Analyzing and Predicting Viral Tweets

Maximilian Jenders
Hasso Plattner Institute
Prof.-Dr.-Helmert-Str. 2-3
Potsdam, Germany
maximilian.jenders
@hpi.uni-potsdam.de

Gjergji Kasneci
Hasso Plattner Institute
Prof.-Dr.-Helmert-Str. 2-3
Potsdam, Germany
gjergji.kasneci
@hpi.uni-potsdam.de

Felix Naumann
Hasso Plattner Institute
Prof.-Dr.-Helmert-Str. 2-3
Potsdam, Germany
felix.naumann
@hpi.uni-potsdam.de

ABSTRACT

Twitter and other 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 paper addresses important questions concerning the spread of information on Twitter. What makes Twitter users retweet a tweet? Is it possible to predict whether a tweet will become "viral", i.e., will be frequently retweeted? To answer these questions we provide an extensive analysis of a wide range of tweet and user 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. All experiments are performed on a real-world dataset, extracted through a public Twitter API based on user IDs from the TREC 2011 microblog corpus.

Categories and Subject Descriptors

H.4 [**Information Systems Applications**]: Miscellaneous-Collaborative and social computing systems and tools

Keywords

Prediction; model; microblog; Twitter; tweet; retweet; spread; analysis

1. VIRAL TWEETS

Social networks, offering microblogging services, such as Twitter, Google+, Facebook, etc., enable the rapid spread of information from a handful of users to hundreds of millions of people around the globe. It is not surprising that many of these platforms are being sentineled and harnessed by various organizations with the goal of discovering the most recent news or sentiments regarding products, companies, events, socio-political issues, celebrities, and many more. Understanding the main factors that make certain microblogs "viral", i.e., spread quickly and widely, in these kinds of networks can be crucial for the analysis of public opinion formation and many other opinion mining tasks; insights gained here can be used to implement prediction tools for advertisers, opinion leaders, and decision makers. For example, imagine a popular product that is to be extended by a new feature. In such a case, it would be helpful to infer from beta versions of the product, as early on as possible, what the general opinion on the introduced feature will be. This knowledge would be essential for deciding whether the new feature should be maintained or omitted. A more abstract question is whether it is possible to design viral microblogs, which could shape the general opinion on a certain product, celebrity, or socio-political matter. In a society where the web is becoming the prime source of information, we should be aware of biases and manipulation strategies embodied in pieces of information that are widely spread throughout the web.

This work analyzes and highlights different factors that lead to the quick spread of information in social networks. The focus of the paper is on Twitter as a platform representative of many social microblogging services enabling common users, celebrities, institutions, and companies to publish text messages of restricted length. On Twitter, these messages are called *tweets* and are restricted to 140 characters each. Tweets are typically about recent news, status updates, or comments on specific topics. A user can re-post, i.e., "retweet", tweets from other users and may also choose to "subscribe" to another user's tweets by becoming a "follower" of that user. Despite the restricted length of tweets, Twitter's retweet mechanism and its social network, consisting of the "follower" relationships, provide an unprecedented mechanism for the spread of information in form of tweets.

In this paper, we analyze this mechanism and present our findings:

- We start out by analyzing "obvious" tweet and user features (e.g., number of followers, tweet length, number of hashtags, number of mentions, etc.) with respect to their impact on and correlation with the retweet frequency.
- We then analyze "latent" features, such as the sentiment and emotional divergence contained in a tweet, and unveil their correlations with the retweet frequency of the tweet.
- Finally we address the question whether viral tweets can be predicted. We answer this question affirmatively and provide a comparison between a Naive Bayes model, which assumes conditional feature independence, and a generalized linear model that avoids such simplifying assumptions.

WWW 2013 Companion, May 13–17, 2013, Rio de Janeiro, Brazil.
ACM 978-1-4503-2038-2/13/05.

