Bad News Travel Fast: A Content-based Analysis of Interestingness on Twitter

Nasir Naveed Thomas Gottron Jérôme Kunegis Arifah Che Alhadi
WeST – Institute for Web Science and Technologies
University of Koblenz-Landau
{naveed,gottron,kunegis,alhadi}@uni-koblenz.de

ABSTRACT

On the microblogging site Twitter, users can forward any message they receive to all of their followers. This is called a retweet and is usually done when users find a message particularly interesting and worth sharing with others. Thus, retweets reflect what the Twitter community considers interesting on a global scale, and can be used as a function of interestingness to generate a model to describe the content-based characteristics of retweets. In this paper, we analyze a set of high- and low-level content-based features on several large collections of Twitter messages. We train a prediction model to forecast for a given tweet its likelihood of being retweeted based on its contents. From the parameters learned by the model we deduce what are the influential content features that contribute to the likelihood of a retweet. As a result we obtain insights into what makes a message on Twitter worth retweeting and, thus, interesting.

Categories and Subject Descriptors

H.4 [**Information Retrieval**]: Data Mining; D.2.8 [**Text Analysis**]: Metrics—*content quality, ranking measure*

General Terms

Content Quality Measure, Algorithms

Keywords

Microblog, Twitter, Retweet Ranking, Topic Modeling, Sentiment Analysis

1. INTRODUCTION

A microblogging platform such as Twitter allows the users to share information via short messages. A posted message is distributed to those people that subscribe to the information of the author. In the context of Twitter this subscription is known as *following*. This structure of followers forms a large network among the users of Twitter. A particularity is that the receiver of a message has the option to relay it and forward it to her followers. This practice is called *retweet* and normally users will forward a message if they consider it interesting and worth sharing with others.

The question of what causes a message to be retweeted has frequently been addressed, but mainly in a scenario of retweet prediction for a given user and with a focus on the structure of the social network [4, 9, 13]. In this case a typical observation is that a well connected user with active followers is more likely to be retweeted. As the content of a tweet in such a setting is neglected or reduced to a few very simple features, a network-based analysis of retweets may give hints into *who* tends to write interesting messages, but cannot give insights into *what* the community is interested in. Therefore, in this paper, we focus on the content of a tweet and train a prediction model to forecast for a given tweet its likelihood of being retweeted based purely on its contents. From the parameters learned by the model we deduce what are the influential content features that contribute to the likelihood of a retweet – and thereby are characteristics of an interesting message in the context of Twitter.

For this purpose, we analyze a set of high- and low-level content-based features on a large collection of Twitter messages. The low-level features comprise the words contained in a tweet, the tweet being a direct message, the presence of URLs, hashtags, usernames, emoticons, and of question and exclamation marks as well as terms with a strong positive or negative connotation. These features are directly extracted from the text of a message and do not require further processing. The high-level features are formed by associating tweets to topics and by determining the sentiments of a tweet. For prediction we employ a logistic regression analysis model, which is trained and verified on large datasets.

Thereby, in this paper we make two contributions:

- We consider the problem of learning which tweets are retweeted, based on a wide range of content features and independently of context information such as the user's position in the social network and the timestamp of a tweet. We show that it is possible to predict which tweets are retweeted.
- By analyzing the parameters learned in our prediction model, we identify the features that contribute most strongly to the probability of a tweet being retweeted. This allows for a deeper insight into what is of interest in the Twitter community.

Outline We begin by reviewing related approaches for analyzing tweets and give a small overview of Twitter datasets in Section 2. In Section 3 we then study the problem of content-based retweet prediction. The prediction model is trained and evaluated in Section 4. There, we also discuss in detail the results and interpret

the model parameters to deduce the factors rendering a message interesting on Twitter. Finally, in Section 5 we discuss our results and conclude the paper with an outlook at possible applications and future work.

2. BACKGROUND AND RELATED WORK

Twitter is a microblogging service founded in 2006 that allows more than 200 million[1] users to share *tweets* with each other: short messages of no more than 140 characters. Users can *follow* other users in order to receive their tweets. If a user considers a tweet interesting, she may forward it to her own followers. This practice is called *retweeting* and usually users retweet the content of general interest or concerned with the audience who follows their tweets [2]. By convention, retweets are indicated by specific keywords such as RT and via. The purpose of retweeting is often to disseminate information to one's followers. According to Kwak et al. [9], any retweeted tweet can be expected to reach an average of 1,000 users no matter how many followers the first tweeting user had.

