IT Systems Engineering
Universität Potsdam

Master's Thesis

Analyzing and Predicting Viral Tweets

Maximilian Jenders

Maximilian.Jenders@student.hpi.uni-potsdam.de

Submitted: 02.10.2012

Supervision: Dr. Gjergji Kasneci

Abstract

Information on the Internet is being generated and shared not only on traditional, static websites, but also increasingly through blogs and in social networks. Twitter and other social microblogging services have become indispensable sources of information in today's web. Understanding the main factors that make certain pieces of information spread quickly in these platforms can be decisive for the analysis of opinion formation and many other opinion mining tasks.

This thesis addresses important questions concerning the spread of information. All research conducted here is based on Twitter as a representation of such microblogging networks. What makes Twitter users retweet a tweet? How do tweets spread in the social network over time? And most importantly, is it possible to predict whether a tweet will become "viral", i.e., will be frequently retweeted by other Twitter users?

In order to answer these questions we provide an extensive analysis of a wide range of tweet- and user-specific features regarding their influence on the spread of tweets. The most impactful features are chosen to build a learning model that predicts viral tweets with high accuracy. This model is shown to outperform a baseline Naive Bayes model. Further, a method to estimate the expected number of retweets is pointed out and evaluated. All experiments are performed on a real-world dataset that was extracted through a public Twitter API based on user IDs from the TREC 2011 microblog corpus.

Zusammenfassung

Informationen werden im Internet zunehmend nicht nur auf tradionellen, statischen Webseiten, sondern auch auf Blogs und in sozialen Netzwerken erstellt. Twitter und andere Mikroblogdienste haben sich als unverzichtbare Informationsquellen im heutigen Web etabliert. Ein gutes Verständnis der Faktoren, die eine Auswirkung auf die schnelle Verbreitung bestimmter Informationen innerhalb dieser Platformen bewirken, kann für eine Analyse von Meinungsbildungs- und -erhebungsaufgaben entscheidend sein.

Diese Masterarbeit thematisiert wichtige Fragen der Ausbreitung von Informationen auf Twitter. Was bewegt Twitternutzer dazu, einen Tweet zu retweeten? Wie verbreiten sich Tweets in diesem sozialen Netzwerk im Laufe der Zeit? Und - am allerwichtigsten - ist es möglich vorherzusagen, ob ein Tweet "viral"wird, d.h. häufig von anderen Nutzern retweetet werden wird?

Um diese Fragen zu beantworten analysieren wir eine große Auswahl von tweet- und nutzerspezifischen Merkmalen in Hinsicht auf ihren Einflusses auf die Verbeitung von Tweets. Die einflussreichsten Merkmale werden ausgewählt, um ein Lernmodell zu erstellen, das virale Tweets mit hoher Genauigkeit vorhersagt. Das Modell übertrifft ein Basismodell, das auf Naive Bayes beruht. Weiterhin wird eine Methode aufgezeigt, die voraussichtliche Anzahl an Retweets zu schätzen. Alle Experimente werden auf realen Daten ausgeführt, die durch eine öffentliche Twitter-API extrahiert wurden und auf Nutzer-IDs des TREC 2011 microblog corpus basieren.

Acknowledgement

This thesis would not have been possible without the support of a many people. to thank everyone who helped me on the way.

I would like to thank Prof. Dr. Felix Naumann and the staff of the Information Systems chair at Hasso-Plattner Institute. In particular, I would like to express my gratitude to my supervisor, Dr. Gjergji Kasneci, who was abundantly helpful and always spared time to offer invaluable assistance and guidance.
Furthermore, I would like to thank my family, especially my parents, for their continuous support during my whole studies.
Special thanks also to all those who proofread my thesis and provided me with valuable and constructive feedback.

Declaration

I hereby declare in lieu of oath that I composed this thesis independently and without inadmissible help from outside. The sources used are quoted in full and parts that are direct quotations or paraphrases are identified as such.

Potsdam, October 02, 2012

Maximilian Jenders

Contents

1 Introduction

In today's Web 2.0, user-generated content gains evermore importance. Information is no longer purely generated on a small quantity of static, popular and oftentimes corporate websites that mostly have just one author, but increasingly by many autonomous users on a vast multitude of websites. This shift provides users with the means to publish their own content instead of being mere consumers of information. While longer and more detailed information is typically published on blogs[1] and regular websites, short messages and statements are increasingly shared on social networks, such as Twitter[2] Google+[3], Facebook[4], and LinkedIn[5], which can boast of hundreds of millions of users. Such networks provide microblogging services, enabling users to disseminate posts that usually consist of short text messages with less than 200 characters. Analogously to blogs, those posts are usually displayed chronologically reversed. One very important feature of these kinds of networks is that they allow users to share information with their online friends, who in turn can also re-share the information with their acquaintances on the network. This enables a rapid spread of information from a few users to millions of people around the globe.

Motivation and Goal

Since these mechanisms can connect users with big audiences, extending the reach of their message and thereby providing them with influence, it is not surprising that a multitude of corporations and institutions already closely monitor and exploit such social networks. Detecting the latest news and public opinions on various issues, e.g., companies, products, events, socio-political matters, celebrities, or even being able to shift public attention towards these matters can help in solving many issues on one hand and improving reputation on the other. This holds for corporations, celebrities or institutions alike.

[1] A blog, short for web log, is a website used for discussion with or information of other users and consist of seperate "posts" (i.e., articles) that are displayed in reversed chronological order.
[2] `https://twitter.com/`
[3] `https://plus.google.com/`
[4] `https://facebook.com/`
[5] `http://linkedin.com/`

Understanding the main factors that make a certain message *viral*, i.e., spread quickly and widely within and through these networks, can therefore be crucial in the analysis of public opinion formation as well as many other opinion mining tasks. Any insights gained in these fields can be harnessed to deliver better prediction tools to advertisers, opinion leaders and decision makers. With accurate prediction tools, a company could, for example, gather and evaluate public opinion on beta versions of a product before manufacturing the product. Additionally, the product could be promoted more precisely and effectively if the relevant audience was identified.

Already, predictions can be made based on data of social networks. For example, [AH10] harnessed Twitter posts to forecast box-office revenues for movies, outperforming market-based predictors. Furthermore, social media might prove to be a great tool for behavorial researchers. In common research (e.g., in psychology, sociology or psychometric sciences), humans are observed while answering questions or performing tasks. However, the test subjects know that they are being observed and a researcher or recording device is almost always present. This alone might prompt the subjects to subconsciously change their behavior respectively the way they express their opinions. They might even behave differently to maintain a better impression. In contrast, posts on the Internet can be monitored without any interaction or even knowledge of the user that published them, presenting a great opportunity for non-intrusive analyses of how humans shape and express their opinions. Posts on social networks are often shared with the general public, enabling access to a vast pool of opinions, statements, and topics of interest that can be further broken down into categories according to research needs.

This thesis addresses important aspects of public opinion formation by analyzing and highlighting different factors that influence how rapidly and widely information spreads in social networks. The research conducted in this thesis focuses on Twitter, which we chose as a representative and very popular[6] microblogging service. Any insights gained might analogously hold for data from other social or even news networks, as Twitter presents both the characteristics of both social and news networks [KLPM10].

On Twitter, the messages users publish to the general public are called *tweets* and have a restricted maximum length of 140 characters each. Tweets are typically either status updates, detailing feelings and experiences, personal comments on recent news or specific topics, or so-called presence updates, i.e., statuses showing a user's current location accompanied by a personal comment.

While users can browse any tweets, the tweets of a user are usually only viewed by that user's *followers*, i.e., all those who selected to subscribe to that user's tweets.

[6] Twitter is the world's eighth popular website, according to Alexa as of September 26nd, 2012, `http://www.alexa.com/topsites/`

Note that users do not necessarily need to follow their followers, resulting in a directed relationship graph. A user can share another user's tweet by *retweeting*, i.e., reposting, it, even if no follower relationship exists between both users. Hence, the popularity or impact of a tweet can be measured by its retweet frequency, i.e., the number of retweets it receives. Despite the restricted length of tweets, Twitter's retweet mechanism and its social network, based on the "follower" relationships, provide an unprecedented mechanism for the spread of information in the form of tweets.

In this thesis, this spread mechanism is analyzed and the following questions are addressed:

- What are the main factors that make a tweet "viral", so that it is read and retweeted by many Twitter users? To answer this question we analyze different tweet features with respect to their impact on the retweet frequency. Analogously, we analyze the features of the user from whom the tweet originated. These analyses are conducted with the goal of finding out which of these features are the most important ones for the retweet frequency and how they influence and correlate with each other.

- How do viral tweets spread over time? For tweets with different numbers of retweets, we investigate spread patterns according to which of these tweets are retweeted over time.

- Is it possible to predict viral tweets? We answer this question affirmatively and provide a generalized linear model that can predict viral tweets with high accuracy and outperforms a generative baseline model that assumes conditional independence between features. Furthermore, we extend the generalized linear model to approximate the expected number of retweets. All models are extensively evaluated on a real-world dataset that was extracted via a public Twitter API and based on Twitter user IDs from the TREC 2011 microblog corpus.

Chapter Overview

This thesis is organized as follows: Chapter 2 gives an overview of work related to this thesis. Different structural, contextual, and sentimental aspects of related research are pointed out and discussed. Next, the framework and process for the extraction of the dataset on which all subsequent analyses are performed on is described in Chapter 3. User selection and the data retrieved from Twitter are elaborated on as well as the exact definition of viralilty for the purposes of this thesis. In Chapter 4,

we provide a broad analysis of different tweet- and user-specific features and discuss their impact on the number of retweets that tweets receive. The analyzed features encompass, for example, the number of followers a user has, the length of a tweet, and the sentiment used in tweets. Chapter 5 presents a generalized linear model that predicts viral tweets by exploiting the analyzed features. A Naive Bayes model is introduced as a baseline. An extensive evaluation of the models is carried out in Chapter 6. Furthermore, the prediction model is harnessed to create predictions on the expected number of retweets a tweet is likely to receive in Chapter 7. Finally, the work is discussed and future work is pointed out in Chapter 8, and a conclusion is given in Chapter 9.

2 Related Work

Prior work has already gained many insights on social networks in general as well as on Twitter in particular. Research concerning the spread of information on Twitter can be categorized into three broader topics: structural, contextual, and sentimental. We address these topics in the following paragraphs.

2.1 Structural Analysis

Much of the prior work has investigated structural properties of social networks to predict influential users. Link-analysis techniques based on the number of followers were used to derive influence scores, e.g., by exploiting PageRank [KLPM10], by predicting the propagation of such scores through the network, e.g., [KKT03] or by studying information diffusion networks on Twitter through "@username" mentions [YC10].

[HRW08] studied the propagation of messages on Twitter, taking into account social interactions between users. The authors found that the driver of message propagation was not the entire follower network, but a sparse and hidden but influential subset of the network containing the users that often interact with each other.

[KKT03] explored link-analysis algorithms for identifying subsets of social network users who should be addressed in order to trigger a quick and wide spread of information. In [KLPM10], it was shown that the PageRank score of a user can be estimated by the number of followers the user has. However, a ranking of users by their PageRank scores yielded a different order than a ranking by the total number of retweets that the tweets of the users accumulate. This strongly indicates that link and retweet analysis capture user influence in different ways.

Similarly, [CHBG10] analyzed the influence of users by measuring three different parameters: the number of followers, the number of retweets, and the number of mentions. The authors found that most of the influential users hold influence over various topics. Furthermore, the authors observed that the number of followers a user has represents his popularity, but that it is not connected to the user's ability of

generating retweets. The explanation offered is that many users, once being followed by another user, follow that user back out of politeness, but are not interested in the second user's tweets. The number of followers, therefore, should not be as important as the actual amount of users that are actively retweeting tweets or are engaged in discussions. While this explanation seems logical, our findings contradict it, showing that the amount of followers is the most impactful indicator of the expected retweet frequency for the tweets of a user.

[BHMW11] reported that the most influential users on Twitter are those who have already been most influential in the past and who have a large number of followers. Yet, methods that try to predict which specific user will generate influential tweets (i.e., viral tweets) are relatively unreliable. This suggests that a more detailed analysis of tweets is needed to complement structural measures.

[ZHVGS10] used a probabilistic collaborative filtering model to predict whether a specific Twitter user would retweet another user's tweet. The model used user features, such as name and number of followers and showed a degraded performance when the words of the message were additionally used as tweet features. The focus on pairwise predictions (i.e., predicting whether just one specific user will retweet a specific tweet) makes this approach unfeasible for large-scale application.

2.2 Content Analysis

Another stream of research, such as [BGL10], [SHPC10], [ABL12], and [POL11] has analyzed the influence of tweets by examining their content.

[BGL10] investigated the question of why and how people retweet. Various styles of retweets are highlighted. Specifically, the authors point out the use of mentions in retweets as means to attribute the retweet to the author of the original tweet.

[SHPC10] reported that the use of URLs and hashtags in a tweet affects the total number of retweets that the tweet incurs. The authors of [ABL12] pointed out that the typical user reaction to tweets is the retweet of tweets or the unfollowing of users who posted those tweets. The authors further analyzed the details of user reactions by setting up a website that allowed users to voluntarily rate tweets. The study founds that users tolerate large amounts of less-desired content in their Twitter feeds before unfollowing, and that users prefer tweets sharing information or random thoughts to updates about the current location or mood of the user.

