A Comprehensive Analysis of Tweet Content and Its Impact on Popularity

Mohammad Mahdavi
School of Electrical and
Computer Engineering
University of Tehran
Email: moh.mahdavi@ut.ac.ir

Masoud Asadpour
School of Electrical and
Computer Engineering
University of Tehran
Email: asadpour@ut.ac.ir

Seyed Morteza Ghavami
School of Electrical and
Computer Engineering
University of Tehran
Email: m.ghavami@ut.ac.ir

***Abstract*—By the appearance of the online social networks, ordinary people have gained more chance to make and publish content. However, for audiences, as the number of these shared contents grows, the importance of detecting important and related ones increases. So, a significant question is how much would a shared content become popular among audiences, regardless of the source of content?**

In this paper, we investigate this question by performing a comprehensive analysis of tweet content and studying its impact on popularity of tweet. Here, the number of retweets of tweet is used as a popularity measure. We show that tweets with ”social” content, have in general more chance of popularity due to their attraction for society. In contrast, tweets with ”individual” content, have little chance to get popular. We collect a fair data set of tweets. In order to do more detailed investigation and access the semantic features, we set an annotation and labeling process. We analyze the informativeness of content-based features and use them to train predictive models. The results clearly show the importance of content-based features. They specifically support this idea that specifying whether a tweet is speaking about an individual or social subject, is the most informative content-based feature to predict the popularity i.e. the number of retweets.

Keywords—Online Social Network, Twitter, Tweet Content, Popularity Detection, Retweet Prediction.

I. Introduction

Since the dawn of Web 2.0 and its prominent consequence, the online social networks, ordinary people have found more chance to generate and share content. But like the real world, in this virtual environment, a minority gain most of others' attention. Actually, because of the ”preferential attachment” nature of the social networks, celebrities have more opportunities to be heard [1].

In this situation, what chance do ordinary people have to express their opinions in the social networks? They lose the battle of popularity to celebrities but they still have something to prove themselves: the content! Every user, regardless of his/her fame and reputation can generate content in the social networks. If the published contents are rich enough (e.g. interesting, well-written, mentioning an important subject), they would perhaps become popular.

However, what is a good content for popularity? Goodness is an ambiguous concept but in the context of social networks, popularity of content can be quantified by quantitative measures e.g. the number of times a content is shared among users. When a content is shared by a user, that simply means the user has found it important for some reasons [2]. Therefore, this count can be a good measure of content popularity among the society members.

978-1-5090-3435-2/16/$31.00 2016 IEEE

In this paper we address this question by doing a comprehensive analysis of tweet content. We believe that in Twitter content is very important and regardless of the person who has generated it, content can still help predict popularity. Our main contribution in this paper is that we show tweets containing society's concerns (which we call them *social* contents), have more attention of other users. On the other hand, tweets with individual's concerns (which we call them *individual* contents), have less chance to get popular among other users.

We have collected a fair sample of tweets from Twitter as an evidence. With the help of volunteer annotators, we have labeled these tweets. The results of experiments on this data set clearly show the importance of content-based features for popularity prediction. The results support the idea that the social contents are more popular and in contrast, the individual contents are not attractive for the society. We have used the content-based features to train predictive models for predicting popularity classes of tweets. The results of these classification tasks show feasibility of the approach and the importance of content by itself.

The rest of this paper is organized as follows: in section II we look at previous works in this area. In section III we describe the procedure of data collection and annotation. The extracted features are introduced in section IV. In section V we look at results of the experiments and discuss about them. And finally, in section VI we conclude the paper.

II. Related Work

Since we aim at predicting popularity and our measure is the number of retweets, previous works on two subjects are reviewed: *popularity detection* and *retweet prediction.*

A. Popularity Detection

This task deals with the problem of detecting popular new tweets at the first hours of publishing. In this task the popularity measure is not necessarily the number of retweets.

Some of the studies have used features like content, source user, category, subjectivity in the language, and named entities to predict popularity of new tweets [3]. According to the results, although these features are not sufficient, they are effective in predicting popularity.

On the other hand, there are other works that have used more specific approaches for this task like using a three-layer graph to rank tweets according to their trustworthiness and content popularity [4], calculating topics of tweets to evaluate significance of a given topic at each time point [5], or using a variation of HITS algorithm to detect valuable tweets [6]. According to their results, these approaches can improve precision of Twitter ranking scheme and have a better performance in detecting valuable tweets.

