An Analysis of Influential Users for Predicting the Popularity of News Tweets

Krissada Maleewong(✉)

School of Information Technology, Shinawatra University, Pathumthani, Thailand
krissada@siu.ac.th

Abstract. Twitter plays an important role in today social network. Its key mechanism is retweet that disseminates information to broad audiences within a very short time and help increases the popularity of the social content. Therefore, an effective model for predicting the popularity of tweets is required in various domains such as news propagation, viral marketing, personalized message recommendation, and trend analysis. Although many studies have been extensively researched on predicting the popularity of tweets, they mainly focus on the content-based and the author-based features, while retweeter-based features are less concerned. This paper aims to study the impact of influential users who retweet tweets, also called retweeters, and presents simple yet effective measures for predicting the influence of retweeters on the popularity of online news tweets. By analyzing the popularity of news tweets and the impact of the retweeters, a number of useful measures are defined to evaluate influence of users in the retweeter network, and used to establish the prediction model. The experimental results show that the application of the retweeter-based features is highly effective and enhances the performance of the prediction model with high accuracy.

Keywords: Twitter · Retweet · Influential user · Active user · Popular user · News tweet · Social network

1 Introduction

Nowadays, Twitter is considered as the most prominent micro-blogging service available on the Web. It allows people to publish 140-character short messages known as tweets, which can also contain images, videos, or URLs that link to the original online sources. In the Twitter network, users (a.k.a. *followers*) can follow other notable users (a.k.a. *followees*) to gain real-time updates on news and statuses. When a user finds an interesting tweet written by another user and wants to share it with his/her followers, the user (a.k.a. *retweeter*) can retweet such a tweet by either using a retweet button or manually editing the original tweet and adding a text indicator (e.g., RT @*user* or via @*user*) to mention that the original tweet came from the specified user. Therefore, retweeting mechanism is an important technique for information diffusion in Twitter and utilized in several applications such as breaking news detection, personalized message recommendation, viral marketing, trend analysis, and Twitter-based early warning systems. By focusing on online news, Twitter is adopted as a new medium for

R. Booth and M.-L. Zhang (Eds.): PRICAI 2016, LNAI 9810, pp. 306–318, 2016.
DOI: 10.1007/978-3-319-42911-3_26

disseminating news from their websites to the readers. Several news sources such as BBC News, CNN, and Bangkok Post distribute latest news to readers using their Twitter, while readers read the articles and might follow the URLs attached in the news tweets to the original news sources for further reading. A reader who finds an interesting news tweet can retweet the news tweets to his/her followers. This mechanism results in not only the popularity of news tweets but also the popularity of the root of news pages published on the news websites. Accordingly, the retweeting mechanism can help people to access the news faster and empowers the online news channels to widespread their content in social network.

One important factor that affects the popularity of news tweets is users. In this paper, a user who retweets tweets is called a *retweeter*. Since different retweeters have unequally influence, an analysis of the impact of influential retweeters on the popularity of tweets is required. In addition, understanding how a tweet becomes popular can help to gain a better insight into how the information is dispersed over the social network. However, predicting the popularity of news tweet is a challenging task comparing to other kinds of tweet due to theirs short life time. Therefore, this paper aims to study important features that impact the popularity of news tweets based on retweeting mechanism and retweeter-based features. By analyzing the influence of different types of retweeters, a number of useful measures are proposed to evaluate the influence of retweeters. Based on the proposed measures, a number of retweeters-based features are introduced for predicting the popularity of news tweets.

The organization of this paper is as follows: Sect. 2 discusses the related work. Section 3 analyzes the popularity of news tweets and studies the impact of influential retweeters. Section 4 presents the prediction model. Section 5 performs an experiment and reports the experiment results. Section 6 draws conclusions and future research direction.