[POL11] employed structural and content features in a machine learning algorithm to predict whether a tweet will be retweeted. The results were compared with a baseline

where human judges were given two tweets and then asked to predict which of the two will receive be retweeted. However, no analysis of the used features is given, and only few feature weights were presented. Furthermore, the human prediction baseline consisted of insufficient data with just two humans making predictions about 202 tweets.

2.3 Sentiment Analysis

Other work has focused on the sentiment analysis of tweets or other social media, most prominently [AH10], [GGSS12], [HAN+11], [TBP+10], and [PGS12].

[AH10] built a model to make predictions about movie box office revenues based on the content of tweets. A sentiment analysis on movie comments could further increase the predictive power.

[GGS11] examined the emotion of words in the English, German, and Spanish language. The authors discovered that the frequency of word use is not only influenced by the length of words and their average information content, but also by the emotional content of words. Words with a positive emotional content are being used more often, yet negative words contain more information than positive ones.

[BMP11] analyzed six different mood states derived from tweets and related them to records of popular events, such as socio-political, cultural, and economic events. The authors reported that such events have immediate effect on the sentiment expressed in tweets. This finding is supported by [LWLC12], which showed that periodic events such as Christmas and Halloween evoke similar mood patterns every year. Furthermore, significant increases in negative mood indicators coincide with announcements of public spending cuts by the government.

Over the course of one month, [TBP11] analyzed the sentiment in the 30 largest events discussed in Twitter post, finding that on Twitter, important events go along with increases in average negative sentiment. Furthermore, [GGSS12] analyzed emotional expression of users, as derived from their messages, finding that the emotional expression of individual users persists over a long period of time. Finally, [HAN+11] trained a Naive Bayes classifier to detect news in tweets and showed that negative sentiment enhances the virality of news-related tweets, but not of the non-news tweets.

In an effort to automatically classify sentiments in social networks, the authors of [TBP+10] designed *SentiStrength*[1], an algorithm for extracting sentiment strength

[1] `http://sentistrength.wlv.ac.uk/`

from informal English text. The algorithm exploits the grammar and spelling styles in typical microblogs and builds on human-evaluated dictionaries for words connotated with positive or negative sentiments. SentiStrength was tested on MySpace[2] comments, revealing an impressive accuracy concerning the identification of sentiments. An improved version of the algorithm in [TBP12] showed that SentiStrength is robust enough to be applied to a wide variety of different social web contexts.

In [PGS12], the SentiStrength algorithm was used to infer sentiments in tweets. The authors analyzed how the inferred sentiments relate to the retweet probability of a tweet. More specifically, for each tweet, a positive and negative sentiment score was derived by means of the SentiStrength algorithm. According to these scores, tweets were classified according to their overall valence as positive, negative, and neutral tweets. For each of these categories, the retweet distribution was analyzed. Interestingly, the authors reported that the fraction of tweets is similar to that of retweets in each category. This report is contradicted by our analysis. We find that the fraction of retweets is significantly higher than that of tweets for the negative category, implying that tweets with negative sentiments are much more likely to be retweeted. Another interesting notion introduced by [PGS12] is that of *emotional divergence*, which combines the positive and negative score of a tweet to compute the overall emotional strength of the sentiment in a tweet. This score can be used to make predictions about the probability of a tweet being retweeted. We investigate the relation between emotional divergence and retweet probability as well and are able to confirm the findings of [PGS12].

In this thesis, we perform analyses on a multitude of tweet- and user-specific features. Structurual and content-based aspects are taken into account as well as sentimental aspects. All analyses are discussed in detail to reach an extensive understanding of if and to what extent the features influence the number of retweets that tweets receive. The insights are then applied to build a model that accurately predicts whether tweets will become viral. The thesis distinguishes itself from prior work by the depth and breadth of the analyses and considered features.

[2] `https://www.myspace.com/`

3 Data Extraction

The first and most important step of the data extraction was to choose the Twitter users which were to be crawled. All subsequent analyses are based on the tweet and user information extracted from those selected users. Therefore, we carefully paid attention to avoid any bias during the selection process, which is detailed in Section 3.3.

As part of the TREC 2011 microblog track[1], a representative sample of the so-called *Twittersphere* was published under the name "Tweets2011 corpus"[2], containing identifiers for approximately 16 million tweets and their corresponding users that were collected between January 23rd and February 8th, 2011. The initiative was led by the National Institute of Standards and Technology (NIST).

We decided to harness the above TREC dataset to select the users whose information we planned to analyze. Unfortunately, the Twitter Terms of Service prohibit the distribution of tweet and user data, permitting only user and tweet identifiers to be shared. Therefore, we relied on the public Twitter REST API[3] to retrieve the data based on the identifiers provided by the above corpus.

3.1 Twitter API

For the work executed for this thesis, we used the Twitter4J library[4] to handle all API calls. The results returned by Twitter were subsequently processed and then written into a database. It should be noted that Twitter imposes an hourly rate limit of 350 request to the REST API[5]. Twitter provides different request types for retrieving different types of information. For example, one request type allows retrieving information about up to 100 users specified by their ID; another one allows

[1] `https://sites.google.com/site/microblogtrack/`
[2] `http://trec.nist.gov/data/tweets/`
[3] `https://dev.twitter.com/docs/api/`
[4] `http://twitter4j.org/en/index.html`
[5] A seperate rate limit applies to each authenticated user. At the time of this writing, Twitter announced changes to the REST API and the rate limit, see `https://dev.twitter.com/blog/changes-coming-to-twitter-api/`.

fetching 200 of the last 3,200 tweets of a user, and yet another one can be used to obtain 100 identifiers of a user's followers.

Through a sequence of requests using the above API, the following information about users and tweets are accessible to be crawled:

- for users:
  - the list of their followers,
  - the list of the users they follow,
  - the total number of tweets,
  - their last 3200 tweets;
- for tweets:
  - the actual message,
  - the publishing date of the tweet,
  - the number of retweets it has obtained so far,
  - whether the tweet itself is a retweet.

3.2 Data Extraction and Analysis Framework

The software architecture was designed with special focus on having independent, decoupled classes, each serving their own, specific purpose. The resulting architecture is depicted in Figure 3.1 and consists of three main parts: *DatabaseConnection*, *Crawler* and *Analyzer*.

The *DatabaseConnection* encapsulates all access to a DB2 database and holds various SQL statements to store, update or retrieve data. When necessary, *Data Storage Classes* are used to hold the data returned by SQL queries so they can be more easily processed. For example, the *DBStatus* class can be used to instantiate any data from the table in which tweets are stored. Since the machine on which crawling and analyses were performed has multiple processors, efforts were made to allow multiple program instances to operate on the data simultaneously. The parallel execution of multiple program instances on different parts of the data leads to a

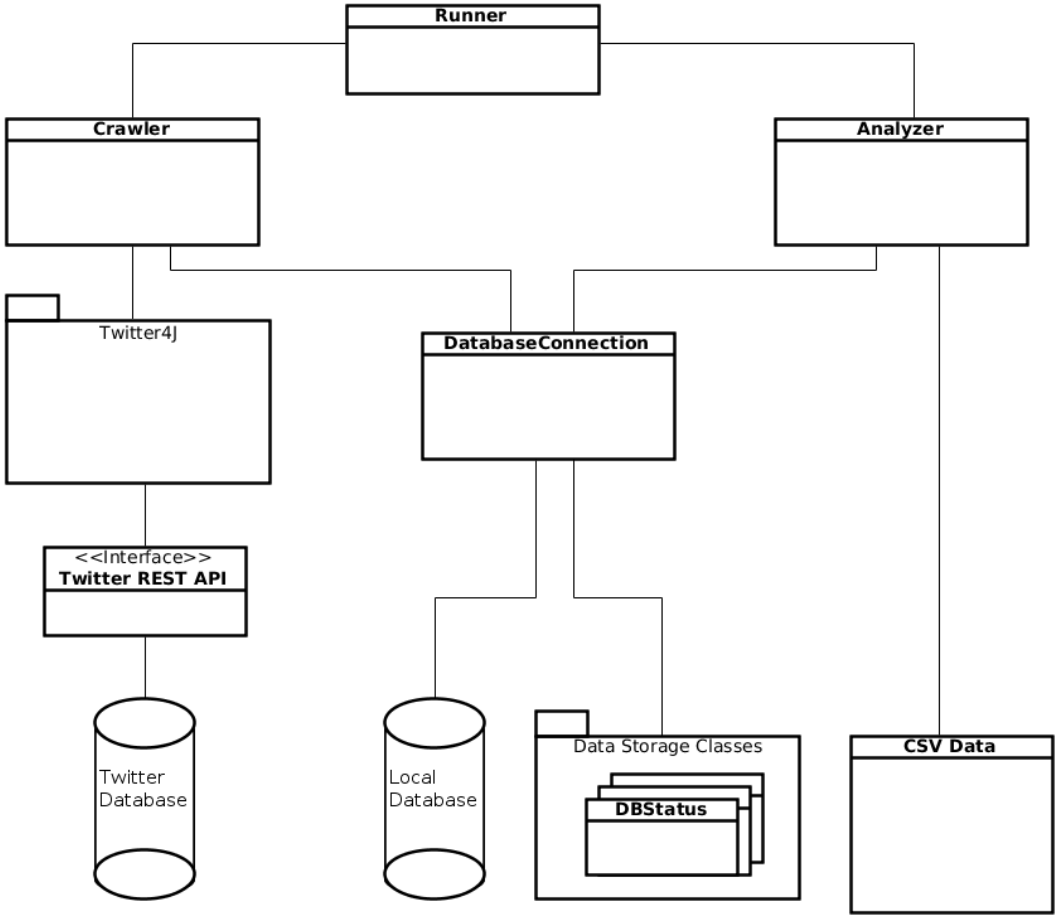

Figure 3.1: Data extraction and analysis framework.

significant reduction of the execution times for many post-crawling processing and analysis jobs.

The *Crawler* relies on the *Twitter4J* library to fetch tweet- and user-specific data from Twitter through the *Twitter REST API*. The *Crawler* fetches the ID of the user that is to be crawled next[6] through the *DatabaseConnection* and then downloads all accessible tweets of that user with several sequential calls to the *Twitter REST API*. When a user's tweets are completely downloaded, they are processed in preparation for analyses (e.g., the tweet's sentiment and hashtags are determined) and subsequently passed to the *DatabaseConnection* for storage in the database. Finally, the timestamp of the latest crawl is updated for the user and the next user is selected.

[6]The next user to be crawled either is a new user that has not been crawled yet or the user that has not been crawled for longest time.

If the rate limit is reached, the *Crawler* pauses until new API calls are permitted. Also, HTTP errors are handled by a short waiting period followed by a repetition of the failed API call.

All analytical procedures are handled by the *Analyzer*. The typical function is structured in the following way: a function loads all prerequisite data through the *DatabaseConnection* and then iteratively analyzes the data. The results are then either written to a *CSV* file or into the database, sometimes updating old data. The data was visualized using *R*[7].

The *Runner* class determines the order in which functions of the *Crawler* and *Analyzer* classes are being executed.

3.3 User Selection

For the selection of users, a reasoning about the way in which the influence of users is distributed within the Twitter network is needed. Without this, any selection of users might establish a systemic bias. For example, only few users are expected to have high influence (e.g., in terms of followers) in the Twitter network. If this is true, sampling randomly from the user space would return only very few users with high numbers of followers. As a result, no meaningful analyses on correlations between a user's number of followers and the number of retweets that user's tweets receive could be performed due to a lack of relevant data.

For an assessment of the distribution of influence, similar features may be helpful. The Power Law underlies many similar distributions, e.g., the PageRank of web sites[PRU06], the number of inbound links for blogs, or, more specifically, the number of friends of LiveJournal users[Shi03]. Therefore, a very likely candidate for a distribution of user influence is the Power Law, which implies that very large values are exceedingly rare, whereas small values are extremely common. At least, a pareto distribution, i.e., a distribution in which about 20% of all users hold roughly 80% of all influence, is very probable. From the user-specific features that are supplied by Twitter (see Section 3.1), we expected the number of followers to have the highest impact on the amount of retweets that tweets receive and thus on the influence of the user posting those tweets. Tweets are mostly seen by the followers of the publishing user, hence having more followers would lead to more people seeing and possibly retweeting a tweet.

Since we expected the number of followers that user have to underly a Power Law or Pareto distribution, we identified two options for user selection. Crawling very

[7] `http://www.r-project.org`

many users would ensure that enough users with high numbers of followers are in the dataset for analyses. However, the REST API rate limits rendered this approach unfeasible. We therefore selected the second option, which is to categorize selected users into different groups according to the numbers of followers they have and then randomly sample users from each category, thereby ensuring that enough data is present for the analysis task.

Hence, we categorized randomly selected user identifiers into nine different groups based on the number of followers each user had. Due to the aforementioned rate limit imposed on API calls, the number of Twitter users for the analyses had to be kept moderate. Out of each category, 100 randomly sampled users that were manually verified to only tweet in English were selected. This limitation on the tweet language simplified the sentiment detection task since there are reasonably reliable sentiment detection algorithms available for the English language.

Table 3.1 depicts the categories along with the amount of users that were ultimately picked as well as the average number of followers and retweets of the users in that category. Note that for the last category, only 48 users were admissible. In addition to those 848 users, a random sample of each user's English-tweeting followers was also selected to be crawled, yielding a total number of approximately 15,000 users.