B. Retweet Prediction

The aim of this task is predicting the number of retweets that a tweet would achieve in future. This problem is usually formulated as a classification or regression task but in both cases, some features that have correlation with the number of retweets should be extracted.

Some works on this category have focused on content-based features [7], [8]. According to their results prediction of retweet count only by pure content-based features is possible.

However, the majority of works have considered both content and contextual features [9], [1], [10]. Despite, there is no agreement on the most informative features. Some works have mentioned that content-based features such as topics of tweets and availability of supplementary information have more impacts on predicting depth of deliberation of tweets [11]. In contrast, some other works have reported that user-related features are more important than those features that are related to content [12], [1]. Going beyond comparing type of features, some works have reported features like URLs, hashtags, number of followers, number of followees, and age of author's account as the most important features to predict retweet rate [13]. But obviously, as mentioned in [14], a combination of structural, content-based, and sentimental features is needed to perform this task effectively.

Some other works have used features of users who retweet the tweets in addition to content and contextual features [15]. The similarity of original tweet and users' previously retweeted tweets is an effective feature for predicting retweet action [16].

There are also some works that have designed some innovative methods like considering initial spread of cascades [17], using image-based features of media links in tweets [18], and proposing a two-phase model that first classifies tweets and then tries to predict retweet count by regression [19].

Besides all of these works, the questions of "what is a good content?" and "regardless of author, how much a good content can achieve popularity?" deserve more researches. In this paper, we focus on content and its lexical and semantic features to address these questions.

III. Data Set

We have collected a data set among the Persian speaking users of Twitter. Creating our data set has been done in two phases: (1) data collection and (2) tweet annotation.

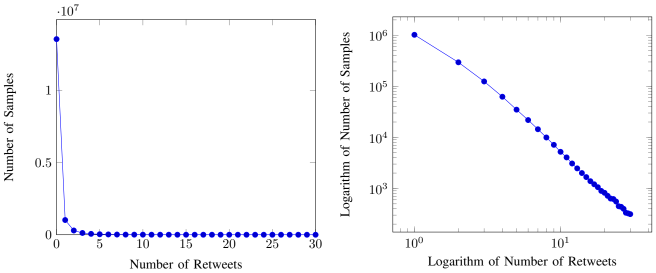

Fig. 1: Retweet distribution of collected tweets

TABLE I: Data set details

Class	Number of Samples	Number of Retweets
Class 1	3350	$= 0$
Class 2	3350	$\in \{1, 2, 3\}$
Class 3	3350	> 3

A. Data Collection

One of the our most important concerns was that tweets collected from Twitter should be a *vast* and *fair* sample with minimum bias. For this reason, we collected 15181525 tweets from Twitter for time period of February 2007 to August 2014 among original tweets (no retweet, reply, etc.) published by Persian speaking users. For all of these tweets, we also collected the number of favorites and the number of retweets as target features for measuring popularity.

After collecting tweets, we drew the histogram of number of retweets. As Figure 1 shows, it is a power law distribution. To choose around 10000 sample in a fair manner from the collected tweets, we calculated the average number of retweets:

$$M_{R>0} = \frac{1}{|R(t) > 0|} \sum_{R(t)>0} R(t) \qquad (1)$$

where $R(t)$ is the number of retweets of tweet t and $M_{R>0}$ is the mean of retweet count of those tweets that have at least one retweet. It was 2.28 on the collected tweets. So, we rounded this average up. Then, we divided the tweets into three parts based on the number of retweets: no rewteets, from 1 to 3 retweets ($\lceil M_{R>0} \rceil = 3$), and bigger than 3 retweets. We selected 3350 tweets randomly from each category summing up to 10050 tweets. Table I shows the details of data set.

B. Tweet Annotation

After sampling, we tried to label these tweets with the help of 77 volunteer annotators. In the labeling process, volunteers were asked to answer five questions related to semantic features of each tweet:

Funniness. How much funny is the tweet?

Individualness. How much does the tweet speak about individual issues and private life of the author?

Socialness. How much related is the tweet to public issues and society problems?

Positive Sentiment. How much positive sentiment does the tweet have?

Negative Sentiment. How much negative sentiment does the tweet have?

The answer could be selected among four options: nothing (= 0), low (= 1), mediocre (= 2), and much (= 3).

IV. Feature Vector

We extracted some features from the content of annotated tweets and prepared the final training set. For each tweet we extracted 42 features. Table II shows the feature vector and description of each feature.