The paper is structured in following way: an overview of related work is given in Section 2. In Section 3, we present the extraction of the dataset, on which all subsequent analyses and experimental evaluations were performed. Section 4 provides a broad analysis of different tweet and user features and their impact on the number of retweets. A model that predicts viral tweets by exploiting the analyzed features is presented in Section 5. An extensive evaluation of the model is carried out in Section 6.

2. RELATED WORK

Prior work has already gained many insights on social networks in general as well as for Twitter in particular.

Research concerning the spread of information on Twitter can be categorized into three broader topics. We address these topics in the following paragraphs.

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 are used to derive influence scores, e.g., by exploiting PageRank [9], or by predicting the propagation of such scores through the network, e.g., [8]. For example. The PageRank score of a user can be estimated by the number of followers of that user [9]. However, the authors also report that ranking users by their PageRank scores is different from ranking users by the total number of retweets they obtain, thus showing that link and retweet analysis capture the influence of users in two different ways. Another work analyzes the influence of users by measuring three different parameters: the number of followers, the number of retweets, and the number of mentions [5]. The authors observe that having a high number of followers does not necessarily lead to an increase in the number of retweets or mentions. Furthermore, influential users seem to be those who focus on specific topics.

Although the most influential users on Twitter are in general those who have already been most influential in the past and who have a large number of followers, methods that try to predict which specific user will generate influential tweets are relatively unreliable [2]. This suggests that a more detailed analysis of tweets and users is needed to complement structural measures. Finally, link-analysis algorithms were explored to identify subsets of social network users who should be addressed in order to trigger a quick and wide spread of information [8]. As our data set contained only a very partial relationship graph between all users, such link analysis techniques were unfeasible for the purpose of this work.

Content analysis. Another stream of research has analyzed the influence of tweets by examining their content. The question of why and how people retweet is analyzed; varying styles of retweets are highlighted [4]. Other work reports that the use of URLs and hashtags in a tweet affects the total number of retweets that the tweet incurs [15]. The authors of [1] let users rate tweets on a website and found out 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 status updates. Another approach employs structural features, such as PageRank, as well as content and metadata features to predict the spread of tweets [7]. However, no detailed analyses of the features is presented. Furthermore, collaborative filtering techniques on structural and content features were used to rank users based on their likelihood of retweeting a given tweet [18]. Other content features used for predicting reaction to a tweet include the amount of verbs, nouns, and adjectives in tweets [14]. In [12], a variety of features, including even punctuation characters, such as exclamation and questions marks, were employed to determine likelihood of a retweet. However, a deeper analysis of the features was not given.

Sentiment analysis. Other work has focused on the sentiments analysis of tweets. In an effort to automatically classify sentiments in social networks, the authors of [17] designed *SentiStrength*[1], an algorithm for extracting sentiment strength 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 comments, revealing an impressive accuracy concerning the identification of sentiments. An improved version of the algorithm shows that SentiStrength is robust enough to be applied to a wide variety of different social web contexts [16]. A fine-grained version of this algorithm outperformed other state-of-the-art approaches for German socia media texts [11]. The SentiStrength algorithm has also been used to infer sentiments in tweets [13]. The authors analyzed how the inferred sentiments relate to the retweet probability of a tweet. More specifically, tweets were classified into predominantly positive, negative, and neutral tweets. For each of these categories, the retweet distribution was analyzed. Interestingly, the authors report that in each category, the fraction of tweets is similar to that of retweets. 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 [13] is that of *emotional divergence*, which combines the positive and negative score of a tweet, and also 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 [13].

Other work analyzes six different mood states derived from tweets and relates them to records of popular events, such as socio-political, cultural, and economic events [3]. The authors report that such events have immediate effect on the sentiment expressed in tweets. This is supported by [10], showing 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.

Finally, the emotional expression of users, as derived from their messages, was analyzed, finding that the emotional expression of an individual user persists over a long period of time [6].