2.1 Analysis of Tweets

Twitter has attracted much research in recent years. Some studies measured the influence of twitterers based on the social network, e.g. using PageRank, the number of followers, the number of retweets and trending topics [4, 6, 9, 13, 14, 15]. These studies look into the correlation between the number of followers and influence. A common finding based on work by [4, 13] is that popular users with large number of followers do not necessarily have more influence. This reveals that the popularity of a user does not automatically imply a higher influence in Twitter.

However this finding is contradicted by [6, 9, 14]. These works argue that the context of a tweet (the twitterer's social graph, the time of the tweet, etc.) does influence the likelihood of the tweet being retweeted. Suh et al. [14] state that contextual features include basic aspects of the graph structure, e.g the number of followers and followees (people who a user follows), the age of the account, the number of favorite tweets, and the number and frequency of tweets seem also affect the retweetability. The analysis also considers content feature factors. The study confirms that the inclusion of URLs and hashtags strongly correlates with retweetability.

Hong et al. [6] use retweets as a measure of popularity and apply machine learning techniques to predict how often new messages will be retweeted. The authors analyze the content of messages, temporal information, metadata of messages and users, and the user's social graph as the features in predicting the messages to be retweeted.

Kwak et al. [9] found that rankings based on the number of followers and PageRank are very similar, while rankings based on the number of retweeted messages differ, concluding that interest does not necessarily correlate with social status. In analogy with PageRank, Weng et al. [15] define the TwitterRank measure to rate users. Although these methods may be used to predict the popularity of a tweet, they cannot be used as a rank for finding interesting tweets, as they are based on user rankings and contextual information instead of content.

In summary, these more recent works indicate that the likelihood of a tweet to be retweeted is based on context of the tweet (number of followers or followees, time of tweet, age of the account) and elementary features of the content of a tweet (presence of URLs, hashtags, trending topics). Instead, we put a much stronger emphasis on the content and analyze a wider set of low-level content-based features as well as derived, high-level content-based features (topic and sentiment of tweet).

[1] as of May 2011

Table 1: Twitter datasets used in our experiments.

Dataset	Users	Tweets	Retweets
CHOUDHURY [5]	118,506	9,998,756	7.89%
CHOUDHURY-EXT [5]	277,666	29,000,000	8.64%
PETROVIĆ [11]	4,050,944	21,477,484	8.46%

Table 2: Patterns used in retweets.

RT @[username] ...
... (via @[username])
retweeting @[username] ...
♲ @[username] ...
retweet @[username] ...

2.2 Retweet Datasets

In our experiments we use three Twitter datasets that have already been employed in related work. They are listed in Table 1 with some key properties and statistics. All datasets consist of a corpus of individual tweets, along with their timestamp and an identification of the user who sent the tweet.

The datasets do not explicitly identify retweets. Therefore, in a preprocessing step we detect the retweets in the data by using the patterns given in Table 2 that capture the different ways people mark retweets. All patterns are applied in a case-insensitive way. The relative amount of retweets in all three datasets is listed in Table 1 as well.

2.3 Term Sparsity of Tweets

Due to the shortness of tweets they contain few words and hardly ever contain a word multiple times in a single tweet. Analyzing several million tweets, we observed that 85% of tweets do not contain any term more than once. This kind of sparsity has to be considered when training a prediction model as it renders the classical term frequency (*tf*) measure essentially ineffective. The histogram of *tf* values of individual tweets is shown for the CHOUDHURY-EXT dataset in Figure 1.

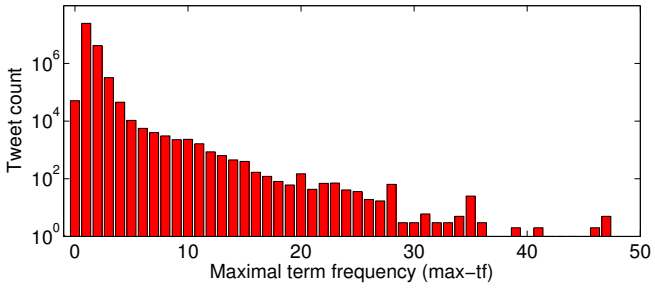

Figure 1: Distribution of maximal term frequencies (*max-tf*) in Twitter messages of the CHOUDHURY-EXT dataset after removing stop words. The y-axis is logarithmic.