First part of the feature vector belongs to features that are related to characters of the tweet. We calculated the number of characters in the original tweet. Then, we removed all components of the tweet except the words in order to calculate the number of characters that had appeared in those words. We also discretized it into 4 levels. Frequency of appearance of the most frequent characters in the tweet is another feature. We also calculated the number of punctuations and numbers that had appeared in the tweet.

Next part of the feature vector consists of features that depend on words. We calculated the number of words and stop words in the tweet. We also counted the number of words that are found in dictionary. We smoothed it by dividing it to the number of words. We also calculated the number and fraction of long words that have more characters than a threshold. Average length of words was calculated too.

Next subset of features belongs to those that are related somehow to the social network around users e.g. the number of hashtags and number of general hashtags in tweet. General hashtags are those tags that had appeared more than a threshold in the whole data set. Number of mentions, number of hyperlinks that point to a text file, number of hyperlinks that point to a media file, and number of all type of hyperlinks that had appeared in the tweet were calculated as well.

Next part of the feature vector consists of features that are related to emoticons appeared in the tweet. We counted the number of positive, negative, neutral emoticons, and the total number of emoticons.

As mentioned in section III, the values of funniness, individualness, socialness, positive sentiment and negative sentiment of each tweet were calculated by annotation. These features form the next part of the feature vector.

We also calculated the five semantic features for any words or hashtags that had appeared in tweets. For example, funniness of a word based on the funniness of tweets containing it, is calculated as:

$$F(w) = \frac{1}{|T^w|} \sum_{t \in T^w} F(t) \tag{2}$$

where T^w is set of tweets that have the word w and $|T^w|$ is its cardinality i.e. the total number of tweets that have w. In a quite similar manner, funniness of hashtag h based on the funniness of tweets containing it, is calculated as:

$$F(h) = \frac{1}{|T^h|} \sum_{t \in T^h} F(t) \tag{3}$$

Based on the funniness of words that have appeared in a tweet t, average funniness score of words of t could be calculated as:

$$\hat{F}^w(t) = \frac{1}{|w \in t|} \sum_{w \in t} F(w) \tag{4}$$

where $w \in t$ refers to all words w in tweet t. In a quite similar manner, average funniness score of hashtags of t could be calculated as:

$$\hat{F}^h(t) = \frac{1}{|h \in t|} \sum_{h \in t} F(h) \tag{5}$$

Other four semantic features can be calculated in a similar way for words and hashtags.

On the other hand, popularity of words and hashtags in terms of their frequency of appearance in the tweets can be a good feature that might be useful in predicting whether a tweet could ultimately get popular or not. So, a feature to measure average frequency of tweet's words is introduced this way:

$$\hat{Q}^w(t) = \frac{1}{|w \in t|} \sum_{w \in t} \log |T^w| \tag{6}$$

Instead of the frequency itself, we use its log for smoothing the effect of words with big frequencies. Average frequency of tweet's hashtags is similar to $\hat{Q}^w(t)$, but it is summed over all hashtags:

$$\hat{Q}^h(t) = \frac{1}{|h \in t|} \sum_{h \in t} \log |T^h| \tag{7}$$

Last two features of the feature vector are the number of favorites and the number of retweets. As mentioned in section III, they have been collected from Twitter in the data collection phase.

V. Experiments

A. Features Evaluation

In order to evaluate informativeness of the extracted features and their impacts on popularity of tweets, we use the Pearson correlation coefficient that shows how much linearly related are two variables:

$$\rho_{XY} = \frac{c_{XY}}{\sigma_X \sigma_Y} \tag{8}$$

where σ_X and σ_Y are population standard deviation and c_{XY} is population covariance of variables X and Y. The first variable could be logarithm of the number of either favorites or retweets. The second variable could be one of the features listed in Table II.

Table III shows the Pearson correlation coefficient of the features and logarithm of favorite count. The most correlated feature with the logarithm of favorite count is the logarithm of retweet count. It can be interpreted as a good upper bound for informativeness of the content-based features.