#Followers	#Users	Average #followers	Average #tweets
0 - 9	100	4	160
10 - 49	100	27	763
50 - 99	100	76	2146
100 - 299	100	198	1,693
300 - 999	100	552	10,547
1,000 - 4,999	100	2,027	21,975
5,000 - 49,000	100	13,264	15,451
50,000 - 499,999	100	131,247	16,978
>500,000	48	1,455,763	16,618

Table 3.1: Users categorized by their number of followers with the average number of followers and tweets provided by Twitter per category.

Finally, we tested the initial hypothesis of how user influence (here, estimated by the number of followers a user has) is distributed. Although the selection of initial users was dependent on their respective number of followers, more than 14,000 followers of these initial users were later randomly selected irregardless of the amount of followers they had. Therefore, a chart describing the distribution of followers in our data can be considered to reflect the real distribution over all users. Figure 3.2 displays this distribution. As it can be seen, very few users have high numbers of followers, whereas most users have just very few followers. This is best described by a Pareto distribution and confirms our hypothesis.

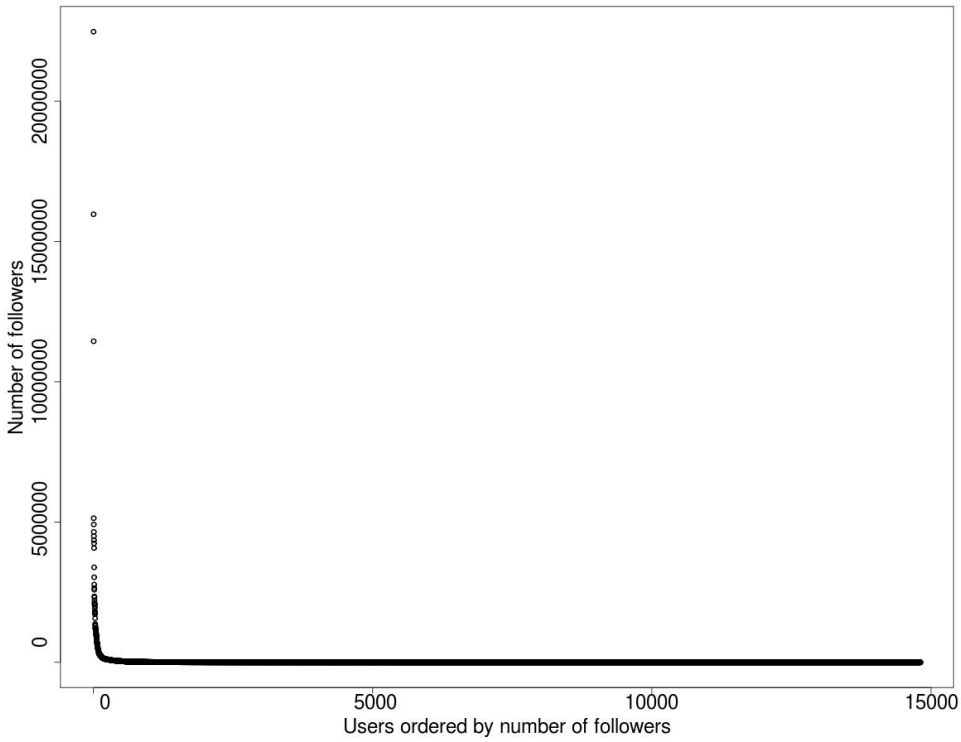

Figure 3.2: Plot showing how the number of followers users have are distributed over the users.

3.4 Data Collection

Over the course of two months, information about the users and their latest 3,200 tweets were constantly recrawled in a least-recently-crawled fashion. Each tweet was stored in a database table along with its information, e.g., current retweet count, publishing date, the text of the tweet, as well as information about the user who posted it. As the number of retweets for a tweet changes over time, the current retweet count at the time of the crawling and a timestamp were stored in a separate table. This table was used to analyze patterns of *retweet spread*, i.e., the cumulative distribution of retweets over time. Overall, more than 465 million entries with a current retweet count and timestamp were collected. To mitigate the rate limit on the API, separate crawlers were utilized in parallel.

Overall, more than 21.8 million *pure* tweets[8] and 4.2 million retweets were collected for approximately 15,000 users, averaging to 1,453 pure tweets and 280 retweets per user. The relatively high percentage of extracted retweets (16.2% as opposed to 9% of retweets in the dataset of [PGS12]) may be the result of using the official information from Twitter on whether a tweet is a retweet, rather than relying on self-developed classification methods based on a tweet's content.

Unfortunately, the Twitter Terms of Service prohibits us from releasing the collected dataset to the research community. We can, however, publish the source code used to conduct the studies described in this thesis. This encompasses the code used to crawl the Twitter data, to store it into and retrieve it from the database, perform the analyses and finally create the evaluation graphs. All the source code will be made available on a project-specific website.

3.5 Viral Tweets

At this time, a reflection about the definition of what exactly constitutes a viral tweet in the scope of this thesis is appropriate. Originally, the term of a "viral" message is an analogy to biology, where a virus quickly spreads from one person to a big group, who in turn spread the virus even further. Ultimately, that virus (or here, message) is transmitted to a large quantity of people - sometimes hundreds of thousands - within a short time. As a measure for the "virality" of a tweet, we use the number of retweets it receives. This number is supplied by Twitter. To be able to classify tweets, a threshold of retweets that tweets have to reach in order to be labeled as viral has to be determined. In order to do so, it would be useful to first

[8] A pure tweet is a tweet that is not a retweet.

get an idea about the relation between retweet numbers to the number of tweets that receive these retweets.

To this extent, we grouped all tweets in our dataset by the number of retweets they received. Figure 3.3 displays each group's respective number of retweets on the x axis and the number of tweets that are in each group, i.e., the number of tweets with that specific number of retweets, on the y axis. Note that a log-log scale is used. As it can be seen, the distribution adheres to a Power Law. Less than 5% of all tweets obtained more than 40 retweets and only about 1.1% more than 1,000. This also implies that there are very few tweets with hundreds of thousands of retweets, which would definitely be considered viral. However, these few tweets would not constitute enough feature data points to formulate solid predictions upon. Instead, we also considered tweets with fewer retweets as viral to be able to make predictions that can generally be applied to all tweets. Chapter 6 explores different thresholds for the distiction between viral and non-viral tweets and the implications on the prediction accuracy.

With the numbers of retweets that a tweet receives, we have a meaningful measure of the amount of people that are influenced by the tweet.

A second aspect for measuring virality is the speed with which the tweet spreads through the Twitter network. This aspect is not addressed in this work, mainly because of feasibility reasons: to calculate the speed with which a tweet gets retweeted, the tweet would either have to be constantly recrawled to measure the number of retweets at many consecutive points in time (since popular tweets are retweeted very frequently), or all tweets that retweeted the original tweet would have to be crawled to compare the retweet time with the publishing time.

To retrieve the needed information, both approaches require a very high number of API calls, which we can not execute for a user base large enough for analyses. As a result, we concentrated our virality assessment solely on the number of retweets. The speed with which retweets occur is briefly discussed in Chapter 7.

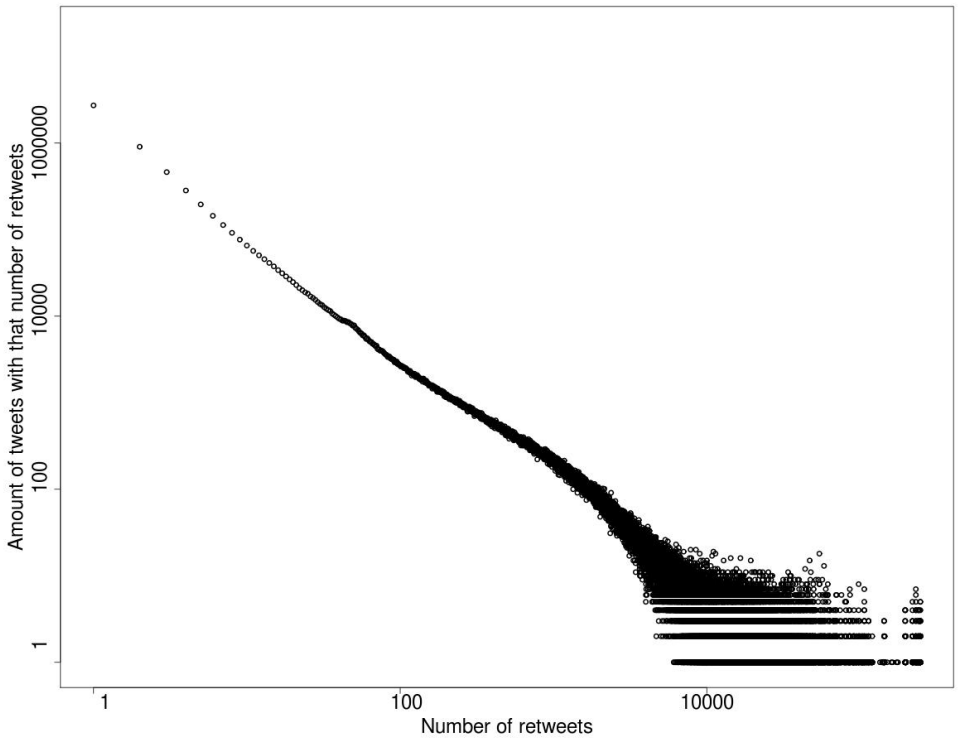

Figure 3.3: Log-log scale plot of the number of retweets a tweet can receive in relation to the number of tweets that received this number of retweets.

4 Features Influencing the Retweet Frequency

In order to identify features that influence the spread of tweets, we have analyzed a wide variety of tweet features. The measure on the tweet spread used in this research is the number of retweets the tweet obtains. The analyzed features can be categorized according to whether they are user- or tweet-specific. The features that proved most impactful were selected to build a model for predicting viral tweets. To the best of our knowledge, this is the first work that deeply analyzes such a broad spectrum of features, encompassing structural, textual, and sentiment analysis.

4.1 User-Specific Features

From the features related to users that can be obtained through Twitter REST API, we analyzed the number of followers a user has, the total number of tweets the user posted, and the Jaccard distance between the hashtags of users and that of their followers.

Number of Followers

One of the first factors that we analyzed with regard to its influence on the spread of tweets was the number of the followers of a user. The rationale for taking a closer look at this factor is the following: the more followers a user has, the higher is the visibility that his tweets will receive and hence, the higher could be the frequency with which his tweets are retweeted. Additionally, one can argue that having a high number of followers does not happen merely by chance; rather, it happens because the tweets of a user are interesting to many people.

In support of this hypothesis, the plot of Figure 4.1 shows that the average number of retweets (per tweet) grows over-proportionally with the number of followers. In order to remove clutter and to emphasize overall trends, logarithmic buckets were used in this analysis: users for which the logarithm of their number of followers is

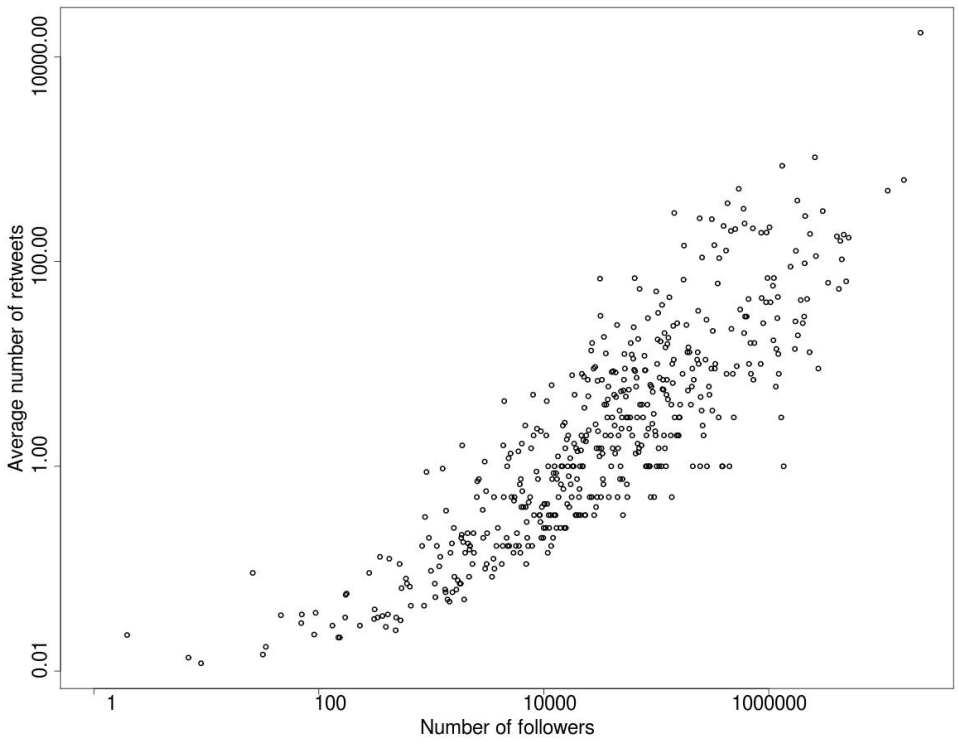

Figure 4.1: Log-log scale plot of the average number of retweets in relation to the number of followers.

approximately similar were put into the same bucket. For each bucket, the average number of followers and the average number of retweets were calculated and are displayed the log-log scale plot.

Number of Tweets

Subsequent to the analysis on the number of followers of a user, we took a closer look at the relation between a user's total number of tweets and the average retweets the posts obtain. The total number of tweets a user has published so far is supplied by Twitter, whereas the average number of retweets these tweets incur is not retrievable. However, the number of retweets for single tweets are made public by Twitter. We therefore calculated the average number of retweets on all tweets that were crawled for the user.

