Temporal Twitter prediction by content and network

Bálint Daróczy[1] Róbert Pálovics[1,2] Vilmos Wieszner[3]
Richárd Farkas[3] András A. Benczúr[1]
[1]Institute for Computer Science and Control, Hungarian Academy of Sciences (MTA SZTAKI)
[2]Technical University Budapest
[3]University of Szeged, Institute of Informatics
{daroczyb, rpalovics, benczur}@ilab.sztaki.hu, {wieszner, rfarkas}@inf.u-szeged.hu

ABSTRACT

In recent years Twitter became *the* social network for information sharing and spreading. By retweeting, users spreading information and build cascades of information pathways. In this paper we investigate the possibility of predicting the future popularity of emerging retweet cascades immediately after the message appears. We introduce a supervised machine learning approach which employs a rich feature set utilizing the textual content of the messages along with the retweet networks of the users. We also propose a temporal evaluation framework focusing on user level predictions in time.

Keywords

Twitter, Retweet prediction, Temporal classification, Language features

1. INTRODUCTION

Twitter, a mixture of a social network and a news media [16], has recently became the largest medium where users may spread information along their social contacts.

In this paper we investigate the temporal influence of messages sent over Twitter. Cha et al. [7] define influence as "...the power of capacity of causing an effect in indirect intangible ways...". In their key observation, the influence of a user is best characterized by the size of the audience who retweets rather than the size of the follower network.

Our goal is to predict the timely success of the information spread, on the individual message level. We analyze how certain messages may reach out to a large number of Twitter users. In contrast to a similar investigation for analyzing the influence of users [3], we investigate each tweet by taking both the author user and the textual content of the message into account.

We characterize the users both by the statistical properties of their follower network and their past retweet counts. The textual content is described by the terms of the normalized text and by several orthographic features along with deeper (psycho)linguistic ones that try to capture the modality of the message in question.

In our experiments we use the data set of [1] that consists of the messages and the corresponding user network of the Occupy movement.

The main contributions of this work is that we carried out an intensive feature engineering both at network and content analysis – instead of focusing on only one of them – and the added value of the two worlds was empirically evaluated. In our results we consider user and network features as defined in [8] and our previous work [19] as baseline and concentrate on the power of content analysis.

1.1 Related results

Social influence in Web based networks is investigated in several results: Bakshy et al. [4] model social contagion in the Second Life virtual world. Ghosh and Lerman [11] compares network measures for predicting the number of votes for Digg posts, who even give an empirical comparison of information contagion on Digg vs. Twitter [17]. In [12, 13], long discussion based cascades built from comments are investigated in four social networks, Slashdot (technology news), Barrapunto (Spanish Slashdot), Meneame (Spanish Digg) and Wikipedia. They propose models for cascade growth and estimate model parameters but give no size predictions.

A number of related studies have largely descriptive focus, unlike our quantitative prediction goals. In [7] high correlation is observed between indegree, retweet and mention influence, while outdegree (the number of tweets sent by the user) is found to be heavily spammed. [16] reports similar findings on the relation among follower, mention and retweet influence. Several more results describe the specific means of information spread on Facebook [5, 2, 6].

Similar to our results, Cheng et al. [8] predict retweet count based on network features. Unlike in our result where we predict immediately after the tweet is published, they consider prediction after the first few retweets. The network features used in their work are similar to the ones in the present paper and in our earlier work [19]. We consider these results as baseline in this paper.

From the content analysis point of view, there has been several studies focusing exclusively on the analysis of the tweet messages' textual content to solve the re-tweet count prediction problem. Besides the terms of the message, Naveed et al. [18] introduced the features of direct message, mention, hashtag, URL, exclamation mark, question mark, positive and negative sentiment, positive and negative emoticons and valence, arousal, dominance lexicon features. Wang et al. [22] proposed deeper linguistic features like verb tense, named entities, discourse relations and sentence similarity.