Table 3: The features and their value range used to represent tweets.

Feature	Values
Direct message	$\{0, 1\}$
Includes username	$\{0, 1\}$
Includes hashtag	$\{0, 1\}$
Includes URL	$\{0, 1\}$
Exclamation mark	$\{0, 1\}$
Question mark	$\{0, 1\}$
Term positive	$\{0, 1\}$
Term negative	$\{0, 1\}$
Emoticon positive	$\{0, 1\}$
Emoticon negative	$\{0, 1\}$
Valence	$[-5, +5]$
Arousal	$[-5, +5]$
Dominance	$[-5, +5]$
Terms	$[0, 1]$
Topics (100)	$[0, 1]$

3. CONTENT-BASED RETWEET PREDICTION

As mentioned in the introduction, we are interested in retweets, because they can be seen as an indicator for interestingness. The rationale behind this hypothesis is, that the motivation of a user for retweeting a message is, that the user considers the original tweet interesting enough to relay it to her own followers. However, whether a particular tweet actually is retweeted depends heavily on context, such as the user's position in the social graph or the time of day the tweet is posted. Generally, a tweet of a user with few or passive followers is less likely to be retweeted. Similarly, tweets posted in the night tend to get retweeted less. Despite this, neither of these contextual pieces of data has any influence on the content of a tweet. To avoid introducing such a contextual bias into our analysis of interestingness, we deliberately ignore such context information and rely only on features extracted from the message itself. We proceed with a detailed description of the features we actually use for the representation of tweets.

3.1 Features

All of the following features are based on the tweets themselves and ignore a tweet's author and timestamp. A complete list of the employed content features is given in Table 3. As can be seen there, most features are binary, i.e. have a value of either 0 or 1. An exception are the features for the term odds, topics and sentiment values.

Direct messages Direct messages are addressed to another user directly. These messages start with the username of the addressee. While other users can still see these messages[2], they are not in the primary focus of the message. Direct messages are meant as kind of public conversation, rather than a general broadcast of information.

Given the rather personal note and intention of direct messages, as well as their different purpose in the interaction among users, we expect them to be much less retweeted. Accordingly, the feature of whether a tweet is a direct message is of importance for our retweet prediction.

[2]Unlike private messages which are visible only to the sender and recipient.

URLs, usernames and hashtags. Without further differentiation we consider the presence of particular items typical for tweets. These are the presence of a URL, the mention of a username or a hashtag. Usernames are used in Twitter to refer to other users directly, either for addressing a user of for talking about him. Hashtags, or simply tags, are used to mark specific topics. They can be either inline in the messages or appended after the message itself. URLs are universally used to indicate the location of the full text being talked about. On Twitter, usernames and hashtags can be identified by their specific syntax using the pattern @username and #hashtag. We use the string http: to identify URLs. These give three binary features.

Related work has already recognized the effect of the presence of URLs, hashtags and usernames on the retweet behavior. As these are purely content-based features we take them also into account in our setting.

Exclamation and question marks. We use the presence of exclamation marks "!" and question marks "?" at the end of tweets as two binary features. Exclamation marks are used in personal communication to mark strong and potentially emotional statements and in general text to mark interjections and exclamations. Question marks indicate questions in all types of text, and are by their nature intended to elicit responses. Due to the multiple uses of both symbols, we cannot easily judge if in all cases a question mark really does indicate a question or an exclamation mark expresses a strong statement. However, using the location at the end of the message as an indicator is a suitable and straightforward heuristic.

Both types of messages might have an influence on the reaction of the users that receive such a tweet. Questions can be passed on in order to extend their reach and find an expert capable of providing an answer. Statements might be retweeted to demonstrate support.

Positive and negative terms. We look for positive and negative words from the short predefined list given in Table 4. Terms expressing positive and negative feelings have previously been found to influence the social interaction in Twitter [10], and we conjecture them to also play a role in making a tweet interesting or uninteresting.