In contrast to prior and any of the mentioned related work, we deeply analyze a broad spectrum of user- and tweet-related features and combine structural, content-based and sentimental aspects to predict viral tweets.

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

3. DATASET EXTRACTION

As part of the TREC 2011 microblog track[2], a representative sample of the so-called "Twittersphere" containing identifiers for approximately 16 million tweets and the users who posted them was published under the name "Tweets2011 corpus"[3]. The initiative was led by the National Institute of Standards and Technology (NIST). As the Twitter terms-of-service permit only the distribution of identifiers for users and tweets, we relied on the public Twitter REST API to retrieve more data based on the identifiers provided by the above corpus.

For this study, we used the Twitter4J library[4] to handle all API calls. The results were subsequently written into a database. It should be noted that the Twitter REST API has a rate limit of 350 requests per hour. Different kinds of request types were used to retrieve different kinds of information. For example, one request type was used to retrieve information about up to 100 users; another one to fetch 200 of the last 3,200 tweets of a user, and yet another one was used to obtain 5,000 identifiers of a user's followers.

Using the above API, the following information about users and tweets can be crawled through a sequence of API calls:

- 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 is a retweet of another tweet.

After analyzing the distribution of followers, we employed stratified random samling to chose user identifiers. For that reason, we categorized randomly chosen user identifiers of the TREC corpus based on their number of followers into the following nine different groups: 0 - 9, 10 - 49, 50 - 99, 100 - 299, 300 - 999, 1,000 - 4,999, 5,000 - 49,000, 50,000 - 499,999, and more than 500,000 followers. The intuition behind this categorization is that different numbers of followers may differently impact the number of retweets. From each group, we sampled 100 users whose tweets were manually verified to be in English. Additionally, we added a random sample of their followers who also tweeted in English.

During April and May 2012, information about users and tweets was constantly recrawled in a least-recently-crawled fashion. As the number of retweets for a tweet changes over time, older tweets had to be re-crawled as well to keep their number of retweets up to date. Each tweet was stored in a database 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.

Overall, more than 21.8 million unique *pure* tweets (i.e., tweets that are no retweets) 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 [13]) may be the result of using the official Twitter classifier for tweets and retweets, rather than relying on self-developed methods.

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 paper. 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.

[2]https://sites.google.com/site/microblogtrack/
[3]http://trec.nist.gov/data/tweets/
[4]http://twitter4j.org/en/index.html

4. FEATURES INFLUENCING RETWEET FREQUENCY

In order to identify features that influence how tweets get retweeted and thus spread in the Twitter network, we have analyzed a wide variety of tweet and user features. Here, we present only the analysis of those features that were most impactful with regard to the retweet frequency of tweets.

This is the first work that deeply looks into such a broad spectrum of features, encompassing structural, textual, and sentiment analysis.

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 enjoy 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, Figure 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 to average over users having similar number of followers, hence the log-log scale of the plot.

Tweet Length.

Another feature worth analyzing is the textual length of a tweet. An intuitive assumption here is that 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 2 depicts the average retweet count for all possible tweet lengths. 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. For up to 120 characters, the expected number of retweets seems to grow almost proportionally with the tweet length. However, this trend is reversed for tweets with more than 120 characters. An explanation for this finding might be that because of the restricted length of tweets, experienced users typically try to condense their message in the tweet, thereby using less than 140 characters, whereas the complete available message length is utilized only by relatively new users.