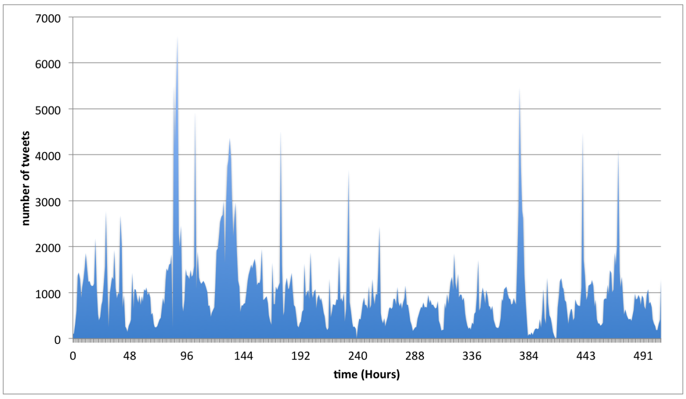

Figure 1: Temporal density of tweeting activity.

Table 1: Size of the tweet time series.

Number of users	371,401
Number of tweets	1,947,234
Number of retweets	1,272,443

Table 2: Size of the follower network.

Number of users	330,677
Number of edges	16,585,837
Average in/out degree	37

Gupta et al. [14] addressed the task of scoring tweets according to their credibility. Credibility is a highly related phenomena to social influence. Moreover, this work is related to our ones as it also combines author, network and content features. The feature set to describe the content of a message included the following novel items: the length of the message, swear words, pronouns and self words.

2. DATA SET

The dataset was collected by Aragón et al. [1] using the Twitter API that we extended by a crawl of the user network. Our data set hence consists of two parts:

- *Tweet dataset:* tweet text and user metadata on the Occupy Wall Street movement[1].
- *Follower network:* The list of followers of users who posted at least one message in the tweet dataset.

Table 1 shows the number of users and tweets in the dataset. One can see that a large part of the collected tweets are retweets. Table 2 contains the size of the crawled social networks. Note that the average in- and outdegree is relatively high. Fig. 1 shows the temporal density of tweeting activity.

For each tweet, our data contains

- tweet and user ID,
- timestamp of creation,
- hashtags used in the tweet, and

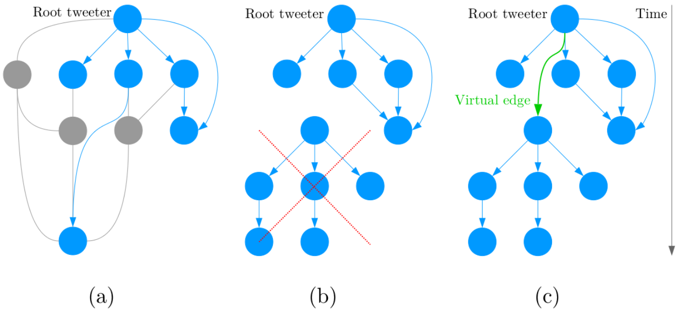

Figure 2: Creation of retweet cascades: Figure (a) shows the computation of the cascade edges. In Figures (b) and (c) we show the possible solutions in case of missing cascade edges.

- the tweet text content.

In case of a retweet, we have all these information not only on the actual tweet, but also on the original *root tweet* that had been retweeted. We define the root tweet as the first occurrence of a given tweet.

3. RETWEET CASCADES

3.1 Constructing retweet cascades

In case of a retweet, the Twitter API provides us with the ID of the original tweet. By collecting retweets for a given original tweet ID, we may obtain the set users who have retweeted a given tweet with the corresponding retweet timestamps. The Twitter API however does not tell us the actual path of cascades if the original tweet was retweeted several times. The information from the Twitter API on the tweet needs to be combined with the follower network to reconstruct the possible information pathways for a given tweet. However it can happen that for a given retweeter, more than one friend has retweeted the corresponding tweet before and hence we do not know the exact information source of the retweeter. The retweet ambiguity problem is well described in [3]. In what follows we consider all friends as possible information sources. In other words for a given tweet we consider all directed edges in the follower network in which information flow could occur (see Fig. 2 (a)).