Following the line of thought on statements marked with an exclamation mark, strong positive and negative terms might foster a retweet as a sign of support among users.

Emoticons. Emoticons or smileys are short character sequences representing emotions. We parse the tweets to find positive emoticons such as :-) and negative emoticons such as :-(, giving two binary features. Table 4 gives the complete list.

As emotions have been observed to influence reaction among users, emoticons might be an indicator of interestingness. Besides transmitting emotions, they are also used to mark jokes, funny comments or irony. These kind of messages have a tendency to be passed on, as can be observed by the behavior of people forwarding emails of that kind.

Sentiments. Many tweets are personal and express sentiments. To detect the sentiments expressed by a tweet, we follow previous Twitter research and select a simple dictionary-based approach [8]. We use the Affective Norms of English Words (ANEW) dictionary [3], which gives for 1,030 English words numerical values

Table 4: Terms and emoticons expressing positive and negative emotions in Twitter messages.

	Positive	Negative
Terms	great like excellent rock on	f**k suck fail eww
Emoticons	:-) :) ;-)	:-(:(

that capture valence (pleasure vs displeasure), arousal (excitement vs calmness) and dominance (weakness vs strength).

In order to deal with inflections of dictionary words, we apply the Porter stemmer [12] to both the dictionary terms and the words extracted from the tweets. The computed values vary from 1 to 9, and we normalize them by subtracting the median value 5. This allows positively and negatively annotated terms to counterbalance each other. The total valence, arousal and dominance of a tweet are computed as the sum of the values associated with each term. Words not contained in the ANEW dictionary are considered neutral and do not affect the score for these features.

The three dimensions we used in this setting capture different notions of sentiments. This allows for a more subtle analysis than the more common sentiment analysis techniques focusing on positive and negative emotions.

Terms. The most obvious content feature in text are the contained terms. We extract terms and normalize them using case folding and the Porter stemmer [12]. Given the sparsity of tweets and the reduced expressiveness of the frequency of a term in a message we only consider presence or absence of each individual term and ignore multiple occurances. For each message M we compute the odds of it being a retweet based on the terms t_i it contains. Assuming independence between the occurrences of terms and employing Bayes' theorem the odds value can be brought into a form that is easier to handle:

$$\begin{aligned} O(\text{retweet} \mid M) &= \frac{P(\text{retweet} \mid M)}{P(\text{non-retweet} \mid M)} \\ &= \frac{P(\text{retweet}) \cdot P(M \mid \text{retweet})}{P(\text{non-retweet}) \cdot P(M \mid \text{non-retweet})} \\ &= O(\text{retweet}) \cdot \frac{P(t_1 \dots t_n \mid \text{retweet})}{P(t_1 \dots t_n \mid \text{non-retweet})} \\ &= O(\text{retweet}) \cdot \prod_{t \in M} \frac{P(t \mid \text{retweet})}{P(t \mid \text{non-retweet})} \end{aligned}$$

where $O(\text{retweet})$ are the a priori odds of a retweet, and the product ranges over the ratios of the probabilities of each contained term to occur in a retweeted or a non-retweeted message. To estimate these probabilities we use maximum likelihood estimation and Laplacian smoothing to handle unseen terms.

Even though the sparsity of tweets makes it difficult to train a prediction model on terms alone, the individual terms are a very good representation of the content. Thus, the contained terms can be seen as a very detailed and narrow description of the tweet's topic. The topic models described below provide a broader approach for capturing the content.

Topics. The topic of a tweet is a latent feature and can be inferred by analyzing a tweet's content. As each tweet is limited to 140 characters with heterogeneous vocabulary written in a language unlike standard written English, many supervised models in machine learning and natural language processing are hard to train and evaluate. Modeling Twitter content requires methods that are suitable for short texts with heterogeneous vocabulary with minimum supervision. Recent work shows that one such method which works well on short texts for modeling topics is Latent Dirichlet allocation (LDA) and its extensions [1, 15]. In LDA a topic is represented as a distribution over words that occur typically for this topic.

To learn latent topics from training and test data we construct a topic model using Gibbs sampling for latent Dirichlet allocation. We use 100 latent topics for our datasets. The number of topics for the corpus is an objective criterion that can be chosen using a number of available methods. A solution with too few topics will generally result in broad topics whereas a solution with too many topics will result in fine grained topics that are hard to interpret. Our approach is to use perplexity to choose the number of topics that leads to the best generalization performance for the task. The perplexity of a model describes its entropy and has been used to assess generalizability of text models to subsets of documents [1].