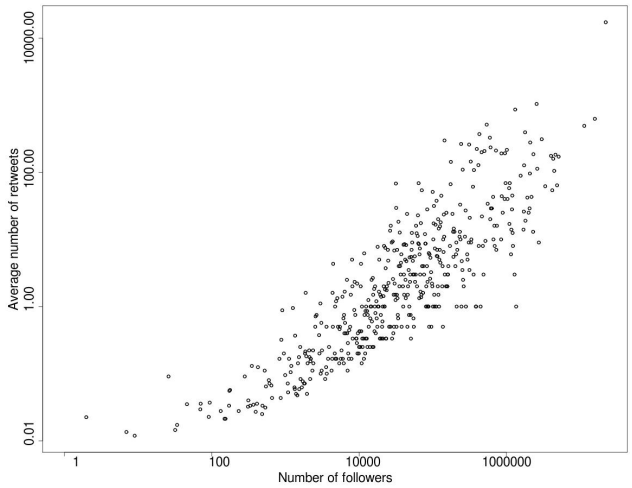

Figure 1: Log-log scale plot of the average number of retweets per tweet in relation to the number of followers.

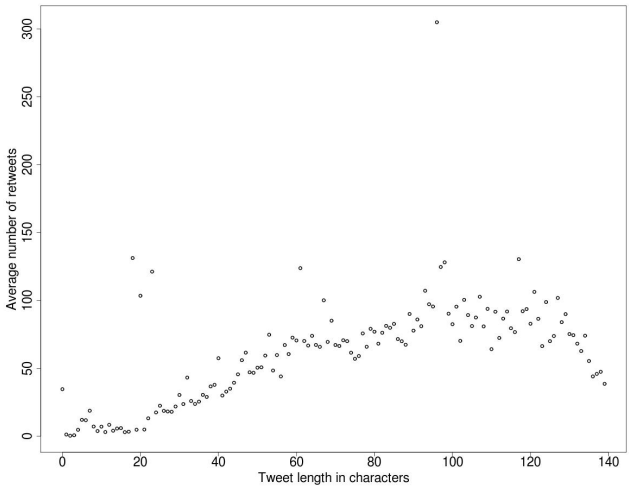

Figure 2: Average number of retweets in relation to the tweet length.

Hashtags and Mentions.

Two other textual parameters that we analyzed are the hashtags and the mentions contained in a tweet. Hashtags are basically keywords that are used to tag tweets so they can be more easily categorized and found by users. Additionally, hashtags can be used to supplementary provide meaning. Given these use cases of hashtags, one would expect that interesting tweets are more likely to contain hashtags, thus enabling users to find and talk about them as well as better understand them. Indeed, the plot of Figure 3 shows that tweets containing 1 to 3 hashtags are more likely to be retweeted than tweets without 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, leaving less space for the actual information.

Analogously to hashtags, mentions are used to tag and address Twitter users whose names occur in a tweet. As in the case of hashtags, using some mentions increases the number of retweets a tweet receives on average, whereas larger number of mentions decreases the average number of retweets, see Figure 4.

Sentiment Analysis.

Another factor possibly influencing 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. To infer the sentiment of a tweet, we used the *SentiStrength* algorithm [17]. For a given input tweet, SentiStrength returns two scores representing respectively the positive and negative sentiment expressed by the tweet. This score ranges from -1 to -5 for negative and 1 to 5 for positive sentiments, with 1 (or -1) standing for a statement void of sentiment and 5 (or -5) for a statement affiliated with the utmost sentiment.

Our data			
tweet type	negative	neutral	positive
pure tweets	17.8%	37.68%	44.48%
retweets	22.4%	37.13%	40.49%
Data from [13]			
tweet type	negative	neutral	positive
pure tweets	19.9%	33.8%	46.3%
retweets	19.8%	31.4%	48.8%

Table 1: Pure tweets and retweets in relation to sentiment valences for out data and as reported in [13].

Sentiment Valence. Pfitzner et al. [13] already employed SentiStrength to analyze the distributions of tweets and retweets according to their *sentiment valence* (i.e., whether they predominantly have a positive, neutral, or negative sentiment). They show that in each of these categories, the fraction of tweets resembles that of retweets. However, the dataset used in [13] has an average of ten pure tweets per user, while our dataset has an average of 1,453 pure tweets per user and a much higher percentage of retweets. We therefore were interested to find out whether the findings of [13] would hold for our data.