3.2 Restoring missing cascade edges

For a given tweet, the computed edges define us a *retweet cascade.* However our dataset contains only a sample of tweets on the given hashtags and hence may not be complete: it can happen that a few intermediate retweeters are missing from our data. As a result, sometimes the reconstructed cascade graphs are disconnected. As detailed in Fig. 2 (b) and (c), we handle this problem in two different ways. One possible solution is to only consider the first connected component of the cascade (see Fig. 2 (b)). Another one is to connect each disconnected part to the root tweeter with one virtual cascade edge (see Fig. 2 (c)). In what follows, we work with cascades that contain virtual edges, therefore every retweeter is included in the cascade.

4. FEATURE ENGINEERING

[1] http://en.wikipedia.org/wiki/Occupy_Wall_Street

To train our models, we generate features for each root tweet in the data and then we predict the future cascade size of the root tweet from these feature sets. For a given root tweet, we compute features about

- the author user (*user features*),
- the the follower network of the author (*network features*) and
- the textual content of the tweet itself (*content features*).

Table 3 gives an overview of the feature templates used in our experiments.

4.1 Network Features

We consider statistics about the user and her cascades in the past as well as the influence and impressibility of her followers. We capture the influence and impressibility of a user from previously observed cascades by measuring the following quantities:

- *Number of tweets in different time frames:* for a given root tweet appeared in time t and a predefined time frame τ, we count the number of tweets generated by the corresponding user in the time interval $[t - \tau, t]$. We set τ for 1, 6, 12, 24, 48 and 168 hours.
- *Average number of tweets in different time frames:* We divide the number of tweets in a given time frame by τ.
- *User influence:* for a given user, we compute the number of times one of her followers retweeted her, divided by the number of the followers of the user.
- *User impressibility:* for a given user, we compute the number of times she retweeted one of her followees, divided by the number of followees of the user.

4.2 Content features

The first step of content processing is text normalization. We converted the text them into lower case form except those which are fully upper cased and replaced tokens by their stem given by the Porter stemming algorithm. We replaced user mentions (starting with '@') and numbers by placeholder strings and removed the punctuation marks.

The *content features* are extracted from the normalized texts. The basic feature template in text analysis consists the *terms* of the message. We used a simple whitespace tokenizer rather than a more sophisticated linguistic tokenizer as previous studies reported its empirical advantage [15]. We employed unigrams and bigrams of tokens because longer phrases just hurt the performance of the system in our preliminary experiments.

Besides terms, we extracted the following features describing the *orthography* of the message:

- *Hashtags* are used to mark specific topics, they can be appended after the tweets or inline in the content, marked by #. From the counts of hashtags the user can tips the topic categories of tweet content but too many hashtag can be irritating to the readers as they just make confusion.
- *Telephone number:* If the tweet contains telephone number it is more likely to be spam or ads.
- *Urls:* The referred urls can navigate the reader to text, sound, and image information, like media elements and journals thus they can attract interested readers. We distinguish between full and truncated urls. The truncated urls are ended with three dot, its probably copied from other tweet content, so it was interested by somebody.
- The *like sign* is an illustrator, encouragement to others to share the tweet.
- The presence of *question mark* indicate uncertainty. In Twitter they are usually a rhetorical question rather than a concrete question (people do not search answer on Twitter). The author more likely want to made the reader to think on what contains the message.
- The *Exclamation mark* highlight the part of the tweet, it express emotions and opinions.
- If *Numerical expressions* are present the facts are quantified then it is more likely to have real information content. The actual value of numbers were ignored.
- *Mentions:* If a user mentioned (referred) in the tweet the content of the tweet is probably connected to the mentioned user. It can have informal or private content.
- *Emoticons* are short character sequences representing emotions. We clustered the emoticons into positive, negative and other categories.

The last group of content features tries to capture the *modality* of the message:

- *Swear words* occurring influence the style and attractiveness of the tweet. The reaction for swearing can be ignorance and also reattacking, which is not relevant in terms of retweet cascade size prediction. We extracted the swear word list from `http://www.youswear.com`.
- *Weasel words and phrases*[2] aimed at creating an impression that a specific and/or meaningful statement has been made when in fact only a vague or ambiguous claim has been communicated. We used the weasel word lexicon of [21].
- We employed the linguistic inquiry categories (LIWC) [20] of the tweets' words as well. These categories describe words from emotional, cognitive and structural points of view. For example the "ask" word it is in Hear, Senses, Social and Present categories. Different LIWC categories can have different effect on the influence of the tweet in question.

