).

Tunkelang, Daniel (2009). *A Twitter Analog to Pagerank*. URL: http://thenoisychannel.com/2009/01/13/a-twitter-analog-to-pagerank (visited on 01/14/2014) (cit. on p. 31).

/u/minimaxir (2013). *Heatmap of all link submissions to Reddit which get a score >3000, by hour and day-of-week of submission [OC]*. URL: http://redd.it/1pe4vm (visited on 11/01/2013) (cit. on p. 99).

Van Mieghem, Piet (2011). "Human Psychology of Common Appraisal: The Reddit Score". In: *Multimedia, IEEE Transactions on* 13.6, pp. 1404–1406. ISSN: 1520-9210. DOI: 10.1109/TMM.2011.2165054 (cit. on pp. 21, 41).

Wagner, Claudia, Matthew Rowe, et al. (2012). "Ignorance Isn't Bliss: An Empirical Analysis of Attention Patterns in Online Communities". In: *Privacy, Security, Risk and Trust (PASSAT), 2012 International Conference on and 2012 International Confernece on Social Computing (SocialCom)*. Amsterdam, The Netherlands, pp. 101–110. DOI: 10.1109/SocialCom-PASSAT.2012.33 (cit. on pp. 2, 22, 39).

Wagner, Claudia, Philipp Singer, et al. (2013). "The Wisdom of the Audience: An Empirical Study of Social Semantics in Twitter Streams". In: *ESWC*. Ed. by Philipp Cimiano et al. Vol. 7882. Lecture Notes in Computer Science. Springer Berlin Heidelberg, pp. 502–516. ISBN: 978-3-642-38288-8. DOI: 10.1007/978-3-642-38288-8_34 (cit. on pp. 21, 31).

Weninger, Tim, Xihao Avi Zhu, and Jiawei Han (2013). "An Exploration of Discussion Threads in Social News Sites: A Case Study of the Reddit Community". Niagara, Ontario, Canada, pp. 579–583. ISBN: 978-1-4503-2240-9. DOI: 10.1145/2492517.2492646 (cit. on pp. 21, 43).

Wu, Fang and Bernardo A. Huberman (2007). "Novelty and collective attention". In: 104.45, pp. 17599–17601. DOI: 10.1073/pnas.0704916104. eprint: http://www.pnas.org/cgi/reprint/104/45/17599.pdf (cit. on pp. 21, 26).

Zhu, Yingwu (2009). "Measurement and Analysis of an Online Content Voting Network: A Case Study of Digg". In: *CoRR* abs/0909.2706. URL: http://arxiv.org/abs/0909.2706 (cit. on pp. 21, 27).