Table 1 shows the proportion of pure tweets and retweets in relation to their sentiment valence in our data and as reported in [13]. As it can be seen, in our data, tweets with a positive valence make up for the largest group of all tweets and 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 Pfitzner et al.

Interestingly, our findings imply that tweets with predominantly negative sentiment have a higher retweet probability. By applying Bayes' Theorem, the probability of a retweet

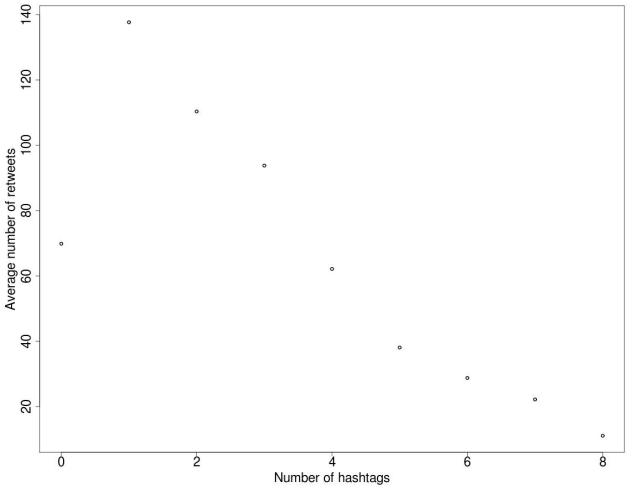

Figure 3: Average number of retweets in relation to the number of hashtags used in the tweet.

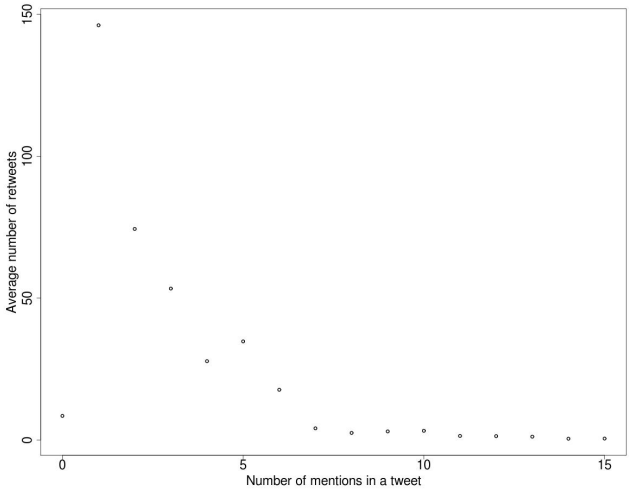

Figure 4: Average number of retweets in relation to the number of mentions used in the tweet.

given a negative sentiment valence is

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

The conditionals in the above formula can be estimated by their maximum likelihood estimations, which results in:

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

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

$$P(retweet|neutral) \approx 0.160$$

$$P(retweet|positive) \approx 0.150$$

As 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 typically relate to negative experiences) are more likely to attract the attention of other users, thus increasing the visibility and the retweet probability.

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 [13]. For a given tweet t, it can be calculated as

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

resulting in a value between 0.2 and 1.0. This score aims at capturing the strength of sentiments in tweets, regardless of their valence and, analogously to the sentiment valence, can be used to calculate retweet probabilities.

The authors of [13] 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 2. 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.

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 2: Emotional divergence scores for different fractions of pure tweets and retweets (in relation to number of total tweets and retweets, respectively.