[2] See `http://en.wikipedia.org/wiki/Wikipedia:Embrace_weasel_words`.

Table 3: Feature set.

user	*number of* {followers, tweets, root tweets}, *average* {cascade size, root cascade size}, *maximum* {cascade size, root cascade size}, *variance* of {cascade sizes, root cascade sizes}, *number of* tweets generated with different time frames, *time average* of the number of tweets in different time frames
network	tweeter's influence and impressibility followers' average influence and impressibility
terms	normalized *unigrams and bigrams*
orthographic	number of # with the values 0, 1, 2 . . . 4 or 4 < number of {like *signs*, ?, !, mentions} number of full and truncated *urls* number of arabic *numbers* and *phone numbers* number of positive/negative/other *emoticons*
modality	number of swear words and weasel phrases union of the *inquiry categories* of the words

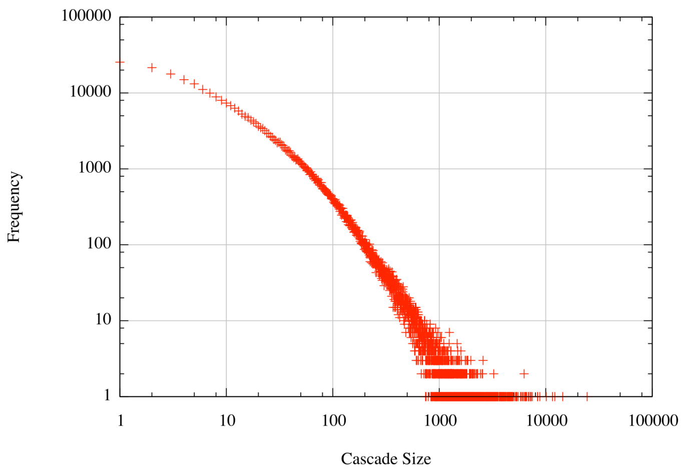

Figure 3: Cascade size distribution.

5. TEMPORAL TRAINING AND EVALUATION

Here we describe the way we generate training and test sets for our algorithms detailed in Section 6. First, for each root tweet we compute the corresponding network and content features. We create daily re-trained models: for a given day t, we train a model on all root tweets that have been generated before t but appeared later than $t - \tau$, where τ is the preset time frame. After training based on the data before a given day, we compute our predictions for all root tweets appeared in that day.

Our goal is to predict cascade size at the time when the root tweet is generated. As the cascade size follows a power law distribution (see Fig. 3), we estimate sizes on the logarithmic scale. In our experiments multi-class classification for ranges of cascade sizes performed better than regression methods for directly predicting the logarithm of the size. We defined three buckets, one with 0. . . 5 (referred as "low"), one with 6. . . 50 ("medium") and a largest one with more than 50 ("high") retweeters participating in the cascade. We trained multiclass random forest classifiers for the three buckets.

We evaluate performance by AUC [10] averaged for the three classes. Note that AUC has a probabilistic interpretation: for the example of the "high" class, the value of the AUC is equal to the probability that a random highly retweeted message is ranked before a random non-highly retweeted one.

By the probabilistic interpretation of AUC, we may realize that a classifier will perform well if it orders the users well with little consideration on their individual messages. Since our goal is to predict the messages in time and not the rather static user visibility and influence, we define new averaging schemes for predicting the success of individual messages.

We consider the classification of the messages of a single user and define two aggregations of the individual AUC values. First, we simply average the AUC values of users for each day (user average)

$$AUC\mathrm{user} = \frac{1}{N}\sum_{i=1}^{N} AUC_i, \qquad (1)$$

Second, we are weighting the individual AUC values with the activity of the user (number of tweets by the user for the actual day)

$$AUC\mathrm{wuser} = \frac{\sum_{i=1}^{N} AUC_i T_i}{\sum_i^N T_i} \qquad (2)$$

where T_i is the number of tweets by the i-th user.