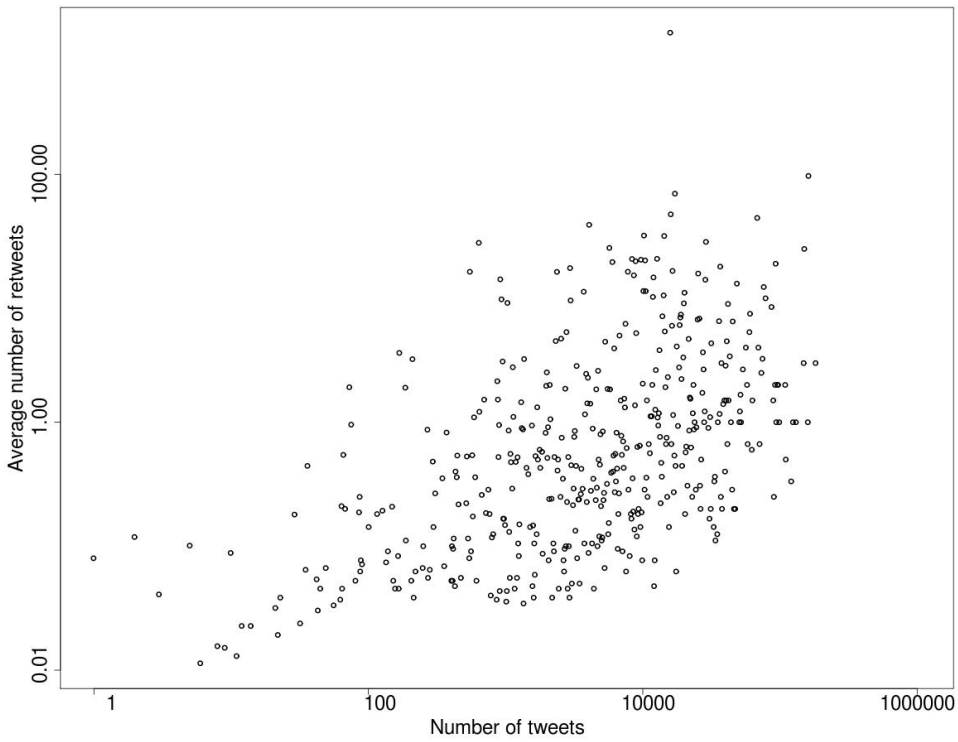

Figure 4.2: Log-log scale plot of the average number of tweets in relation to average number of retweets.

While a user might not need to constantly post tweets to be relevant, we expected users with very few tweets to obtain practically no retweets. For users with high amounts of total tweets, we conceived two conflicting hypotheses: they might either have high amounts of tweets because they are more apt to issue irrelevant statuses, resulting in a lower number of average retweets, or if posting a tweet is viewed as introducing a chance that this tweet gets retweeted, posting many tweets would potentially lead to more retweets of those tweets in total.

Figure 4.2 displays a user's average number of tweets in relation to the average number of retweets that the tweets of the user obtained. Again, logarithmic buckets were used to create averages over similar users, resulting in a log-log scale of the plot. As it can be seen, our first hypothesis seems appropriate: tweets from users with fewer than 100 total tweets obtain very few retweets. However, for higher number of total tweets, the average retweets become very scattered, making general predictions impossible.

Hashtag Jaccard Distance

Tweets are often provided with hashtags, which are basically tags preceded by a hash symbol (#) and are used to label tweets so they can be more easily categorized and found by users. Additionally, hashtags can be used to provide further explanation of a message's meaning. For example, adding *"#fail"* to a message can emphasize a misstep, *"#irony"* underlines that the message is meant to be interpreted ironically, and the *"#Obama2012"* hashtag denotes that a tweet is concerned with the election campaign of Barack Obama.

As mentioned earlier in Chapter 3, we crawled tweets from users as well as a randomly chosen subset of their followers. Here, an interesting question arises: how similar are the topics that users and their followers tweet about? Can different similarity scores have different implications on the number of retweets users receive? In order to answer this questions, we observed the hashtags present in tweets, since they should indicate the topics the tweets relate to. As a metric, we employed the *Jaccard distance*, which measures dissimilarity between two sets A and B and is specified as:

$$J_\delta(A, B) = 1 - \frac{|A \cap B|}{|A \cup B|}$$

For each user, we calculated the average Jaccard distance between the hashtags of their tweets and those of their followers. Figure 4.3 shows the distances in relation to the average retweets users receive; a high value denotes a high dissimilarity. Users generally seem to have a very high hashtag dissimilarity to their followers. This could indicate that users talk about various topics, only very few of which they share common interest in with their followers. On the other hand, hashtags shared between users might not be a meaningful measure for topic similarities.

The Jaccard distances are spread over a very narrow range and a pattern does not seem to emerge when put in relation with the average number of retweets. Therefore, drawing general inferences about the number of retweets that users can expect based upon the Jaccard distance between their hashtags and those of their followers does not seem possible.

4.2 Tweet-Specific features

A variety of tweet-specific features were extracted and analyzed with respect to their influence on the number of retweets the tweet receives. These features are: the length of the tweet, the hashtags, mentions and URLs used in the tweet, and finally the sentiments expressed in the tweet.

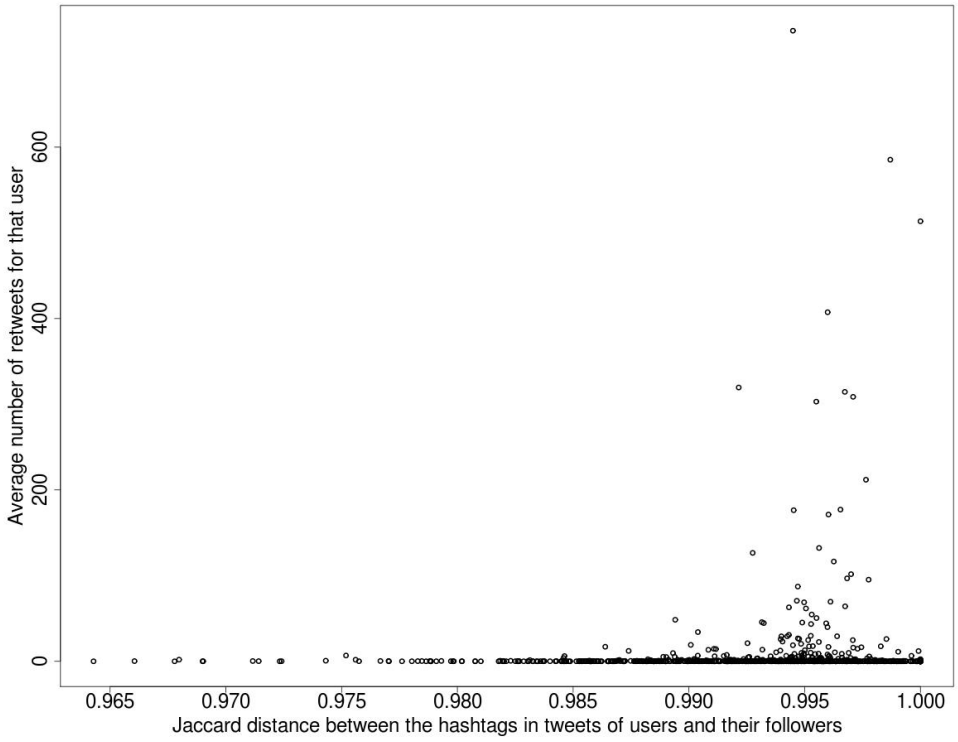

Figure 4.3: Average Jaccard distance between the hashtags of users and their followers in relation to their average number of retweets the users' tweets obtain. Note that the x axis starts at a value of 0.965.

Tweet Length

While no research was done with respect to the intrinsic reasons that motivate users to retweet a tweet, we assumed that interest is the key drive. Interest in a tweet can arise in various ways: a big information gain[1] a user perceives after reading a message can prompt interest in the tweet, causing the user to publish a retweet in order to share the newly gained knowledge. Alternatively, a user might like and concur with the statement underlying a tweet and retweet it to show support.

With this in mind, a feature worth analyzing is the textual length of a tweet. An intuitive assumption here is that the longer a tweet is, the more information it can contain. Furthermore, the more information a tweet contains, the more interesting it could be, and hence the more often it could be retweeted. Consequently, we expected to find that the retweet count for a tweet correlates with the number of characters of that tweet.

The plot of Figure 4.4 depicts the average retweet count for all possible tweet lengths[2]. Disregarding the outliers (which are due to few tweets with a very high retweet count, pulling up the averages) and focusing on the trends, the data seems to support our hypothesis. Up to 120 characters, the expected number of retweets seems to grow almost proportionally with the tweet's length. However, this trend is reversed for tweets with more than 120 characters. An explanation for this finding might be that experienced Twitter users are very adapt at condensing information into short tweets. Therefore, the complete available tweet length might mostly be utilized by relatively new users that are not popular or experienced enough to publish tweets that many people find interesting.

Hashtags

Another textual parameter that we analyzed are the hashtags contained in a tweet. As mentioned above, hashtags can be used to add context to a tweet. Given such use cases of hashtags, one would expect that interesting tweets are more likely to contain hashtags, thus enabling users to more easily find and talk about them.

Indeed, the plot of Figure 4.5 shows that tweets containing one to three hashtags are more likely to be retweeted than tweets without any hashtags. However, as the number of hashtags in a tweet grows, the expected number of retweets decreases. This can be explained by the increased character consumption of multiple hashtags,

[1] The concept of information gain is based on information theory and is to be interpreted as the surplus of information a user has after taking new information into account.

[2] Keep in mind that tweets are limited to 140 characters.

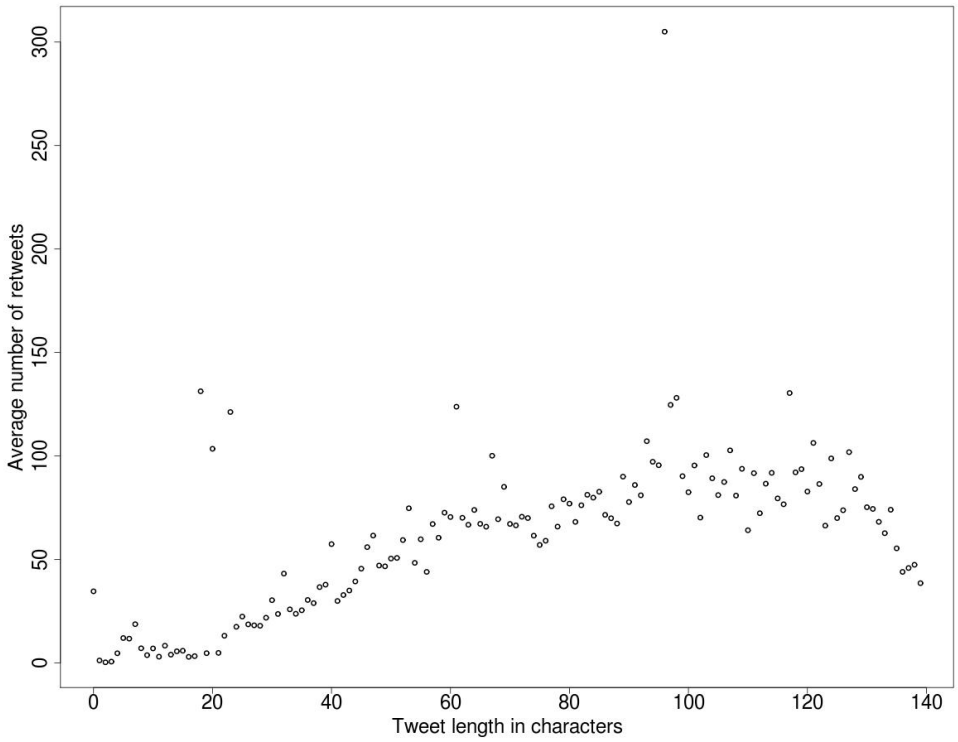

Figure 4.4: Average number of retweets in relation to the tweet length in characters.

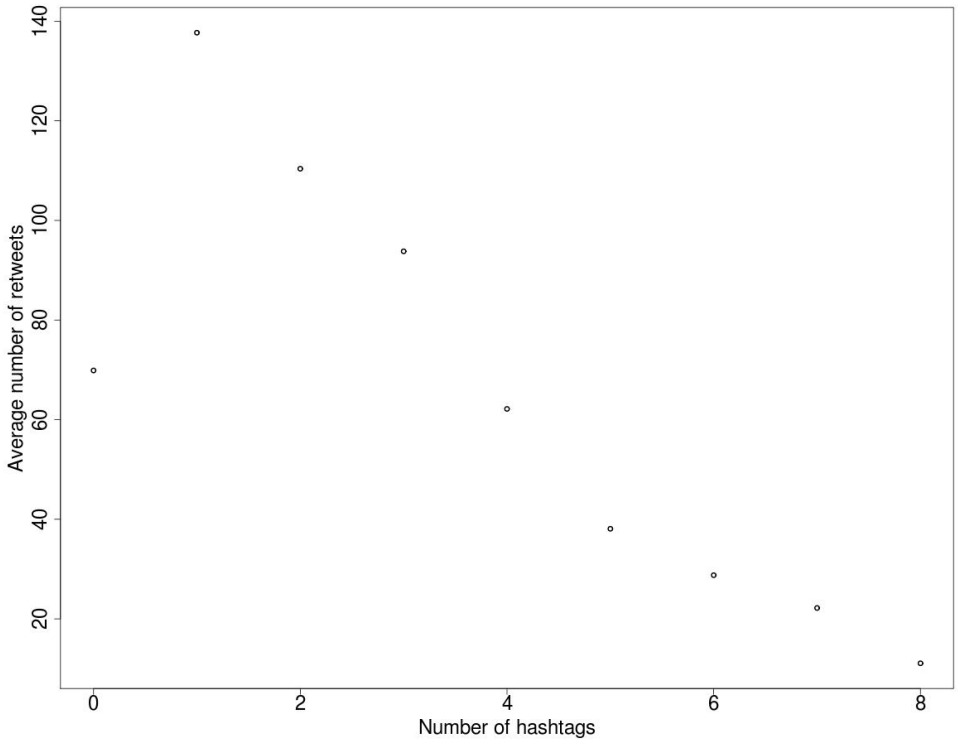

Figure 4.5: Average number of retweets in relation to the number of hashtags used in the tweet.

leaving less space for the actual information.