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 5 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

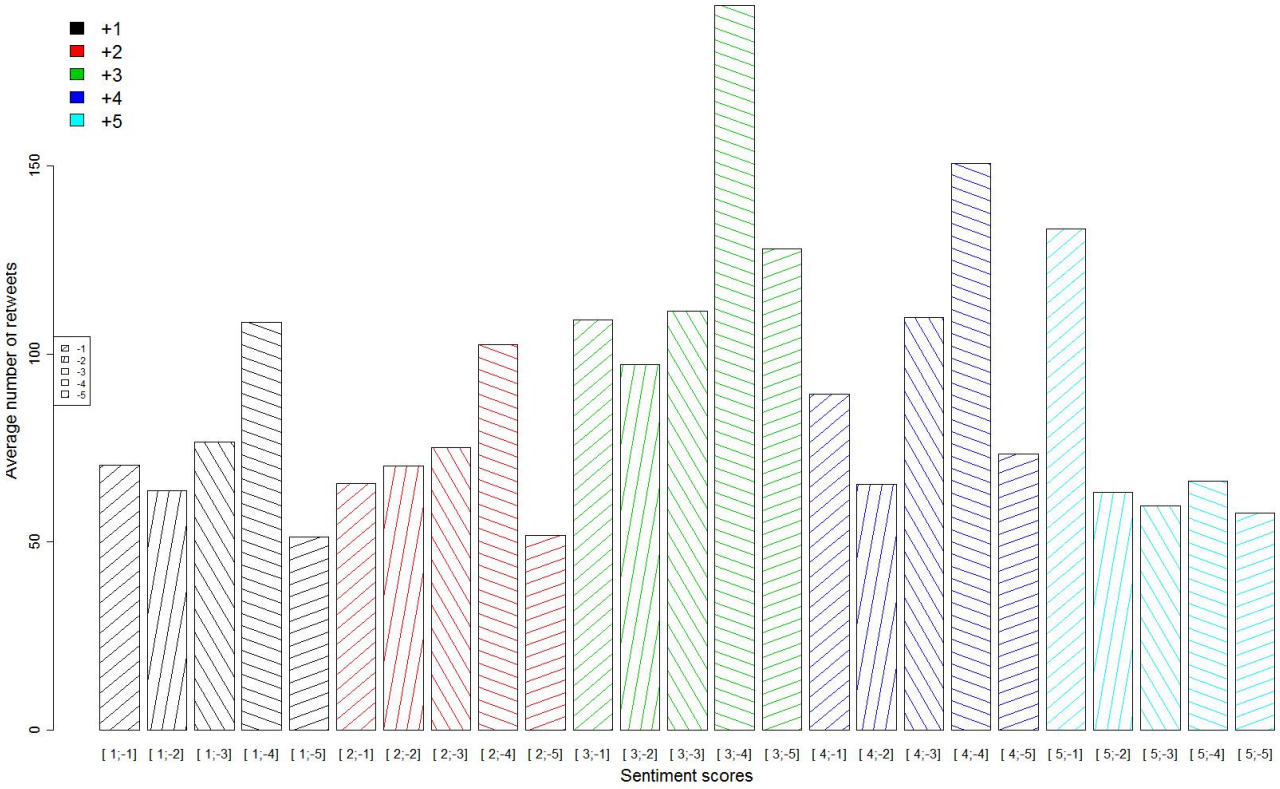

Figure 5: Average number of retweets for individual sentiment scores. The sentiments are ordered from +1 to +5 and, for each positive score (indicated by different colors), by negative score from -1 to -5 (indicated by different shading).

Twitter community. Generally, tweets with a sentiment of +3 seem to incur larger number of retweets.

5. PREDICTING VIRAL TWEETS

In the previous section, we analyzed the impact of different tweet and user features on the expected retweet frequency. Here, we propose different probabilistic models for predicting whether a given tweet is viral (i.e., whether it will be more frequently retweeted than a certain threshold T) by taking into account all the features we have discussed so far.