The next most correlated features are the number of hyperlinks of any type and the number of hyperlinks to web pages. Those have negative correlation with the number of favorites. It seems tendency of people to rapidly skim the tweets and postpone opening hyperlinks and reading long texts to future

TABLE II: Feature vector calculated for each tweet

Feature Name	Description
$N_C(t)$	Number of characters of tweet
$N_{WC}(t)$	Number of word-only characters of tweet
$N_{DWC}(t)$	Number of word-only characters of tweet discretized in $\theta_{DWC} = 4$ categories of equal size
$Q_{MFC}(t)$	Frequency of appearance of the $\theta_{MFC} = 2$ most frequent characters in tweet
$N_P(t)$	Number of punctuations in tweet
$N_N(t)$	Number of numbers in tweet
$N_W(t)$	Number of words in tweet
$N_{SW}(t)$	Number of stop words in tweet
$N_{DW}(t)$	Number of dictionary words in tweet
$D_{DW}(t)$	Fraction of dictionary words in tweet
$N_{LW}(t)$	Number of words in tweet that have more than $\theta_{LW} = 6$ characters
$D_{LW}(t)$	Fraction of words in tweet that have more than $\theta_{LW} = 6$ characters
$\hat{L}_W(t)$	Average length of words in tweet
$N_H(t)$	Number of hashtags in tweet
$N_{GH}(t)$	Number of general hashtags appeared more than $\theta_{GH} = 4$ times in the whole data set
$N_M(t)$	Number of mentions in tweet
$N_L(t)$	Number of all hyperlinks in tweet
$N_{TL}(t)$	Number of hyperlinks in tweets that point to a text file
$N_{ML}(t)$	Number of hyperlinks in tweet that point to a media file
$N_E(t)$	Number of all emoticons in tweet
$N_{+E}(t)$	Number of positive emoticons in tweet
$N_{-E}(t)$	Number of negative emoticons in tweet
$N_{NE}(t)$	Number of neutral emoticons in tweet
$F(t)$	Funniness of tweet
$I(t)$	Individualness of tweet
$S(t)$	Socialness of tweet
$P_+(t)$	Positive sentiment of tweet
$P_-(t)$	Negative sentiment of tweet
$\hat{F}^w(t)$	Average funniness score of tweet's words
$\hat{I}^w(t)$	Average individualness score of tweet's words
$\hat{S}^w(t)$	Average socialness score of tweet's words
$\hat{P}^w_+(t)$	Average positive sentiment score of tweet's words
$\hat{P}^w_-(t)$	Average negative sentiment score of tweet's words
$\hat{F}^h(t)$	Average funniness score of tweet's hashtags
$\hat{I}^h(t)$	Average individualness score of tweet's hashtags
$\hat{S}^h(t)$	Average socialness score of tweet's hashtags
$\hat{P}^h_+(t)$	Average positive sentiment score of tweet's hashtags
$\hat{P}^h_-(t)$	Average negative sentiment score of tweet's hashtags
$\hat{Q}^w(t)$	Average frequency of tweet's words
$\hat{Q}^h(t)$	Average frequency of tweet's hashtags
$V(t)$	Number of favorites of tweet
$R(t)$	Number of retweets of tweet

is the cause of these results. Users of Twitter have actually got used to read short texts. They might spend their time on reading large text articles but they need social evidences i.e. recommendation of other users through retweet or favorite.

The next correlated features are the average funniness score of tweet's words and the funniness of tweet. Not surprisingly, when a tweet has higher funniness score it would receive more favorites. So, according to the feature definition, words that appear more in funny tweets, have more funniness scores too and both of these features have almost the same effects. Other lexical and semantic features have less correlation with the logarithm of favorite count.

Table IV shows the Pearson correlation coefficient of the features and the logarithm of retweet count. Aside from the logarithm of favorite count that is the most correlated feature, the socialness of tweet is the most correlated feature with the logarithm of retweet count. This clearly confirms the idea that when a tweet has social content, it is more attractive for users and gets more retweets.

The next most correlated features are three features all related to hashtags. Hashtags perform an important role in diffusion of information because they are traceable. Tweets that have popular hashtags can be searched by users and thus have more chance to be seen by other users and gain more retweets. Again between hashtag-related features, the average socialness score of hashtags has larger correlation with the logarithm of retweet count.

Even after these features, the average socialness score of tweet's words, also connected to the socialness of tweets, has a large correlation. The average frequency of tweet's hashtags is the next correlated feature. It indicates that common hashtags get more retweets.

However, negative correlation of the individualness of tweet and the logarithm of retweet count supports the other idea that tweets with individual content are less interesting for society. Even the average individualness score of tweet's words as the next most correlated feature confirms this phenomena.