Mentions

Quite analogously to hashtags, mentions are used to tag and address Twitter users whose names occur in a tweet. They are specified by an @ sign followed by a user name. As in the case of hashtags, employing mentions increases the number of retweets a tweet receives on average. However, a larger number of mentions also leaves less space for the actual content, thus a similar behavior concerning the expected number of retweets can be observed, see Figure 4.6.

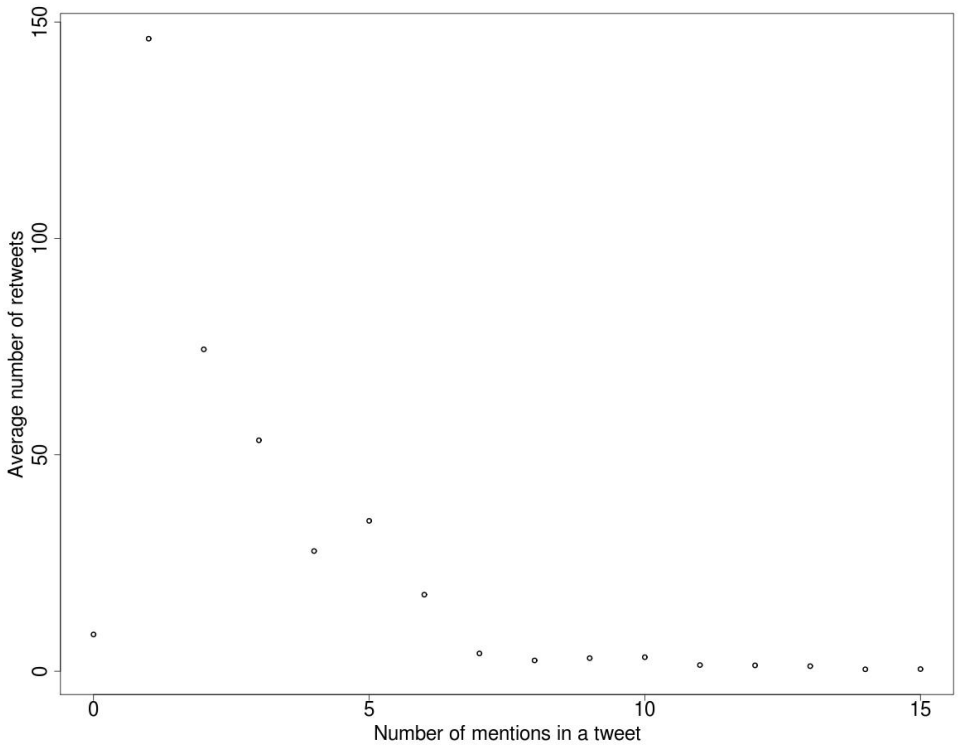

Figure 4.6: Average number of retweets in relation to the number of mentions used in the tweet.

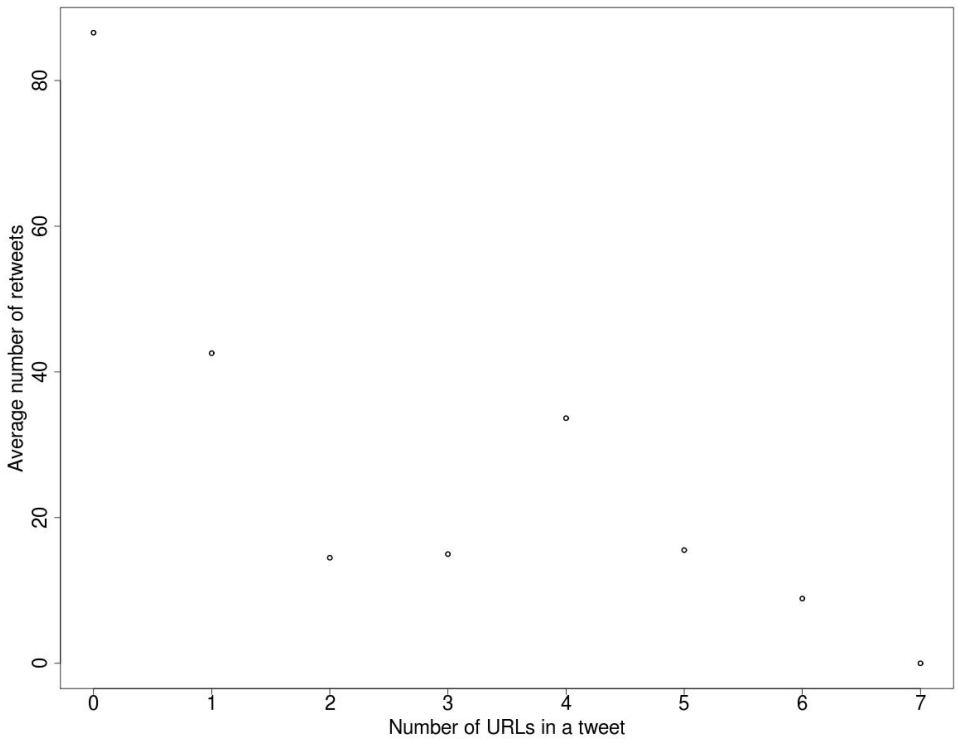

Figure 4.7: Average number of retweets in relation to the number of URLs used in the tweet.

URLs

Yet another tweet-specific feature we analyzed was the number of links embedded in a tweet. We expected a result similar to the use of hashtags and mentions in a tweet: larger numbers of URLs should be accompanied with fewer retweets since they take up space otherwise usable for content. Also, one to three URLs should induce more retweets than none whatsoever, since a URL should link to a relevant website or document that provides useful information. However, Figure 4.7 disproves that theory, showing that tweets without any URLs at all receive, on average, most retweets. The spike for tweets with four URLs can be explained by a few tweets with high numbers of retweets pulling up the averages, which is amplified since the amount of tweets that contain a certain number of URLs decreases for increasing numbers of URLs. A possible explanation is that URLs are mainly used in tweets that, for example, just communicate the user's current position by posting a link to it and are therefore uninteresting to many users[ABL12].

Sentiment Analysis

Another factor analyzed with respect to its influence on the retweet frequency is the sentiment of a tweet. The goal of this analysis was to see whether tweets with positive sentiment undergo a different diffusion process than tweets with negative sentiment. The sentiment analysis of Twitter messages proves difficult because most common sentiment detection algorithms need to analyze texts that are longer than tweets, which are heavily limited in their length. Additional difficulties are introduced since messages on Twitter often contain abbreviations that are common on the Internet, but are not found in conventional dictionaries. To infer the sentiment of a tweet, we used the *SentiStrength* algorithm[TBP+10].

SentiStrength was specifically developed for sentiment analysis of short Internet messages. Apart from detecting negating[3] and boosting[4] terms, it also takes into account typical Internet communication styles like the use of emoticons[5] and colloquial jargon[6]. For a given input tweet, SentiStrength returns a score representing both the positive and negative sentiment expressed in the tweet. This score ranges from -1 to -5 for negative and +1 to +5 for positive sentiments, with 1 standing for a statement void of sentiment and 5 for a statement affiliated with the utmost sentiment. For example, the message "Dogs are fantastic but cats are awful" has a positive sentiment score of +3 and a negative sentiment score of -4, whereas the message "Dogs are fantastic" again has a positive score of +3, but a negative score of -1 (indicating that no negative sentiment is present in the tweet). Based on the sentiment scores, we calculated the predominant overall *sentiment valence* (i.e., whether a tweet has a predominantly positive, neutral, or negative sentiment).

Sentiment Valence

Pfitzner et al. already employed SentiStrength to analyze the distributions of tweets and retweets for predominantly positive, negative, and neutral tweets ([PGS12]). They show that in each of these categories, the fraction of tweets resembles that of retweets. However, the dataset used in [PGS12] has an average of ten pure tweets for each of the approximately three million users, while our dataset has an average of 1,453 pure tweets for 15.000 users, and a much higher percentage of retweets. We therefore were interested to find out whether the findings of Pfitzner et al. would hold for our data as well.

[3] A negating term reverses the sentiment polarity of following terms. For example, in "not happy", the "not" reverses the positive sentiment of "happy", turning it into a negative statement.

[4] A boosting term is one that increases the sentiment of other words. For example, "very happy" is more positive than "happy". "Very" serves as a booster, increasing the sentiment.

[5] Emoticons are, for example, ":-)" or ":-/', and also reflect the user's sentiment).

[6] This encompasses common abbreviations on the Internet, such as "lol" for "laughing out loud'.

	Negative valence	Neutral valence	Positive valence
Pure tweets	17.8%	37.68%	44.48%
Retweets	22.4%	37.13%	40.49%

Table 4.1: Pure tweets and retweets in relation to sentiment valences.

	Negative valence	Neutral valence	Positive valence
Pure tweets	19.9%	33.8%	46.3%
Retweets	19.8%	31.4%	48.8%

Table 4.2: Pure tweets and retweets in relation to sentiment valences, as reported in [PGS12].

Table 4.1 shows the proportion of pure tweets and retweets in relation to their sentiment valence in our data. As it can be seen, tweets with a positive valence make up for the largest group of all tweets while those with negative valence form the smallest group. For the negative sentiment valence, the fraction of retweets is higher than that of tweets, while it is the opposite for the positive sentiment valence. This is in contrast to the findings of [PGS12], which are shown in Table 4.2. However, their finding that the data confirms the *Pollyanna Hypothesis* ([BO69]), which states that human language shows a bias towards positivity, is still supported by our data.

Interestingly, our findings imply that tweets with predominantly negative sentiment have a higher retweet probability. This can be shown by applying Bayes' theorem: the probability that a negative tweet gets retweeted, i.e., the probability of a retweet given a negative tweet sentiment valence, is

$$P(retweet|negative) = \frac{P(negative|retweet) \times P(retweet)}{P(negative)}$$

The conditional probabilities in the above formula can be calculated by their maximum likelihood estimations, which results in:

$$P(negative|retweet) = \frac{\text{Number of retweets with a negative valence}}{\text{Number of retweets}} \approx 0.224$$

$$P(retweet) = \frac{\text{Number of retweets}}{\text{Number of pure tweets + number of retweets}} \approx 0.162$$

$$P(negative) = \frac{\text{Number of pure tweets and retweets with a negative valence}}{\text{Number of pure tweets + number of retweets}} \approx 0.186$$

Therefore:

$$P(retweet|negative) = \frac{0.224 \times 0.162}{0.186} \approx 0.195$$

The probabilities of a retweet given a neutral or positive sentiment valence can be estimated analogously:

$$P(retweet|neutral) = \frac{0.371 \times 0.162}{0.376} \approx 0.160$$

$$P(retweet|positive) = \frac{0.405 \times 0.162}{0.438} \approx 0.150$$

As it can be seen, a tweet with a negative sentiment valence has a higher probability of being retweeted than a tweet with neutral or positive valence. In fact, a tweet with a positive sentiment valence is least likely to be retweeted. A possible explanation for this finding might be that predominantly negative statements (which could typically relate to negative experiences) are more likely to attract the attention of other users, thus increasing the visibility and the retweet probability.

Note that "retweet" in the above formulas only indicates whether a tweet gets retweeted at least once. No quantitative statment concerning the actual number of retweets is made.

Emotional Divergence

Another way to use the positive and negative sentiment scores of a tweet as returned by the SentiStrength algorithm is to combine them into a single score reflecting the *emotional divergence* of the tweet. The notion of emotional divergence was introduced by [PGS12]. For a given tweet t, it can be calculated as

$$divergence(t) = \frac{\text{postive-score}(t) + |\text{negative-score}(t)|}{10}.$$

Remember that the SentiStrength algorithm returns scores in the range of [+1, +5] for positive and [-1, -5] for negative sentiments. Hence, $0.2 \leq divergence(t) \leq 1.0$. This score aims to capture how strongly sentiments are expressed in a tweet regardless of the overall sentiment valence. Analogously to the sentiment valence, the emotional divergence score can be used to calculate retweet probabilities.

The authors of [PGS12] report that an increase in emotional divergence yields a higher retweet probability. The analysis on our dataset confirms these findings. The results are shown in Table 4.3. Note that there are very few tweets and retweets with divergence values between 0.9 and 1.0, hence the calculated retweet probabilities in these cases might be susceptible to outliers.

Emotional divergence	Tweets	Retweets	Retweet probability
0.2	30.61%	29.16%	15.56%
0.3	36.73%	33.78%	15.09%
0.4	20.81%	22.36%	17.20%
0.5	8.25%	9.86%	18.77%
0.6	2.72%	3.51%	19.98%
0.7	0.75%	1.13%	22.63%
0.8	0.12%	0.18%	22.69%
0.9	$\approx 0.01\%$	$\approx 0.01\%$	20.03%
1.0	$\approx 0.001\%$	$\approx 0.001\%$	30.03%

Table 4.3: Emotional divergence scores for different fractions of pure tweets and retweets.