Our first proposal is a generalized linear model. Without going into details, the model is trained on a set of tweets (i.e., user and tweet features for each tweet $\mathbf{x}$, specified through a feature vector $\boldsymbol{\phi}(\mathbf{x}) = (x_1, ..., x_n)$) and the information which tweets received more retweets than a predefined threshold T. Based on this data, the model learns weights w_i for each feature and, given a tweet's feature vector, it is able to calculate a "virality" score $v(\mathbf{x}) = \sum_{i=1}^{n} w_i x_i$. This score, together with a feature-independent intercept weight w_0, can then be applied to a general sigmoid activation function f (here, we use the logistic function) to determine the probability that the number $N_R(\mathbf{x})$ of retweets that a tweet receives is higher than the threshold T:

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

The strength of this 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, the number of hashtags in a tweet and the tweet length seem to depend on each other, as hashtags consume characters. To showcase the implications of simplifying independence assumptions, we introduce a Naive Bayes model. This model calculates a joint probability of $N_R(\mathbf{x}) > T$ and the feature vector $\boldsymbol{\phi}(\mathbf{x})$ for a tweet $\mathbf{x}$ as:

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

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:

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

The conditionals $P(x_i \mid N_R(\mathbf{x}) > T)$ as well as the probability $P(N_R(\mathbf{x}) > T)$ are estimated by their maximum likelihood estimations.

T	F-Measure NB	F-Measure GLM
50	0.916	0.936
100	0.927	0.940
500	0.947	0.963
1000	0.951	0.968

Table 3: F-Measure scores for the Naive Bayes (NB) and the generalized linear model (GLM) for different thresholds T.

6. EVALUATION

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

Table 4: Learned feature weights for the generalized linear model and a threshold $T = 1000$.

To train and evaluate our prediction model, we used the Weka toolkit[5] on a random sample from our dataset. For every tweet in this sample, we knew the total number of retweets as well as all features used in our model. Note that the generalized linear model can learn only from values that are often repeated. Hence, each feature domain was discretized in a reasonable way, 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 and those with four hashtags formed another group, and tweets with five or more hashtags made up the last group.

Next, we chose different thresholds T to separate viral from non-viral tweets. As a measure for the virality of a tweet, we chose the number of retweets that the tweet induces and then labeled tweets which number of retweets was greater than T as viral. A preceding analysis showed that the retweet distribution over the tweets follows a Pareto distribution. For example, only 4% of all tweets receive more than 50 retweets. Hence, we initially chose T to be 50. We expected higher thresholds to lead to better prediction accuracy, since for tweets with many retweets, the feature values should be more discriminative. To test this hypothesis, we ran the experiments with $T \in \{50, 100, 500, 1{,}000\}$. Both the Naive Bayes and the generalized linear model, were tested using tenfold cross-validation. Table 3 shows the resulting F-Measure scores. For all T, the random sample contained equal amounts of viral and non-viral tweets to better keep the scores comparable. The evaluations were executed on a sample of 20,000 tweets, with equal numbers of tweets labeled as viral and non-viral in the ground truth.

Table 3 shows the resulting F-Measure scores, showing that the generalized linear model generally provides more reliable predictions than the Naive Bayes model. As discussed in Section 5, this can be attributed to the fact that the generalized linear model does not assume conditional independence between the features. Note that the F-Measure score 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 and URLs. The weights for the discretized feature values as learned by the generalized linear model for a threshold $T = 1{,}000$ are given in Table 4.

7. CONCLUSION

In this paper, we gave an extensive analysis of "obvious" and "latent" tweet and user features with respect to their impact on the spread of tweets. For reliable prediction of viral tweets, it is not enough to consider structural, content-based, or sentimental aspects in isolation. Rather, a combination of features covering all these aspects and a learning model that avoids simplifying independence assumptions are the key to high prediction quality. We expect these findings to generalize across different social networks and microblog platforms, yet an extensive analysis is needed to verify this hypothesis.

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.

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

8. REFERENCES

[1] P. André, M. S. Bernstein, and K. Luther. Who Gives A Tweet? Evaluating Microblog Content Value. In *Proceedings of the 16th ACM Conference on Computer Supported Cooperative Work and Social Computing*, CSCW, pages 471–474, 2012.