6. RESULTS

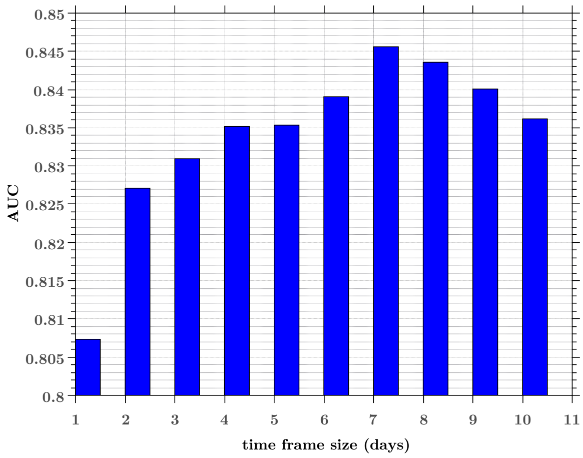

Figure 4: Daily average AUC of classifiers trained with different set of features.

For each day in the testing period, we train a random forest [9] classifier to predict the future retweet size of tweets appearing on that day.

First, we measure classifier performance by computing the average AUC values of the final results for the three size ranges.

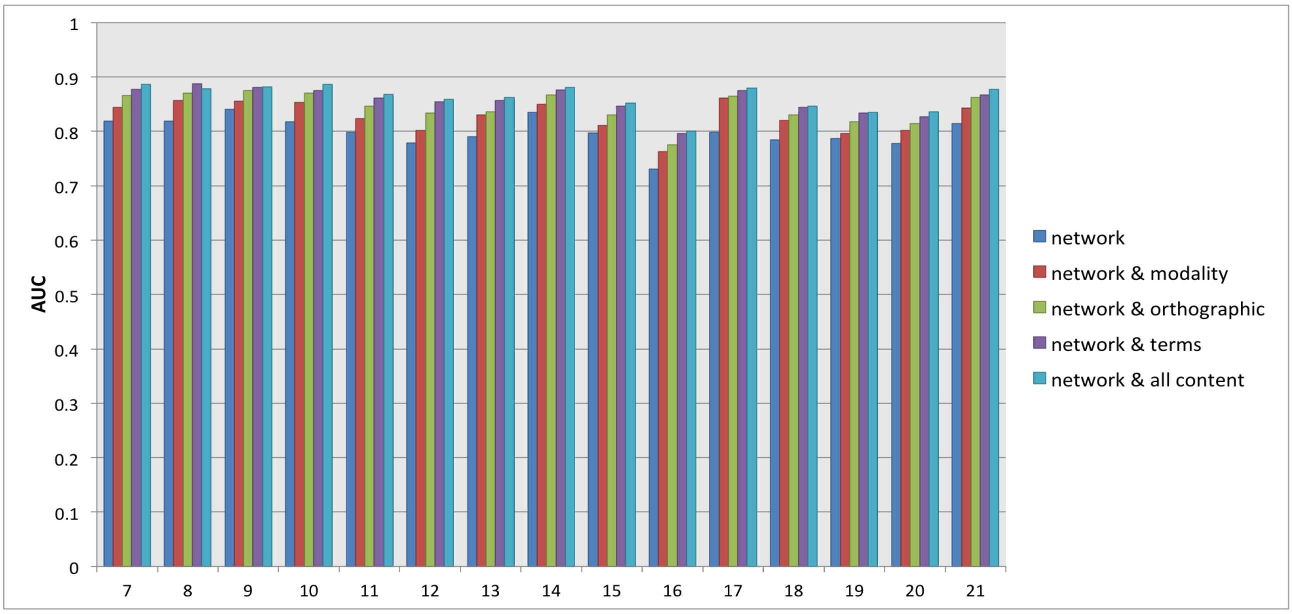

Figure 5: Daily average AUC of classifiers trained with different set of features.