Individual Sentiments

So far, we have introduced the sentiment valence as a measure for the influence of negative, neutral and positive sentiments on the probability that a tweet will be retweeted. The emotional divergence reflects how the amount of sentiment that is expressed influences the number of retweets. However, we did not yet examine the effects of individual sentiment score pairs as they are returned by the SentiStrength algorithm.

Figure 4.8 depicts all 25 combinations of the 5 positive and 5 negative sentiment scores and the average number of retweets for tweets associated with these sentiments. For all but the strongest positive sentiment scores, the average number of retweets rises with an increase of negative sentiment from -2 to -4 and then drops down if the negative sentiment is -5. This implies that negative statements are more interesting to users resulting in more retweets, but also that overly negative statements are not well received in the Twitter community. Generally, tweets with a sentiment of +3 seem to incur larger number of retweets.

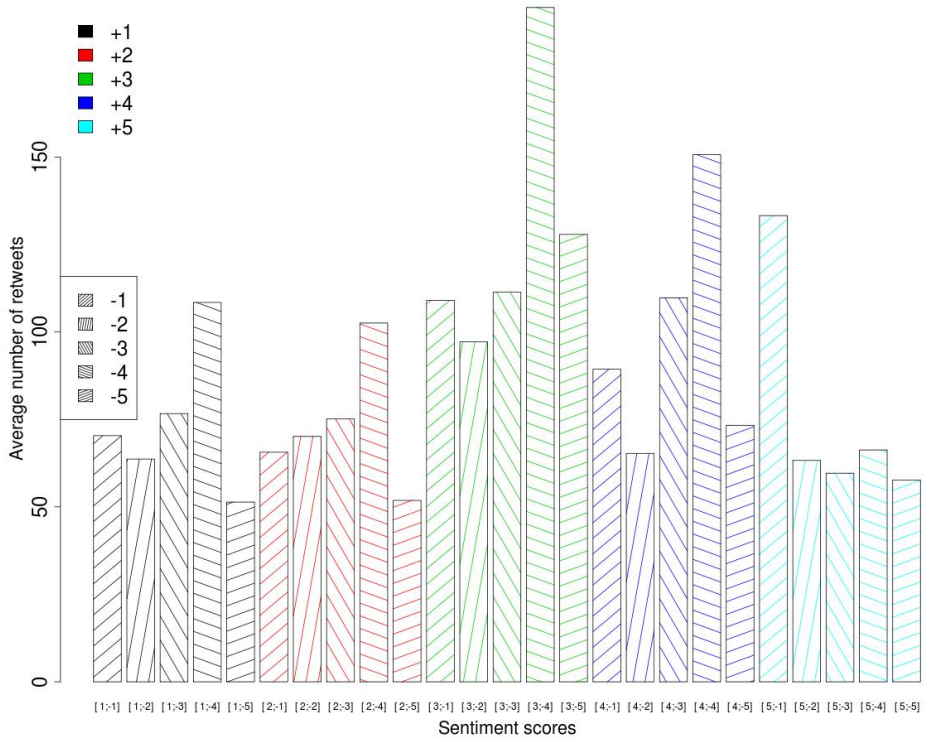

Figure 4.8: Average number of retweets for individual sentiment scores. The sentiments are ordered from +1 to +5 and, for each positive score (indicated by a different color), by negative score from -1 to -5.

5 Predicting Viral Tweets

In the previous section, we analyzed the impact of different tweet- and user-specific features with respect to their impact on the number of retweets tweets receive. In this section, we propose a probabilistic model that predicts whether a given tweet will become viral (i.e., whether it will be frequently retweeted) by taking into account all significant features we have discussed so far. The strength of the proposed model is that it avoids over-simplifying assumptions, e.g., conditional independence between features (given the class), as the interdependencies between the features can be crucial for the prediction task addressed in this section. For example, it seems that on Twitter, the number of tweets and the number of followers of a user strongly depend on each other: the more a user tweets, the more users follow him. Also, the number of hashtags in a tweet influences the tweet length to some extent, as hashtags consume characters in addition the actual message. Hence, for the task of predicting viral tweets, it is important to avoid over-simplifying assumptions, such as independence between features. To show that the proposed model outperforms models that make such simplifying assumptions on the feature interdependencies, we compare it to a Naive Bayes prediction model which serves as a baseline.

5.1 Generalized Linear Model

Let $\mathbf{x}$ be a tweet described by a feature vector[1] $\boldsymbol{\phi}(\mathbf{x}) = (x_1, ..., x_n)$. First of all, we call a tweet viral if its total number $N_R(\mathbf{x})$ of retweets is greater than a threshold T. Since we do not know beforehand the exact impact of each feature on the total number of retweets, we first assign to each feature an unknown weight w_i and define an overall latent "virality" score for a tweet as

$$v(\mathbf{x}) = \sum_{i=1}^{n} w_i x_i \tag{5.1}$$

[1] Note that the vector representation of a tweet comprises tweet-related features such as the tweet length, the number of hashtags, the sentiment valence, etc., as well as user-related features, such as the number of followers.

Now we can estimate the probability that a tweet $\mathbf{x}$ will have more than T retweets given its score by

$$P(N_R(\mathbf{x}) > T \mid v(\mathbf{x})) = f(w_0 + v(\mathbf{x})) \tag{5.2}$$

where w_0 is the intercept of the model and f is a general sigmoid activation function. In the implementation of the model, we define f as the logistic function, i.e.,

$$f(\alpha) = \frac{1}{1 + e^{-\alpha}}. \tag{5.3}$$

Figure 5.1 shows the logistic function as defined in Equation (5.3). For small x values, the function displays an approximately exponential growth, which slows down to a linear growth as x increases and finally comes to a near halt. Since it has a lower bound of 0 and an upper bound of 1, it can be interpreted as a probability distribution. For $x = w_0 + v(\mathbf{x}) = 0$, the probability value $y = f(w_0 + v(\mathbf{x}))$ is 0.5. Hence, this point can be interpreted as a decision boundary that indicates which class a tweet falls into.

Note that this choice of the activation function leads to a logistic regression model and goes hand in hand with the assumption that the random variable

$X_j :=$ Number of samples $\mathbf{x}$ with $\boldsymbol{\phi}(\mathbf{x}) = \boldsymbol{\phi}(\mathbf{x}_j)$ until $N_R(\mathbf{x}) > T$ is found

follows a Binomial distribution $Bin(k, p)$ with parameters $k \in \mathbb{N}$ and $p \in [0, 1]$, where

$$p := \frac{1}{1 + e^{-w_0 - v(\mathbf{x}_j)}} = E[\frac{X_j}{k} | \boldsymbol{\phi}(\mathbf{x}_j)]. \tag{5.4}$$

Note that $logit(p) = ln(\frac{p}{1-p}) = w_0 + v(\mathbf{x})$, thus making the retweet odds log-linearly dependent on the features.

Assuming that X_j follows a Binomial distribution makes the model quite flexible as the Binomial distribution is relatively versatile and can approximate the normal as well as many skewed distributions.

Finally, the goal of the model is to learn the values of the weights $w_0, w_1, ..., w_n$ from given training data. We use the Weka toolkit[2] to learn the weights with a Quasi-

[2] `http://www.cs.waikato.ac.nz/ml/weka/`

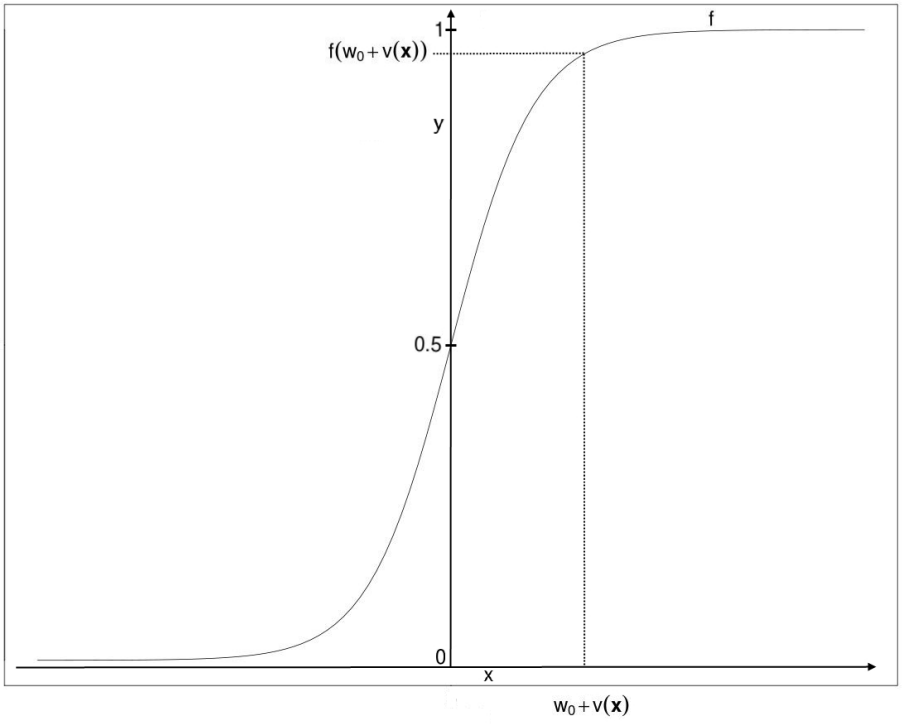

Figure 5.1: Graph of the logistic function.

Newton method and evaluate the prediction accuracy of the model in comparison to that of a Naive Bayes approach in the next chapter.

5.2 Baseline: Naive Bayes Model

As discussed earlier, making over-simplifying assumptions concerning the interdependencies between features could impair the prediction of viral tweets. To see whether this is true, we implemented a Naive Bayes prediction model as follows:

The joint probability $P(N_R(\mathbf{x}) > T, x_1, ..., x_n)$ for a tweet $\mathbf{x}$ with features $x_1, ..., x_n$ can be written as

$$P(N_R(\mathbf{x}) > T, \boldsymbol{\phi}(\mathbf{x})) = P(\boldsymbol{\phi}(\mathbf{x}) \mid N_R(\mathbf{x}) > T)P(N_R(\mathbf{x}) > T). \tag{5.5}$$

In order to avoid the curse of dimensionality, for the estimation of $P(\boldsymbol{\phi}(\mathbf{x}) \mid N_R(\mathbf{x}) > T)$ one can assume that the features are conditionally independent given the class. More specifically

$$P(\boldsymbol{\phi}(\mathbf{x}) \mid N_R(\mathbf{x}) > T) = \prod_{i=1}^{n} P(x_i \mid N_R(\mathbf{x}) > T). \tag{5.6}$$

The conditional probabilities $P(x_i \mid N_R(\mathbf{x}) > T)$ as well as the probability $P(N_R(\mathbf{x}) > T)$ can be estimated by their maximum likelihood estimations

$$P(x_i \mid N_R(\mathbf{x}) > T) = \frac{\#\{\mathbf{x} | (x_i \neq 0) \wedge N_R(\mathbf{x}) > T\}}{\#\{\mathbf{x} | N_R(\mathbf{x}) > T\}} \tag{5.7}$$

and analogously

$$P(N_R(\mathbf{x}) > T) = \frac{\#\{\mathbf{x} | N_R(\mathbf{x}) > T\}}{\#\{\mathbf{x} | N_R(\mathbf{x}) \geq 0\}}. \tag{5.8}$$

Despite the above independence assumptions, the Naive Bayes model is a very popular one and often leads to satisfactory results ([Ris01]). Here we are interested in its performance in comparison to the previously described general linear model.

6 Evaluation of the Prediction Models

To train and evaluate our prediction models described in Chapter 5, we used the Weka toolkit on a random sample of viral and non-viral tweets from our dataset. For every tweet in this sample, the total number of retweets was known as well as all features used in our model. Note that each feature domain was discretized in a reasonable way with respect to the insights from Chapter 4, so that feature values incurring a similar number of retweets fell into the same group. For example, in the case of hashtags, tweets with one to three hashtags were put into one group, tweets without any and those with four hashtags formed another group, and tweets with five or more hashtags made up the last group. Table 6.2 lists all features, how they were discretized according to their values, and the respective weights that were learned by the logistic regression model.

We chose different thresholds T to separate viral from non-viral tweets. As mentioned in Section 3.5, T had to be chosen with the respect to the resulting amount of tweets whose number of retweets surpass the threshold. If T is chosen too high, not enough tweets fall into the "viral" category and the data will not be sufficient for meaningful analyses. We expected higher thresholds to lead to better prediction accuracy, since the feature values should be more discriminative for tweets with many retweets. To test this hypothesis, we ran the experiments with $T \in \{50, 100, 500, 1000\}$. Both the Naive Bayes and the generalized linear model were evaluated based on tenfold cross-validation.

For the evaluation, two different measured are used: accuracy and F-Measure. The accuracy conveys the proportion of properly classified outcomes and is defined as the ratio of true positives and true negatives in relation to all classifications (i.e., the true positives, true negatives, false positives, and false negatives). The F-Measure (also known as F1-score or harmonic mean) considers both the precision[1] and the recall[2]. It is defined as

$$F = 2 \times \frac{precision \times recall}{precision + recall}.$$

Table 6.1 shows the resulting accuracy and F-Measure scores, showing that the gen-

[1] The precision is the proportion of true positives in relation to the true positives and false positives.
[2] The recall is the proportion of true positives in relation to to the number of true positives and false negatives.