The average negative and positive sentiment score of

TABLE III: Pearson correlation coefficient between all of the features and logarithm of favorite count

Feature Name	Pearson Correlation Coefficient	P-Value
R	0.692	0
N_L	-0.307	$6.03E-221$
N_{TL}	-0.291	$5.86E-198$
$\hat{F}^w$	0.238	$2.16E-130$
F	0.205	$4.84E-97$
N_C	-0.157	$2.31E-57$
N_{SW}	0.137	$1.27E-43$
N_{WC}	0.103	$3.80E-25$
$\hat{Q}^h$	-0.1	$7.11E-24$
N_{DWC}	0.098	$7.07E-23$
N_{ML}	-0.097	$9.43E-23$
N_W	0.093	$9.75E-21$
Q_{MFC}	0.093	$9.76E-21$
$\hat{P}^w_-$	0.092	$2.31E-20$
N_{GH}	-0.089	$2.13E-19$
N_N	-0.079	$1.47E-15$
$\hat{I}^w$	0.068	$8.99E-12$
N_{DW}	0.062	$3.02E-10$
P_-	0.057	$7.49E-09$
$\hat{F}^h$	0.056	$1.35E-08$
N_M	-0.055	$3.46E-08$
D_{LW}	-0.054	$6.69E-08$
S	0.053	$7.88E-08$
N_P	-0.052	$1.72E-07$
D_{DW}	-0.052	$1.85E-07$
N_H	-0.051	$3.38E-07$
$\hat{P}^h_-$	-0.046	$3.11E-06$
$\hat{P}^h_+$	0.042	$2.76E-05$
$\hat{S}^h$	-0.04	$5.05E-05$
$\hat{P}^w_+$	0.038	$1.11E-04$
P_+	0.037	$1.73E-04$
$\hat{I}^h$	0.032	$1.39E-03$
$\hat{S}^w$	-0.031	$1.67E-03$
N_E	0.027	$6.25E-03$
N_{+E}	0.022	$2.50E-02$
I	-0.021	$3.13E-02$
N_{NE}	0.02	$4.41E-02$
$\hat{L}_W$	0.012	$2.30E-01$
N_{LW}	-0.009	$3.83E-01$
N_{-E}	-0.002	$8.63E-01$
$\hat{Q}^w$	-0.001	$9.02E-01$

TABLE IV: Pearson correlation coefficient between all of the features and logarithm of retweet count

Feature Name	Pearson Correlation Coefficient	P-Value
V	0.692	0
S	0.243	$3.00E-136$
$\hat{S}^h$	0.233	$1.91E-125$
N_H	0.229	$2.06E-120$
N_{GH}	0.224	$8.22E-116$
$\hat{S}^w$	0.224	$2.20E-115$
$\hat{Q}^h$	0.211	$4.32E-102$
I	-0.2	$4.10E-92$
$\hat{I}^w$	-0.191	$8.59E-84$
$\hat{P}^h_-$	0.163	$5.01E-61$
N_{ML}	0.138	$3.48E-44$
$\hat{P}^h_+$	0.133	$3.16E-41$
N_{TL}	-0.126	$3.53E-37$
N_W	0.124	$2.44E-36$
N_{WC}	0.121	$1.76E-34$
N_{DWC}	0.119	$1.65E-33$
N_{DW}	0.117	$3.20E-32$
Q_{MFC}	0.11	$1.08E-28$
N_{LW}	0.084	$2.35E-17$
N_{SW}	0.072	$3.46E-13$
$\hat{Q}^w$	0.069	$4.69E-12$
$\hat{F}^h$	0.059	$3.36E-09$
F	0.057	$7.72E-09$
$\hat{I}^h$	0.053	$1.10E-07$
P_-	0.049	$7.66E-07$
N_C	0.048	$1.64E-06$
$\hat{P}^w_-$	0.046	$3.08E-06$
N_E	-0.045	$5.51E-06$
N_{+E}	-0.039	$9.15E-05$
N_{-E}	-0.02	$4.46E-02$
N_N	0.019	$5.56E-02$
D_{LW}	0.019	$5.61E-02$
N_{NE}	-0.017	$8.68E-02$
$\hat{P}^w_+$	-0.016	$1.05E-01$
D_{DW}	0.015	$1.30E-01$
N_L	-0.012	$2.27E-01$
$\hat{L}_W$	-0.012	$2.32E-01$
$\hat{F}^w$	0.009	$3.53E-01$
N_M	-0.005	$6.38E-01$
N_P	-0.002	$8.59E-01$
P_+	0	$9.98E-01$

tweet's hashtags are the next most correlated features. They show that embedding sentiments in hashtags is a good tactic to achieve more retweets. Other lexical and semantic features have less correlation with the logarithm of retweet count.