Topic features are broader in concept than individual words, since a single topic consists of an entire collection of related words. Thus, the LDA topics can be used to understand which larger topics are influential on the retweeting behavior of users.

3.2 Regression Analysis

We use logistic regression to compute the probability of a new tweet being retweeted. Logistic regression is a generalized linear regression method for learning a mapping from any number of numeric variables to a binary or probabilistic variable [7]. In the Twitter setting, we learn a mapping from the features of a tweet to the binary value indicating retweets.

Let f_{ij} be the feature i of tweet j, and retweet_j the $0/1$ variable indicating whether the tweet j was retweeted. Logistic regression learns weights w_i under the following model:

$$P(\text{retweet}_j \mid f) = \frac{1}{1 + e^{-(w_0 + \sum_i w_i f_{ij})}} \qquad (1)$$

The weights w_i learned by logistic regression can be interpreted as the log-odds for the feature i. Therefore, positive weights denote a higher probability of retweet for tweets having this feature of $P > 1/2$.

4. EVALUATION AND DISCUSSION OF RESULTS

Once we have trained the logistic regression model we obtain feature weights that indicate their influence on the probability of a message being retweeted. By looking at these weights, we can understand what influences the retweet behavior in Twitter and in conclusion can deduce assumptions on what the users consider interesting on a global scale.

By calculating the features for a new message and applying the function defined in Equation (1) we obtain a probability for this

new message to be retweeted. The computed probabilities can be used for two applications: as a measure for predicting whether a tweet will be retweeted, and as a measure for interestingness.

4.1 Accuracy of Retweet Prediction

In order to verify the learned model parameters, we measure the accuracy of retweet prediction. Therefore, we split the set of tweets into a training and a test set based on the timestamps of the tweets. The training set consists of all tweets with the lowest timestamp values and contains 75% of the available dataset. The remaining 25% of the data are retained for the test set on which we evaluate the prediction quality.

As described in Section 3 we then compute all features for the tweets in the training and test sets. For features that require a model such as word odds and topics, we compute this model only for the training set. Logistic regression is then applied to the features in the training set. The resulting weights are finally used to compute the probability of tweets in the test set to be retweeted.

Figure 2 shows the accuracy of retweet prediction in form of a ROC (receiver operating characteristic) curve. A ROC curve is a method to visualize the prediction accuracy of ranking functions showing the number of true positives in the results plotted against the number of results returned. A ROC curve generated by a random rank would result in a straight diagonal line and rankings performing better than a random rank result in a line going over that diagonal. Figure 2 shows the ROC curve for prediction by logistic regression on the PETROVIĆ dataset. The plot also contains a separate curve for each feature used separately. For features i that have a negative weight w_i learned by logistic regression as shown in Table 5, we show the ROC curve of the inverse ranking.

As expected, prediction is most accurate when taking into account all features. Individual features that perform well for retweet prediction are term odds and the detection of direct messages. We interpret this as terms playing a role in distinguishing types of tweets such as news, personal messages, etc. We conclude that only certain types of messages are likely to be retweeted.

4.2 Analysis of the weights

Now, that we have verified that our model does not make random predictions, but does capture the probability of a tweet being retweeted, we can analyze the weights we have obtained for our model. Table 5 lists the weights learned using logistic regression for different features for the CHOUDHURY-EXT dataset. The weight w_i of a binary feature i with possible values 0/1 learned by logistic regression can be interpreted as the log-odds of a tweet having that feature:

$$w_i \quad = \quad \ln \left[\frac{P(\text{retweet}_j \mid f_{ij} = 1)}{P(\text{retweet}_j \mid f_{ij} = 0)} \right]$$

From the learned regression weights for features, we can make some interesting observations:

- Direct messages are unlikely to be retweeted, which is indicated by the strong negative weight associated to the according feature. This observation corresponds to our intuition, that personal messages are not of interest on a global scale.
- Messages with hashtags, usernames and URLs are likely to be retweeted. This observation has already been made in related approaches which considered theses features alone. However, looking at the prediction performance of these features individually as shown in Figure 2, we can see that they cannot be applied in isolation but that for predicting retweets they need to be combined with other features.
- We also observe sentiments to play an important role for retweeting. Note that the weights need to be interpreted in a slightly different manner in this case. As the features can have negative and positive values (corresponding to the two poles for each sentiment feature), a negative weight does not imply a negative impact on the probability for a retweet. Rather, a negative weight is a sign that negative values for this feature increase the probability for a retweet, while positive weights indicate a better chance for a message to be retweeted if also the feature shows a positive value. Thus, tweets with negative valence values, i.e. annoying or displeasant contents, tend to get retweeted more often. Likewise tweets with positive arousal and dominance values, i.e. exciting and intense tweets, are more likely to be retweeted. This seems to confirm the idiom that bad news travels fast.
- Also, including a positive emoticon such as `:-)` lowers the probability of retweet, whereas adding a negative one such as `:-(` increases the probability. By relating the negative emoticons to negative and displeasant emotions, this seems to support the observations made above for sentiments.
- Positive and negative terms from our short list in Table 4 both render a tweet more likely to be retweeted. In this case, positive words have a stronger effect. One possible explanation is, that users are slightly more reluctant to retweet messages containing rude terms. In any case, these extreme and strong words seem to stimulate a reaction in the followers.
- Tweets ending in an exclamation mark are not likely to be retweeted, but tweets ending in a question mark are. This is an interesting observation and would motivate a deeper analysis of the social aspects on Twitter in question answering, i.e. if questions are really passed on to find an expert capable of answering them.

Table 5: Weights of features learned by logistic regression on the CHOUDHURY-EXT dataset. Positive values denote a positive contribution to a tweet being retweeted; negative weights denote a negative contribution to a tweet being retweeted.

Weight w_i	**Feature** i
-5.45	Constant
-147.89	Direct message
146.82	Includes username
42.27	Includes hashtag
249.09	Includes URL
-16.85	Exclamation mark
23.67	Question mark
13.66	Term positive
8.72	Term negative
-21.80	Emoticon positive
9.94	Emoticon negative
-26.88	Valence
33.97	Arousal
19.56	Dominance
19.79	Term odds

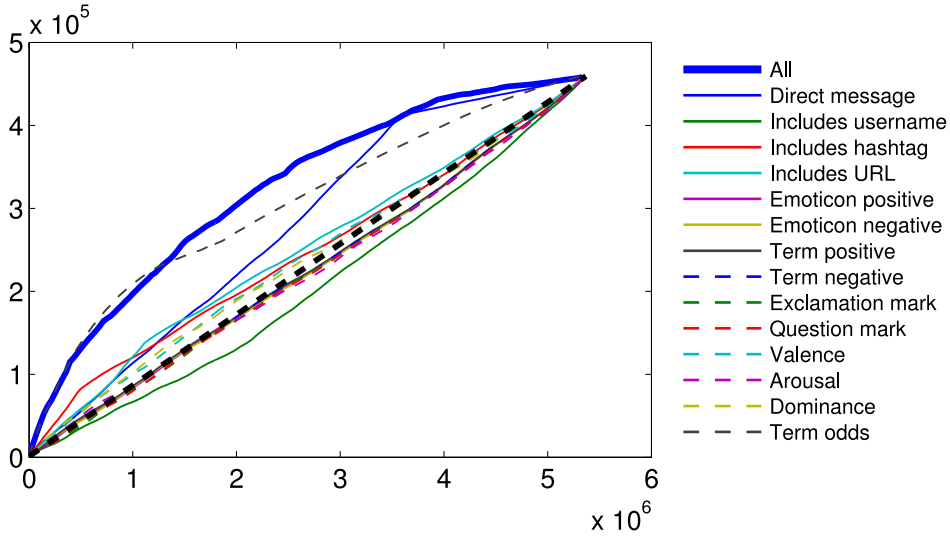

Figure 2: The accuracy of retweet prediction using logistic regression based on all features, and of each feature separately, in the PETROVIĆ dataset. The accuracy is represented as a ROC curve. For clarity, ROC curves for topics are not shown.

- Terms are a strong indicator for a retweet. As already seen in the evaluation of the prediction quality, the content has a strong influence on the probability of a message to be retweeted.