Table 4: Retweet size classification daily average performance of different feature sets

Retweet range		Low	Medium	High	Weighted
Method					Average
network	AUC	0.799	0.785	0.886	0.799
network & modality	AUC	0.827	0.814	0.905	0.827
network & orthographic	AUC	0.844	0.829	0.912	0.843
network & terms	AUC	0.857	0.847	0.914	0.857
network & all content	AUC	0.862	0.849	0.921	0.862

Table 5: Retweet size classification daily average performance of different feature sets evaluated on the user level as defined in equations (1) **and** (2)**.**

Retweet range		Low		Medium		High		Average	
Method		Uniform	Weighted	Uniform	Weighted	Uniform	Weighted	Uniform	Weighted
network	AUC	0.684	0.712	0.752	0.800	0.746	0.796	0.719	0.756
network & modality	AUC	0.700	0.722	0.751	0.796	0.737	0.756	0.726	0.757
network & orthographic	AUC	0.702	0.731	0.753	0.797	0.768	0.782	0.730	0.764
network & terms	AUC	0.705	0.732	0.757	0.800	0.767	0.786	0.733	0.766
network & all content	AUC	0.740	0.783	0.763	0.812	0.769	0.820	0.752	0.797

As mentioned in Section 5, we may train our model with different time frames. In Figure 4 we show the average AUC value with different time frames. As Twitter trends change rapidly, we achieve the best average results if we train our algorithms on root tweets that were generated in the previous week (approximately seven days).

We were interested in how different feature sets affect classifier performance. For this reason we repeated our experiments with different feature subsets. Figure 5 shows our results. For each day, the network features give a strong baseline. The combination of these features with the content result in strong improvement in classifier performance. In Table 4 we summarize the average AUC values for different feature subsets over all four datasets. Our results are consistent: in each case the content related features improve the performance.

Our main evaluation is found in Table 5 where we consider the user level average AUC values as described in Section 5. As expected, since the new evaluation metrics give more emphasis on distinguishing between the tweets of the same user, we see even stronger gain of the modality and orthographic features.

7. CONCLUSIONS AND FUTURE WORK

In this paper we investigated the possibility of predicting the future popularity of a recently appeared text message in Twitter's social networking system. Besides the typical user and network related features, we consider hashtag and linguistic analysis based ones as well. Our results do not only confirm the possibility of predicting the future popularity of a tweet, but also indicate that deep content analysis is important to improve the quality of the prediction.

In our experiments, we give high importance to the temporal aspects of the prediction: we predict immediately after the message is published, and we also evaluate on the user level. We consider user level evaluation key in temporal analysis, since the influence and popularity of a given user is relative stable while the retweet count of her particular messages may greatly vary in time.

Acknowledgments

We thank Andreas Kaltenbrunner for providing us with the Twitter data set [1].

The publication was supported by the KTIA_AIK_12-1-2013-0037 and the PIAC_13-1-2013-0197 projects. The projects are supported by Hungarian Government, managed by the National Development Agency, and financed by the Research and Technology Innovation Fund. Work conducted at the Institute for Computer Science and Control, Hungarian Academy of Sciences (MTA SZTAKI) was supported in part by the EC FET Open project "New tools and algorithms for directed network analysis" (NADINE No 288956), by the Momentum Grant of the Hungarian Academy of Sciences, and by OTKA NK 105645. Work conducted at the Technical University Budapest has been developed in the framework of the project "Talent care and cultivation in the scientific workshops of BME" project. This project is supported by the grant TÁMOP - 4.2.2.B-10/1–2010-0009.

8. REFERENCES

[1] P. Aragón, K. E. Kappler, A. Kaltenbrunner, D. Laniado, and Y. Volkovich. Communication dynamics in twitter during political campaigns: The case of the 2011 spanish national election. *Policy & Internet*, 5(2):183–206, 2013.

[2] E. Bakshy, D. Eckles, R. Yan, and I. Rosenn. Social influence in social advertising: evidence from field experiments. In *Proceedings of the 13th ACM Conference on Electronic Commerce*, pages 146–161. ACM, 2012.