[2] E. Bakshy, J. M. Hofman, D. J. Watts, and W. A. Mason. Everyone's an Influencer : Quantifying Influence on Twitter. In *Proceedings of the 4th ACM International Conference on Web Search and Data Mining*, WSDM, pages 65–74, 2011.

[3] J. Bollen, A. Pepe, and H. Mao. Modeling Public Mood and Emotion: Twitter Sentiment and Socio-economic Phenomena. In *Proceedings of the Fifth International AAAI Conference on Weblogs and Social Media*, ICWSM, pages 450–453, 2011.

[4] D. Boyd, S. Golder, and G. Lotan. Tweet, Tweet, Retweet: Conversational Aspects of Retweeting on Twitter. In *Proceedings of the 43rd Hawaii International Conference on System Sciences*, number 6 in HICSS, pages 1–10, 2010.

[5] M. Cha and K. P. Gummadi. Measuring User Influence in Twitter : The Million Follower Fallacy. In *Proceedings of the 4th International AAAI Conference on Weblogs and Social Media*, ICWSM, 2010.

[6] A. Garas, D. Garcia, M. Skowron, and F. Schweitzer. Emotional Persistence in Online Chatting Communities. *Scientific Reports*, 2, 2012.

[7] L. Hong, O. Dan, and B. D. Davison. Predicting popular messages in Twitter. In *Proceedings of the 20th international conference companion on World wide web*, WWW, 2011.

[8] D. Kempe, J. Kleinberg, and E. Tardos. Maximizing the Spread of Influence through a Social Network. In *Proceedings of the 9th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, volume 41 of *KDD*, 2003.

[9] H. Kwak, C. Lee, H. Park, and S. Moon. What is Twitter, a Social Network or a News Media? In *Proceedings of the 19th International Conference on World Wide Web*, number 2 in WWW, pages 591–600, 2010.

[10] T. Lansdall-Welfare, V. Lampos, and N. Cristianini. Effects of the Recession on Public Mood in the UK. In *Proceedings of the 21st International Conference Companion on World Wide Web*, WWW, pages 2–7, 2012.

[11] S. Momtazi. Fine-grained German Sentiment Analysis on Social Media. In *Proceedings of International Conference on Language Resources and Evaluation*, LREC, pages 1215–1220, 2012.

[12] N. Naveed, T. Gottron, J. Kunegis, and A. C. Alhadi. Bad News Travel Fast : A Content-based Analysis of Interestingness on Twitter. In *Proceedings of the 3rd International Conference on Web Science*, WeSci, 2011.

[13] R. Pfitzner, A. Garas, and F. Schweitzer. Emotional Divergence Influences Information Spreading in Twitter. In *Proceedings of the 6th International Conference on Weblogs and Social Media*, ICWSM, 2012.

[14] M. Rowe, S. Angeletou, and H. Alani. Predicting Discussions on the Social Semantic Web. In *Proceedings of the 8th Extended Semantic Web Conference*, ESWC, pages 405–420, 2011.

[15] B. Suh, L. Hong, P. Pirolli, and E. H. Chi. Want to be Retweeted? Large Scale Analytics on Factors Impacting Retweet in Twitter Network. In *Proceedings of the 2nd IEEE International Conference on Social Computing*, SOCIALCOM, pages 177–184, 2010.

[16] M. Thelwall, K. Buckley, and G. Paltoglou. Sentiment Strength Detection for the Social Web. *Journal of the American Society for Information Science and Technology*, 63(1):163–173, 2012.

[17] M. Thelwall, K. Buckley, G. Paltoglou, D. Cai, and A. Kappas. Sentiment in Short Strength Detection Informal Text. *Journal of the American Society for Information Science and Technology*, 61(12):2544–2558, 2010.

[18] I. Uysal and W. B. Croft. User Oriented Tweet Ranking : A Filtering Approach to Microblogs. In *Proceedings of the 20th ACM International Conference on Information and Knowledge Management*, CIKM, pages 2261–2264, 2011.