B. *Prediction of Popularity Class*

As mentioned in section III, we have three popularity classes based on the number of retweets of samples in our data set. We defined three classification tasks based on these class labels. In the classification task 1, we aim to predict the original class labels that were defined in section III. In the classification task 2, we attempt to classify samples into two classes: those that have at least one retweet and those with no retweets. Finally, in the classification task 3, we classify samples into two classes: those that have more than average retweets i.e. $\lceil M_{R>0} \rceil$ and those that have less than equal to average retweets.

Before classification, we performed the Principal Component Analysis (PCA) and Linear Discriminant Analysis (LDA) transformation on the data set in order to minimize correlation of the features and maximize their separability power in classifying different classes. We train four classifiers and evaluate them by 10-fold cross validation. As a baseline, we use a Naive Bayes classifier trained only with features that are independent of annotations. Obviously, we do not use the number of favorites and retweets in the classification tasks.

Table V shows the results of classification tasks. According to the results, popularity classes of tweets can be predicted to some extents using only the content-based features. However, the semantic features that we introduced could improve the performance of models. It is worth mentioning that the purpose of using only the content-based features is to find out how much content-based features can contribute to the prediction of popularity classes of tweets. It is obvious that considering only the content-based features are not sufficient for predicting the number of retweets.

C. *Sentiment Features*

We also found that correlations between the number of positive, negative, and neutral emoticons in the tweets and the actual positive and negative sentiments of the tweets obtained by annotation process are not significant. Table VI shows the Pearson correlation coefficient between these features. According to the results, emoticons are too noisy and using them for labeling sentiment of tweets is not a suitable approach.

TABLE V: Accuracy of the classifiers in different classification tasks

Classifier	Task 1	Task 2	Task 3
Baseline	0.443	0.702	0.684
KNN	0.480	0.706	0.687
Binary Decision Tree	0.413	0.635	0.617
Naive Bayes	**0.486**	**0.711**	**0.694**
SVM	0.484	0.528	0.681

TABLE VI: Pearson correlation coefficient of the sentiment features

Number of Emoticons	Actual Positive Sentiment	Actual Negative Sentiment
Number of Positive Emoticons	$0.304\ (3.54E-19)$	$-0.480\ (3.41E-49)$
Number of Negative Emoticons	$-0.136\ (8.76E-05)$	$0.258\ (4.46E-14)$
Number of Neutral Emoticons	$-0.224\ (6.89E-11)$	$0.347\ (5.78E-25)$

VI. Conclusion

In this paper, we performed a comprehensive analysis of the lexical and semantic features of tweet content and their impacts on popularity of tweet. We collected a fair sample of tweets from Twitter and chose the number of retweets of tweets as popularity measure. Then, we defined five semantic features including *funniness*, *individualness*, *socialness*, *positive sentiment*, and *negative sentiment* and extracted them from tweets through annotation process. In order to evaluate the extracted features, we used the Pearson correlation coefficient. We also trained predictive models to predict retweet class by only the content-based features.

We showed that tweets with social content are more popular among users. In contrast, tweets with individual content have very little chance to gain users' attention. In addition, the results of the classification tasks showed predicting popularity classes of tweets by only the content-based features is possible. We also found that correlation between emoticons and the actual sentiments of tweets is not significant and emoticons are too noisy to be used as a gold standard for sentiment analysis.

There are plenty of works that can be done in future. To name a few, extracting comprehensive features from authors of the tweets, adding contextual and structural features to improve predictive models, and designing a method for extracting the semantic features in an automated manner could be done in future.

References

[1] S. Petrovic, M. Osborne, and V. Lavrenko, "Rt to win! predicting message propagation in twitter." in *ICWSM*, 2011.

[2] S. Y. Syn and S. Oh, "Why do social network site users share information on facebook and twitter?" *Journal of Information Science*, p. 0165551515585717, 2015.