The topic features are not included in the previous list because they need to be discussed in a more differentiated way. As there are 100 different topics, we cannot address all of them individually. Rather we report the trends we have observed with respect to the topic features.

Table 6 shows the four topics having highest log-odds with positive weights (top high-probability terms for each topic) based on the logistic regression score of the training data that are most likely to be retweeted, and four topics having lowest log-odds with negative weights that are least likely to be retweeted based on regression analysis of training data. From the analysis results it is clear that topics that are very likely to be retweeted address broader public interests such as social media and social networking in general, economy and Christmas-like holidays and public events. Topics that are least likely to be retweeted based on regression scores are more specific and individual in nature, reflecting personal tasks, moods and observations.

4.3 Example: Interesting Tweets

Given the notion of interestingness we can obtain from the odds for a tweet to be retweeted allows for realizing practical applications. For instance, it is possible to get the most interesting tweets from a dataset about a specific topic. As an example, we have listed the top ten most interesting tweets with respect to the log-odds of predicted retweet probability for the term Recipe in Table 7.

5. CONCLUSION

In this paper we introduced and evaluated a method to determine the interestingness of microblog messages. We based our method on the retweet function of Twitter as a measure for messages with a wider interest. To overcome the context bias of, e.g. a user's social network or time, we used a learning approach based on pure content features to predict the probability of a message to be retweeted. To capture the content we used low-level features such as the presence of URLs, hashtags, usernames, question and exclamation marks, emoticons, positive and negative words, as well as high-level feature such as sentiments and latent topics.

Table 6: Logistic regression weights and corresponding high probability terms that describe the particular topic in the CHOUDHURY dataset. The weights can be interpreted as the log-odds of a tweet from a given topic to be retweeted. Positive weights indicate topics that are likely to be retweeted and negative weights indicate topics that are unlikely to be retweeted.

Weight w_i	Topic i
27.54	social media market post site web tool traffic network
16.08	follow thank twitter welcome hello check nice cool people
15.25	credit money market business rate economy home
2.87	christmas shop tree xmas present today wrap finish
−14.43	home work hour long wait airport week flight head
−14.43	twitter update facebook account page set squidoo check
−26.56	cold snow warm today degree weather winter morning
−75.19	night sleep work morning time bed feel tired home

We made the following observations about the retweeting behavior of Twitter users: As a general rule, a tweet is likely to be retweeted when it is about a general, public topic instead of a narrow, personal topic. For instance, a tweet is unlikely to be retweeted when it is addressed to another Twitter user directly, while our topic analysis revealed that general topics affecting many users like social media or Christmas are more likely to be retweeted. This can be understood as the Twitter platform being better suited as a news and announcement channel rather than a personal communication platform, complementing the description of Twitter as news media in [9]. A further interesting observation in this context is the tendency that bad news seem to travel fast in Twitter.

In ongoing work we are extending the approach to take into consideration further features, like the text contained in web pages linked through URLs in tweets. Given our observations, especially the sentiments in tweets are a topic worth investigating further. Also

Table 7: Top 10 interesting tweets by the log-odds of predicted retweet probability for the query Recipe in the CHOUDHURY dataset.

Log-odds	Tweet
3245.0091	How to make potato latkes video recipe by @handmadekitchen http://tinyurl.com/n22t4p #cooking #recipe
2455.3082	Recipe for Chinese Chicken Congee inspired by a painting from the Sung Dynasty http://bit.ly/16V5L0 #art #food #foodie #recipe
2439.568	Have a great idea for a recipe using @greensbury organic meats? You could win free #meat and get your recipe posted!
2385.6006	New Raw Food World S Raw Ice Cream Recipe, Episode #134: We've got a Raw Ice Cream Recipe JU.. http://tinyurl.com/pdt7cq
2362.9419	Recipe looks good - Potatoes Gribiche Recipe: I've not really been in the mood for winte.. http://tinyurl.com/cay294
2337.9173	what to pack for a day at the beach with the fam (plus a yummy beach pasta salad recipe) http://is.gd/1sBKM #ocmom #recipe
2301.8325	Tasty pasta cake recipe's :-) Bub Hub Pregnancy & Parenting Forum: Tasty pasta cake recipes Recipe.. http://bit.ly/ONk9l
2294.2587	It's Taco Tuesday! How about making some Buffalo Sausage Tacos at home, great recipe: http://bit.ly/jwPDT yummm! #food #recipe
2285.9819	Great grilling recipe for this weekend: Cranberry-Onion Pork Roast, Check out the recipe in the Hotlanta Forum: http://tr.im/s8HA #food
2200.9418	RT @nytimesdining: NYT Recipe Challenge #nytrc: Tweet this recipe in as few characters as possible. Serial tweets ok. http://bit.ly/bhf92

the suitability of using Twitter as a social question answering system seems a promising direction for further analysis.