[3] E. Bakshy, J. M. H., W. A. Mason, and D. J. Watts. Everyone's an influencer: quantifying influence on twitter. In *Proceedings of the fourth ACM international conference on Web search and data mining*, pages 65–74. ACM, 2011.

[4] E. Bakshy, B. Karrer, and L. A. Adamic. Social influence and the diffusion of user-created content. In *Proceedings of the 10th ACM conference on Electronic commerce*, pages 325–334. ACM, 2009.

[5] E. Bakshy, I. Rosenn, C. Marlow, and L. Adamic. The role of social networks in information diffusion. In *Proceedings of the 21st international conference on World Wide Web*, pages 519–528. ACM, 2012.

[6] M. S. Bernstein, E. Bakshy, M. Burke, and B. Karrer. Quantifying the invisible audience in social networks. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, pages 21–30. ACM, 2013.

[7] M. Cha, H. Haddadi, F. Benevenuto, and K. Gummadi. Measuring user influence in twitter: The million follower fallacy. In *4th International AAAI Conference on Weblogs and Social Media (ICWSM)*, 2010.

[8] J. Cheng, L. Adamic, P. A. Dow, J. M. Kleinberg, and J. Leskovec. Can cascades be predicted? In *Proceedings of the 23rd international conference on World wide web*, pages 925–936. International World Wide Web Conferences Steering Committee, 2014.

[9] FastRandomForest. Re-implementation of the random forest classifier for the weka environment. `http://code.google.com/p/fast-random-forest/`.

[10] J. Fogarty, R. S. Baker, and S. E. Hudson. Case studies in the use of roc curve analysis for sensor-based estimates in human computer interaction. In *Proceedings of Graphics Interface 2005*, GI '05, pages 129–136, School of Computer Science, University of Waterloo, Waterloo, Ontario, Canada, 2005. Canadian Human-Computer Communications Society.

[11] R. Ghosh and K. Lerman. Predicting influential users in online social networks. *arXiv preprint arXiv:1005.4882*, 2010.

[12] V. Gómez, H. J. Kappen, and A. Kaltenbrunner. Modeling the structure and evolution of discussion cascades. In *Proceedings of the 22nd ACM conference on Hypertext and hypermedia*, pages 181–190. ACM, 2011.

[13] V. Gómez, H. J. Kappen, N. Litvak, and A. Kaltenbrunner. A likelihood-based framework for the analysis of discussion threads. *World Wide Web*, pages 1–31, 2012.

[14] A. Gupta, P. Kumaraguru, C. Castillo, and P. Meier. Tweetcred: Real-time credibility assessment of content on twitter. In *Social Informatics*, volume 8851 of *Lecture Notes in Computer Science*, pages 228–243.

2014.

[15] V. Hangya and R. Farkas. Filtering and polarity detection for reputation management on tweets. In *Working Notes of CLEF 2013 Evaluation Labs and Workshop*, 2013.

[16] 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*, pages 591–600. ACM, 2010.

[17] K. Lerman and R. Ghosh. Information contagion: An empirical study of the spread of news on digg and twitter social networks. In *Proceedings of 4th International Conference on Weblogs and Social Media (ICWSM)*, 2010.

[18] 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 Web Science Conference*, WebSci '11. ACM, 2011.

[19] R. Palovics, B. Daroczy, and A. Benczur. Temporal prediction of retweet count. In *Cognitive Infocommunications (CogInfoCom), 2013 IEEE 4th International Conference on*, pages 267–270. IEEE, 2013.

[20] J. Pennebaker, C. Chung, M. Ireland, A. Gonzales, and R. Booth. The development and psychometric properties of liwc2007. Technical report, University of Texas at Austin, 2007.

[21] Gy. Szarvas, V. Vincze, R. Farkas, Gy. Móra, and I. Gurevych. Cross-genre and cross-domain detection of semantic uncertainty. *Computational Linguistics*, 38(2):335–367, 2012.

[22] A. Wang, T. Chen, and M.-Y. Kan. Re-tweeting from a linguistic perspective. In *Proceedings of the Second Workshop on Language in Social Media*, pages 46–55, 2012.