T	Accuracy NB	Accuracy GLM	F-Measure NB	F-Measure GML
50	91.6%	93.6%	0.916	0.936
100	92.7%	94.0%	0.927	0.940
500	94.7%	96.3%	0.947	0.963
1000	95.1%	96.8%	0.951	0.968

Table 6.1: Prediction accuracy and F-Measure scores for the Naive Bayes (NB) and the generalized linear model (GLM) for different thresholds T.

eralized linear model generally provides more reliable predictions than the baseline Naive Bayes model. As discussed in Chapter 5, this can be attributed to the fact that the generalized linear model does not assume conditional independence between the features. Note that the prediction accuracy indeed increases with T, indicating that the features become more discriminative between viral and non-viral tweets with higher T.

Another advantage of the generalized linear model is that it learns weights for the different features. These weights correspond to the importance of each feature for predicting viral tweets. Considering these weights, our model suggests that the number of followers of a user has the largest influence on the prediction. It is followed by the number of mentions, URLs, the emotional divergence, the number of hashtags, the individual sentiments and sentiment valences, and finally the length of a tweet.

Besides by considering the different feature weights, the impact of features can also be determined through a separate generalized linear model that was trained on a restricted feature set. To find out the importance of the number of followers, for example, weights can be learned for all features except the number of followers. For a threshold T of 1000 retweets, the resulting model has a prediction accuracy of 68.8%. While this accuracy is better than guessing randomly, the gap to the complete model with an accuracy of 96.8% is evident. In contrast, restricting the features to just the number of followers yields an impressive accuracy of 91.6%. This shows that the number of followers indeed is very important, yet the classification can still be substantially improved by taking into account other features.

Feature	Feature values v	Weight
Sentiment valence	$v = positive/negative$	0.1395
Sentiment valence	$v = negative$	-0.1395
Tweet length	$v < 40$	0.6074
Tweet length	$40 <= v < 60, v > 130$	0.0811
Tweet length	$60 <= v <= 130$	-0.4202
Number of mentions	$v = 1$	1.2278
Number of mentions	$2 <= v <= 3$	1.4240
Number of mentions	$5 <= v <= 6$	1.3618
Number of mentions	$v = 0 \mid v >= 7$	-1.4466
Number of hashtags	$v >= 5$	0.2106
Number of hashtags	$v = 0 \mid v = 4$	0.0808
Number of hashtags	$1 <= v <= 3$	-0.0816
Number of followers	$v < 10,000$	6.5304
Number of followers	$10,000 >= v < 300,000$	-0.8486
Number of followers	$v >= 300,000$	-6.0313
Emotional divergence	$v <= 0.3 \mid v >= 0.9$	0.2233
Emotional divergence	$0.4 <= v <= 0.6$	-0.1951
Emotional divergence	$0.7 <= v <= 0.8$	-0.6556
Number of URLs	$v = 0$	-1.2933
Number of URLs	$v = 1 \mid v = 4$	1.2794
Number of URLs	$2 <= v <= 3 \mid v >= 5$	1.2602
Individual Sentiment	$v \in \{(1;-5),(2;-1),(2;-5),(4;-2),(4;-5)\}$	-0.1468
Individual Sentiment	$v \in \{(1;-4),(2;-4),(3;-1),(3;-2),(3;-3),(3;-4);(3;-5),(4;-4),(5;-1)\}$	0.2082
Individual Sentiment	$v \in \{(1;-1),(1;-2),(1;-3),(2;-2),(2;-3),(4;-1),(4;-3),(5;-2),(5;-3),(5;-4),(5;-5)\}$	-0.0214
Intercept (w_0)	-	2.4624

Table 6.2: Learned feature weights for the generalized linear model and a threshold $T = 1000$.

7 Predicting Numbers of Retweets

So far, we have focused on the task of predicting whether a tweet will have more than a certain number of retweets. While this information itself is already valuable, it is only a binary prediction. For practical purposes, it would be even more useful to forecast the approximate number of retweets a tweet will incur.

7.1 Possibilities for Predicting the Number of Retweets

To this end, we looked into the time span during which a tweet is "active" (i.e., is still being retweeted). This time span is of special interest for analyzing the spread distributions of tweets over time. If such a distribution was known for a given tweet, one could make exact predictions about its number of retweets at arbitrary time points after its posting. Initially, we expected to find different patterns of spread distributions based on the overall retweet counts of tweets, e.g., tweets that receive high numbers of retweets would need substantially more time to amass their retweets than tweets that receive only few retweets. However, this assumption was disproved by the subsequent analysis.

As described in Chapter 3, during the crawling phase for the tweets in our dataset, we also collected retweet counts along with corresponding timestamps. Using this information, for each timestamp and each tweet, we calculated the fraction of retweets at that time from the overall number of retweets. To make sure that the number of retweets was almost final and not due to change significantly, we only considered tweets that had a posting date that was at least two weeks old. We then determined the time that passed between the posting of a tweet and the different timestamps. Subsequently, for a time resolution of one hour, each retweet was allocated to the nearest hour since the posting of the original tweet. To analyze different spread patterns, we finally categorized the tweets into classes depending on their overall retweet count. The retweet thresholds we chose were 50, 100, 500, and 1000 (these are the thresholds we also used in Chapter 6), resulting in five classes (one class for less than 50 retweets, one class between each threshold, and one class for more than 1000 retweets).

The result of the analysis are depicted in Figure 7.1. As it can be seen, the spread

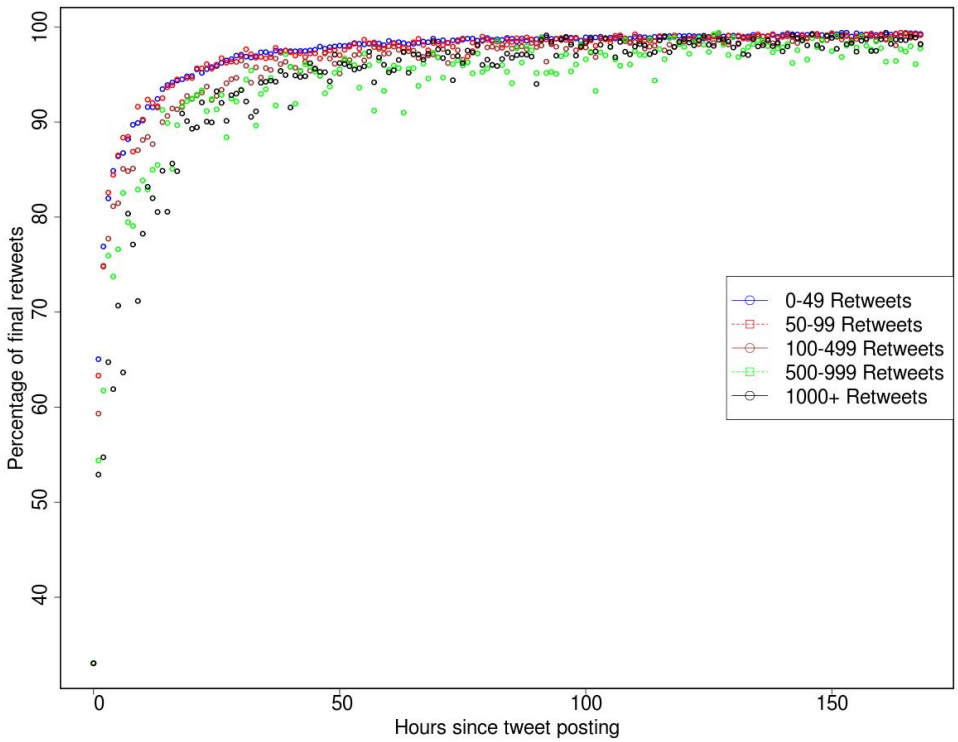

Figure 7.1: Percentage of total retweets obtained so far over time for tweets with different numbers of retweets.

distributions look alike for all tweets, regardless of their class. Tweets with a high number of retweets need just a little longer to receive their retweets than tweets with very few retweets.

Interestingly, a tweet receives a substantial proportion of its retweets within one hour (less than 50 retweets: 51%, more than 1000 retweets: 33%) and most within 10 hours (less than 50 retweets: 92%, more than 1000 retweets: 78%), rising to more than 97% for all classes after 72 hours.

This impressively quick dissemination of tweets can be partly explained by the wide distribution of smartphones, allowing users to constantly monitor their Twitter news feed and publish retweets even when they are not in front of a computer. However, this rapid spread of tweets also diminishes the use of this prediction since the distribution patterns are alike for all classes and most retweets occur already within the first hours.

Hence, we decided to take another look at the generalized linear model from Chapter 5. Recall that this model learns an intercept weight and specific weights for each feature, given a threshold T on the number of retweets. The linear combination of these weights[1] yields a "virality" score for each tweet. The higher this score is for a given tweet, the higher is the belief that this tweet will receive more than T retweets. The higher this belief is, the higher should be the expected number of retweets for that tweet.

By exploiting the relation between these scores and the expected numbers of retweets, we could approximately predict the desired retweet counts of tweets. By applying the activation function of the model[2] to the linear combination of the feature weights of a given tweet, the resulting value $f(w_0 + v(\mathbf{x}))$ represents the probability of that tweet being viral, i.e., incurring more than T retweets. These probabilities can then be aligned with the corresponding known numbers of total retweets for the tweets in our dataset. If this alignment could be interpolated by a function we would have a tool for predicting the expected number of retweets.

7.2 Evaluation

Based on our training set with known number of retweets, we calculated the retweet probability for each tweet and aligned it with the number of retweets that the tweet received. The resulting scatter plot is depicted in Figure 7.2. For probability values of less than 0.6, the numbers of retweets are generally close to each other. For values between 0.6 and 0.8, some tweets have much higher numbers of retweets than others, resulting in some vertical scattering of the number of retweets. For values larger than 0.9, a considerable scattering exists, ranging from thousands of retweets to more than 100,000 retweets.

Next, we calculated the average number of retweets for similar probability values and aligned them with their average number of retweets. As shown in Figure 7.3, this alignment can be approximated through a polynomial function. For probability values up to 0.9, this function[3] is given by $y = -8289x^3 + 13,800x^2 - 3452x + 232$ and predicts the numbers of retweets with average prediction errors depicted in Table 7.1. Two different error measures are used. The average absolute error $\sum_i |retweets(t_i) - prediction(t_i)|$ describes how far off the prediction is from the actual value, whereas the average bias error $\sum_i retweets(t_i) - prediction(t_i)$ displays whether the predictions are biased, i.e., are generally off in favor of higher or lower predictions.

[1] The weights used here were learned based on a threshold $T = 1000$ on the retweet count.
[2] The activation function is specified in Equation (5.3).
[3] The function was calculated as a best fit by Wolfram Alpha, `http://www.wolframalpha.com/`.

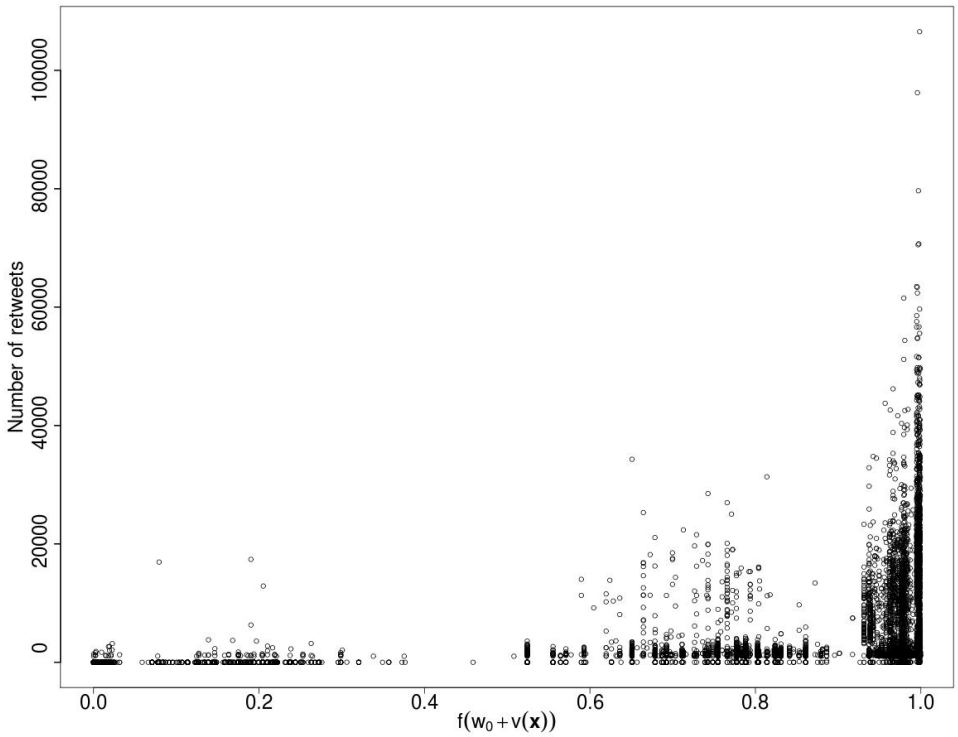

Figure 7.2: Scatter plot of the probability value $f(w_0 + v(\mathbf{x}))$ in relation to numbers of retweets.