Our analysis of content quality could also be interesting in the field of measuring influence among Twitter users or more accurate retweet prediction by taking again context into consideration. Another interesting topic is that of spam. As spammers also use the retweet function to feign relevance of their messages, our methods may be susceptible to spam. So far we employed only basic methods to filter out spam and more sophisticated methods might improve performance.

Finally, we plan to use interestingness as a static quality measure in information retrieval on microblogs. There, interestingness might overcome sparsity and quality issues inherent to microblogs that pose a challenge when searching for tweets.

6. ACKNOWLEDGMENTS

This work was supported by the project WeGov (www.wegov-project.eu) funded by the European Commission under EC Grant number 248512 in the 7th Framework Programme, ICT2009.7.3.

7. REFERENCES

[1] D. M. Blei, A. Y. Ng, and M. I. Jordan. Latent Dirichlet allocation. *J. of Machine Learning Research*, 3:993–1022, 2003.

[2] D. Boyd, S. Golder, and G. Lotan. Tweet, tweet, retweet: Conversational aspects of retweeting on Twitter. In *Hawaii Int. Conf. on System Sciences*, pages 1–10, 2010.

[3] M. M. Bradley and P. J. Lang. Affective norms for English words (ANEW): Instruction manual and affective ratings. Technical report, The Center for Research in Psychophysiology, University of Florida, 1999.

[4] M. Cha, H. Haddadi, F. Benevenuto, and K. P. Gummadi. Measuring user influence in Twitter: the million follower fallacy. In *Proc. Int. Conf. on Weblogs and Social Media*, pages 10–17, 2010.

[5] M. D. Choudhury, Y.-R. Lin, H. Sundaram, K. S. Candan, L. Xie, and A. Kelliher. How does the data sampling strategy impact the discovery of information diffusion in social media? In *Proc. Conf. on Weblogs and Social Media*, pages 34–41, 2010.

[6] L. Hong, O. Dan, and B. D. Davison. Predicting popular messages in twitter. In *WWW (Companion Volume)*, pages 57–58, 2011.

[7] D. W. Hosmer and S. Lemeshow. *Applied logistic regression*. John Wiley and Sons, 2000.

[8] E. Kim, S. Gilbert, M. J. Edwards, and E. Graeff. Detecting sadness in 140 characters: Sentiment analysis and mourning Michael Jackson on Twitter. Technical report, Web Ecology Project, Aug 2009.

[9] H. Kwak, C. Lee, H. Park, and S. Moon. What is Twitter, a social network or a news media? In *Proc. Int. World Wide Web Conf.*, pages 591–600, 2010.

[10] A. Pepe, H. Mao, and J. Bollen. Modeling public mood and emotion: Twitter sentiment and socio-economic phenomena. *CoRR*, abs/0911.1583, 2009.

[11] S. Petrović, M. Osborne, and V. Lavrenko. The Edinburgh Twitter corpus. In *Proc. Workshop on Computational Linguistics in a World of Social Media*, pages 25–26, 2010.

[12] M. F. Porter. An algorithm for suffix stripping. *Readings in Information Retrieval*, 14(3):130–137, 1980.

[13] D. M. Romero, W. Galuba, S. Asur, and B. A. Huberman. Influence and passivity in social media. *CoRR*, abs/1008.1253, 2010.

[14] 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 *Proc. Int. Conf. on Social Computing*, pages 177–184, 2010.

[15] J. Weng, E.-P. Lim, J. Jiang, and Q. He. TwitterRank: Finding topic-sensitive influential twitterers. In *Proc. Int. Conf. on Web Search and Data Mining*, pages 261–270, 2010.