[3] R. Bandari, S. Asur, and B. A. Huberman, "The pulse of news in social media: Forecasting popularity," *CoRR*, vol. abs/1202.0332, 2012. [Online]. Available: http://arxiv.org/abs/1202.0332

[4] S. Ravikumar, R. Balakrishnan, and S. Kambhampati, "Ranking tweets considering trust and relevance," *CoRR*, vol. abs/1204.0156, 2012. [Online]. Available: http://arxiv.org/abs/1204.0156

[5] C.-H. Lee and T.-F. Chien, "Leveraging microblogging big data with a modified density-based clustering approach for event awareness and topic ranking," *Journal of Information Science*, p. 0165551513478738, 2013.

[6] M.-C. Yang, J.-T. Lee, S.-W. Lee, and H.-C. Rim, "Finding interesting posts in twitter based on retweet graph analysis," in *Proceedings of the 35th International ACM SIGIR Conference on Research and Development in Information Retrieval*, ser. SIGIR '12. New York, NY, USA: ACM, 2012, pp. 1073–1074. [Online]. Available: http://doi.acm.org/10.1145/2348283.2348475

[7] 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*, ser. WebSci '11. New York, NY, USA: ACM, 2011, pp. 8:1–8:7. [Online]. Available: http://doi.acm.org/10.1145/2527031.2527052

[8] Y. Arakawa, A. Kameda, A. Aizawa, and T. Suzuki, "Adding twitter-specific features to stylistic features for classifying tweets by user type and number of retweets," *Journal of the Association for Information Science and Technology*, vol. 65, no. 7, pp. 1416–1423, 2014.

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

[10] 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*, ser. WWW '11. New York, NY, USA: ACM, 2011, pp. 57–58. [Online]. Available: http://doi.acm.org/10.1145/1963192.1963222

[11] L. Zhang, T. Peng, Y. Zhang, and X. Wang, "Content or context: Which carries more weight in predicting popularity of tweets in china," in *Proceedings of the 65th Annual Conference of World Association for Public Opinion Research*, 2012.

[12] Z. Xu and Q. Yang, "Analyzing user retweet behavior on twitter," in *Proceedings of the 2012 International Conference on Advances in Social Networks Analysis and Mining (ASONAM 2012)*, ser. ASONAM '12. Washington, DC, USA: IEEE Computer Society, 2012, pp. 46–50. [Online]. Available: http://dx.doi.org/10.1109/ASONAM.2012.18

[13] 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 *Social Computing (SocialCom), 2010 IEEE Second International Conference on*, Aug 2010, pp. 177–184.

[14] M. Jenders, G. Kasneci, and F. Naumann, "Analyzing and predicting viral tweets," in *Proceedings of the 22Nd International Conference on World Wide Web Companion*, ser. WWW '13 Companion. Republic and Canton of Geneva, Switzerland: International World Wide Web Conferences Steering Committee, 2013, pp. 657–664. [Online]. Available: http://dl.acm.org/citation.cfm?id=2487788.2488017

[15] *Predicting Information Spreading in Twitter*, December 2010. [Online]. Available: http://research.microsoft.com/apps/pubs/default.aspx?id=141866

[16] Z. Luo, M. Osborne, J. Tang, and T. Wang, "Who will retweet me?: finding retweeters in twitter," in *Proceedings of the 36th international ACM SIGIR conference on Research and development in information retrieval*. ACM, 2013, pp. 869–872.

[17] A. Kupavskii, L. Ostroumova, A. Umnov, S. Usachev, P. Serdyukov, G. Gusev, and A. Kustarev, "Prediction of retweet cascade size over time," in *Proceedings of the 21st ACM International Conference on Information and Knowledge Management*, ser. CIKM '12. New York, NY, USA: ACM, 2012, pp. 2335–2338. [Online]. Available: http://doi.acm.org/10.1145/2396761.2398634

[18] E. F. Can, H. Oktay, and R. Manmatha, "Predicting retweet count using visual cues," in *Proceedings of the 22nd ACM international conference on Conference on information &#38; knowledge management*, ser. CIKM '13. New York, NY, USA: ACM, 2013, pp. 1481–1484. [Online]. Available: http://doi.acm.org/10.1145/2505515.2507824

[19] G. Liu, C. Shi, Q. Chen, B. Wu, and J. Qi, "A two-phase model for retweet number prediction," in *Web-Age Information Management*, ser. Lecture Notes in Computer Science, F. Li, G. Li, S.-w. Hwang, B. Yao, and Z. Zhang, Eds. Springer International Publishing, 2014, vol. 8485, pp. 781–792. [Online]. Available: http://dx.doi.org/10.1007/978-3-319-08010-9_84