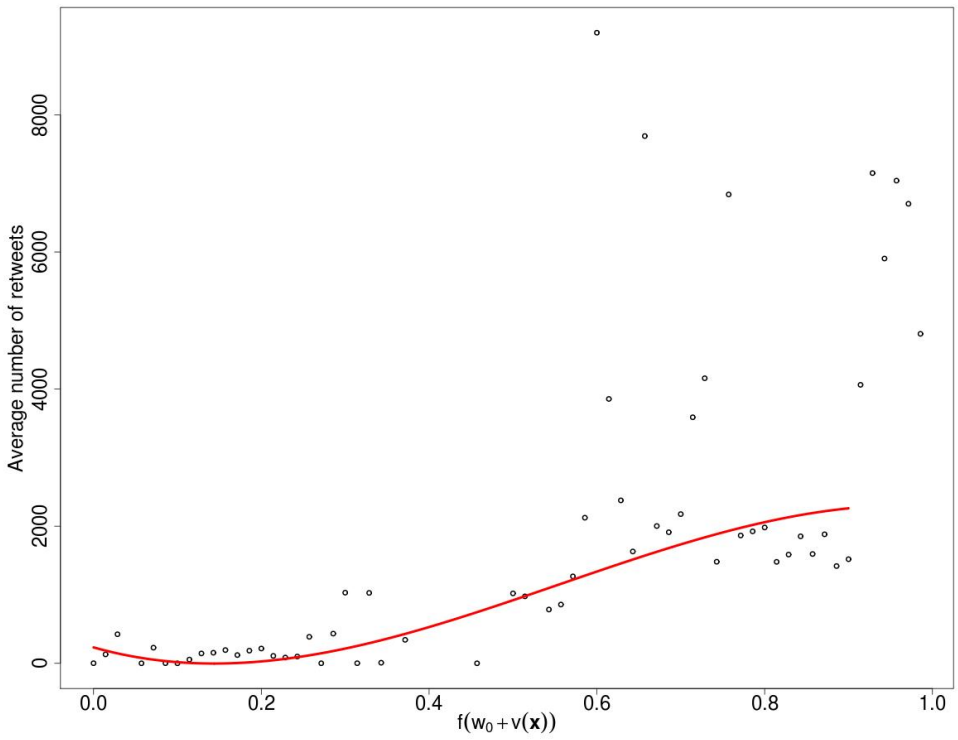

Figure 7.3: Average numbers of retweets for probability values less than 0.8. The polynomial that best fits the data is $y = -8289x^3+13800x^2-3452x+232$.

Unfortunately, this method for approximating retweet counts becomes useless for probability values greater than 0.9. As Figure 7.2 shows a high variation of retweet counts for these values, the average prediction error becomes relatively large. A similar effect can be observed for probability values between 0.6 and 0.8. Although the values are not scattered as much, the average prediction error is considerably higher than for probability values between 0.8 and 0.9, where there is no such scattering.

The scattering of retweet counts can be explained by the Power Law underlying the distribution of numbers of retweets mentioned in Section 3.3. In the long tail, many tweets receive only few retweets. These tweets have similar features and thus are easy to classify and to approximate. However, few tweets have substantially lager numbers of retweets than the rest, yielding too few data points to make build a classifier that can, for example, distinguish between tweets that will receive more than 50,000 retweets and tweets that will receive more than 100,000. Hence, all of these highly viral tweets are assigned a very high probability of being viral for the thresholds used in this paper. However, other tweets that receive fewer retweets also

Probability	Average absolute error	Average bias error
0.0 - 0.4	224	-188
0.5 - 0.6	940	26
0.6 - 0.8	1556	473
0.8 - 0.9	1005	-359

Table 7.1: Average absolute error and bias error for different probability ranges.

receive a similar probability, resulting in the scatterplot presented in Figure 7.2.

8 Discussion and Future Work

As already mentioned in Chapter 3, we defined a viral tweet as a tweet that receives a relatively high number of retweets (e.g., $T \geq 1000$ retweets). This choice influences all analyses and predictions performed in this thesis, as all features are analyzed with respect to the average number of retweets their respective values correlate with. If more tweets with very high numbers of retweets were to be crawled, it would be interesting to see whether the general findings of this thesis also hold for those tweets and where their features differ. This would also allow to increase the threshold for viral tweets, which could further increase the prediction accuracy and might even help improve the estimation of expected retweets for tweets.

Alternatively, with more data, the assessment of a tweet's "virality" could be based on more than the mere number of retweets it obtains. Section 3.5 already mentioned the speed with which a tweet spreads through the Twitter network and points out possible ways of capturing the speed. One could monitor all followers of a user (or, if possible, even the followers' followers) for if and when they retweet a tweet. Thus, the exact speed with which tweets spread could be analyzed as well as the influence of single users - e.g., whether retweets are performed mainly by the same users or sporadically by many different users. An exhaustive graph analysis of the "following" relationships might yield additional valuable insights into the factors that make people retweet. In this thesis, however, it was practically unfeasible to crawl all followers of the users due to the Twitter REST API rate limit (for example, the 15,000 users in our dataset have a total of 237 million followers). If elevated Twitter access could be obtained, more data could be collected for analytical purposes.

Another aspect of tweet virality are the retweet cascades, i.e., the sequence in which a tweet gets retweeted through multiple users. If those cascades could be monitored together with the users that retweeted them, differentiated spread paths could be revealed. It seems very likely that hugely popular and influential Twitter users can make a tweet viral by referencing it, thereby exposing it to a larger number of other users. While tweets might become popular without being retweeted by such influential users that can multiply a tweet's spread, identifying the factors that make such users retweet a message can prove very useful.

A closer investigation of the follower network could refine the techniques of [HRW08] to determine the subset of followers that users actually communicate with and an-

alyze whether different retweet patterns emerge between persons of the subset and the rest of the followers. Since the "following" relationship is directed, the resulting graph can be interpreted as an interest graph of a user that might yield clues on the tweets specific users will likely retweet. An evaluation of how different types of users communicate with and retweet each other is given in [WHMW11].

Crawling of a substantial portion of a user's followers might also provide different results for the analysis on the Jaccard distance between hashtags of users and their followers (see Section 4.1). Although no usable results were obtained for the current data, this might be due to the fact that only few followers of each user were crawled. A bigger portion of followers might change the average distances and average numbers of retweets significantly. Alternatively, a comparison of all words (with a bag-of-words-approach) used in the users' tweets as opposed to just the hashtags might yield different results.

In our analysis on URLs in the tweets, we confined our attention on the number of links that were used. A possible extension of this analysis could encompass features of the links, such as whether an URL shortener was used, the type of the linked document (e.g., HTML, PDF, JPG), the number of words used, top level domain, and web site popularity.

The initial investigations of Chapter 7 are part of further work: the average error for the prediction of expected retweets could be further reduced. Any improvement of the general liner model, for example by additional feature analyses or a careful reexamination of how features are discretized (listed in Table 6.2 in Chapter 6) is also likely to improve the prediction of expected retweets. Another mitigation step for probability values with much scattering of the number of retweets is to build another prediction model with a higher threshold T. Tweets whose probability value in the first model fall into a probability range where the numbers of retweets are difficult to approximate due to high scattering might have a probability value in the second model that is allows a more reliable prediction of the number of retweets the tweet will incur. Of course, the prediction accuracy would have to be Bonferroni-corrected: the error rate increases since the error rates for both (binary) prediction models are combined.

Yet another interesting aspect of further work addresses the semantic content of tweets. It is conceivable that factual statements undergo a different retweet process than personal opinions. Furthermore, different topics might also impact the retweet rate: for some people, politics might not be as interesting as celebrity gossip. [HAN+11] already performed research on different effects of negative sentiment in news and social tweets on their virality and discovered that negative sentiment enhances the virality of news-related tweets, but not of the non-news tweets.

9 Conclusion

In this thesis, we show that a prediction of viral tweets is possible. To achieve best prediction results, it is not enough to consider structural, content-based, or sentimental aspects in isolation. Rather, a combination of features covering all these aspects is important to make reliable predictions. Also, it is important to make as few over-simplifying assumptions concerning feature interdependencies as possible. Our generalized linear model generally outperforms a Naive Bayes one which assumes conditional independence between features for different thresholds concerning the classification of viral tweets.

The evaluation of the generalized linear model shows that the number of followers has the highest impact on the prediction of viral tweets, followed by URLs, mentions and hasthags in tweets. Other non-obvious but still important factors for the prediction are the tweet length and sentiment (sentiment valence and emotional divergence as well as the individual sentiment scores) expressed in a tweet. For instance, we find that a tweet is more likely to be retweeted when connotated with a negative sentiment. The generalized linear model also shows that while accurate predictions of viral tweets can be made based on the number of followers alone, substantial improvements are still possible when further features are included. We expect these findings to generalize across different social networks and microblog platforms, yet an extensive analysis is needed to verify this hypothesis.

We also show a way to harness the prediction model in order to estimate the number of retweets a tweet will induce. We are positive that future work on this topic will provide a better approximation of the number of retweets that a tweet is expected to induce.

In any case, we hope that our findings will broaden the view and ignite new discussions on the spread of information in today's web.

Bibliography

[ABL12] Paul André, Michael S. Bernstein, and Kurt Luther. Who gives a tweet?: Evaluating microblog content value. In *CSCW*. ACM, 2012.

[AH10] Sitaram Asur and Bernardo A. Huberman. Predicting the future with social media. In *CoRR*. IEEE, 2010.

[BGL10] Danah Boyd, Scott Golder, and Gilad Lotan. Tweet, tweet, retweet: Conversational aspects of retweeting on twitter. In *HICSS*. IEEE Computer Society, 2010.

[BHMW11] Eytan Bakshy, Jake M. Hofman, Winter A. Mason, and Duncan J. Watts. Everyone's an influencer: Quantifying influence on twitter. In *WSDM*. ACM, 2011.

[BMP11] Johan Bollen, Huina Mao, and Alberto Pepe. Modeling public mood and emotion: Twitter sentiment and socio-economic phenomena. In *ICWSM*. The AAAI Press, 2011.

[BO69] Jerry Boucher and Charles E. Osgood. The pollyanna hypothesis. In *Journal of Verbal and Learning Behavior*, 1969.

[CHBG10] Meeyoung Cha, Hamed Haddadi, Fabricio Benevenuto, and Krishna P. Gummadi. Measuring user influence in twitter: The million follower fallacy. In *ICWSM*, 2010.

[GGS11] David Garcia, Antonios Garas, and Frank Schweitzer. Positive words carry less information than negative words. *CoRR*, 2011.

[GGSS12] Antonios Garas, David Garcia, Marcin Skowron, and Frank Schweitzer. Emotional persistence in online chatting communities. *Scientific Reports*, 2012.

[HAN+11] Lars Kai Hansen, Adam Arvidsson, Finn Nielsen, Elanor Colleoni, and Michael Etter. Good friends, bad news - affect and virality in twitter. *CoRR*, 2011.

[HRW08] Bernardo A. Huberman, Daniel M. Romero, and Fang Wu. Social networks that matter: Twitter under the microscope. *CoRR*, 2008.

[KKT03] David Kempe, Jon M. Kleinberg, and Éva Tardos. Maximizing the spread of influence through a social network. In *KDD*. ACM, 2003.

[KLPM10] Haewoon Kwak, Changhyun Lee, Hosung Park, and Sue Moon. What is Twitter, a social network or a news media? In *WWW*. ACM, 2010.

[LWLC12] Thomas Lansdall-Welfare, Vasileios Lampos, and Nello Cristianini. Effects of the recession on public mood in the uk. In *WWW*. ACM, 2012.

[PGS12] René Pfitzner, Antonios Garas, and Frank Schweitzer. Emotional divergence influences information spreading in twitter. In *ICWSM*. The AAAI Press, 2012.

[POL11] Sasa Petrovic, Miles Osborne, and Victor Lavrenko. Rt to win! predicting message propagation in twitter. In *ICWSM*. The AAAI Press, 2011.

[PRU06] Gopal Pandurangan, Prabhakar Raghavan, and Eli Upfal. Using pagerank to characterize web structure. *Internet Mathematics*, 2006.

[Ris01] Irina Rish. An empirical study of the naive bayes classifier. In *IJCAI*, 2001.

[Shi03] Clay Shirky. Power laws, weblogs, and inequality. *http://shirky.com/writings/powerlaw_weblog.html*, 2003.

[SHPC10] Bongwon Suh, Lichan Hong, Peter Pirolli, and Ed H. Chi. Want to be retweeted? large scale analytics on factors impacting retweet in twitter network. In *SocialCom/PASSAT*. IEEE Computer Society, 2010.

[TBP+10] Mike Thelwall, Kevan Buckley, Georgios Paltoglou, Di Cai, and Arvid Kappas. Sentiment in short strength detection informal text. *JASIST*, 2010.

[TBP11] Mike Thelwall, Keven Buckley, and Georgios Paltoglou. Sentiment in twitter events. *JASIST*, 2011.

[TBP12] Mike Thelwall, Kevan Buckley, and Georgios Paltoglou. Sentiment strength detection for the social web. *JASIST*, 2012.

[WHMW11] Shaomei Wu, Jake M. Hofman, Winter A. Mason, and Duncan J. Watts. Who says what to whom on twitter. In *WWW*. ACM, 2011.

[YC10] Jiang Yang and Scott Counts. Predicting the speed, scale, and range of information diffusion in twitter. In *ICWSM*. The AAAI Press, 2010.

[ZHVGS10] Tauhid R. Zaman, Ralf Herbrich, Jurgen Van Gael, and David Stern. Predicting information spreading in twitter. *NIPS*, 2010.