2 Related Work

Many studies have been researched on predicting the popularity of tweets by proposing a variety of features such as content-based features, contextual or author-based features, network structural features, and temporal features. These features describe the characteristics and the past evolution of tweets, as well as their social interaction. By applying a generalized linear model, the content-based features (i.e., numbers of URLs and hashtags), and the author-based features (i.e., number of followers and followees, and age of the account) are investigated in order to calculate the retweetability of tweets (Suh et al. 2010). Later, various types of features are introduced including the content-based and temporal information of tweets, metadata of tweets and users, as well as structural properties of the users in social network for estimating the number of future retweets using binary and multi-class classification models (Hong et al. 2011). By concerning a set of social features (i.e., number of followers, friends, statuses, favorites, and number of times the user was listed), the propagation of tweets is predicted based on the passive-aggressive algorithm (Petrovic et al. 2011). In addition, the evolution of retweets is predicted based on the size of the retweeter network and the depth from the

source of tweets using a probabilistic model (Zaman et al. 2014). By investigating the network of SinaWeibo, the biggest microblogging system in China, the *structural characteristics* (Bao et al. 2013) incorporates the early popularity with the link density and the diffusion depth of early adopters for predicting the popularity of short messages. To predict the popularity of newly emerging hashtags in Twitter (Ma et al. 2013), a set of content-based and author-based features is applied to five standard classification models (i.e., Naïve bayes, k-nearest neighbors, decision trees, support vector machines, and logistic regression). The results show that the logistic regression model performs the best but the experiment relaxed the problem and predicted the range of popularity instead of the exact value of the popularity.

With emphasis on analyzing the influence of users in social network, various measures have been introduced. *FollowerRank* (Nagmoti et al. 2010) and *Structural Advantage* (Cappelletti and Sastry 2012) adapt the traditional *in-degree measure* (Hajian and White 2011) for determining the popularity of a user. *Follower-Followee ratio* (Bigonha et al. 2012) and *Paradoxical Discounted* (Gayo-Avello 2013) consider numbers of followers and followees for detecting spammers (users with many followees and few followers). By concerning influential users as celebrities, *StarRank* (Khrabrov and Cybenko 2010) applies PageRank algorithm for determining the acceleration of mentions over time, while *Acquaintance Score* (Srinivasan et al. 2013) utilizes numbers of mentions, replies, and retweets. In order to rank users based on their activities, *TweetRank* (Nagmoti et al. 2010) counts the number of tweets of the user, and *TweetCountScore* (Neves et al. 2015) counts the number of tweets of a user plus the number of retweets. Based on time concerning, *Effective readers* (Lee et al. 2010) measures the speed of a user to tweet about a new topics, while *ActivityScore* (Yuan et al. 2013) counts the number of followers, followees, and tweets of each user during a period of time. *IP Influence* (Romero et al. 2011) evaluates the influence of the users and their passivity using metrics of retweets, followers, and followees. However, most researches have focused on the content-based features of the tweets and the author-based features of the authors who post the tweets, while the the retweeter-based features of users who help disseminate the tweets are less concerned in the prediction models.

3 Analysis of Influential Users in Popular News Tweets

This section describes the dataset and presents data analysis of the characteristics of popular news tweets as well as the impact of influential retweeters.

3.1 Dataset and Popularity of News Tweets

The dataset was collected from BBC News Twitter using Twitter Streaming API. It provides various topics including business, entertainment, science, sport, and weather, etc. The dataset contains 192,821 retweet data collected from 3,336 news tweets posted before December 2015 to make sure that there were likely to be no more retweet occurring.

In most of the recent works, the popularity of tweets is evaluated by the number of retweets because it is the most effective measure to disseminate messages comparing to other metrics such as the number of favorites or replies. In this paper, the *popularity* of news tweets, therefore, refers to the total number of retweets that the tweets receive. Due to problems with identifying the connection between the root tweet and the subsequent manual retweets in the dataset retrieved from Twitter's API, and the previous studies report that retweet graphs typically have most vertices at depth one, suggesting that root tweets (posted by the authors) get retweeted much more often than the retweets get retweeted (Kwak et al. 2010; Goel et al. 2012). This study, therefore, considers only the retweets made using the retweet button and the reduced dataset contains 163,312 retweets collected from 2,545 news tweets. This whole dataset was randomly separated into two smaller sets with a 70:30 ratio for training and testing the models, respectively. Figure 1 shows the frequency distribution of the popularity of the reduced dataset. The histogram demonstrates the skewed right distribution indicating that a few number of news tweets (1 %) receive high popularity (greater than 1,000 retweets), while about 90 % of news tweets gain the popularity between 0–150 retweets. The most popular tweet gains 2,865 retweets, and the median and the average of the popularity are 50 and 75 retweets, respectively.

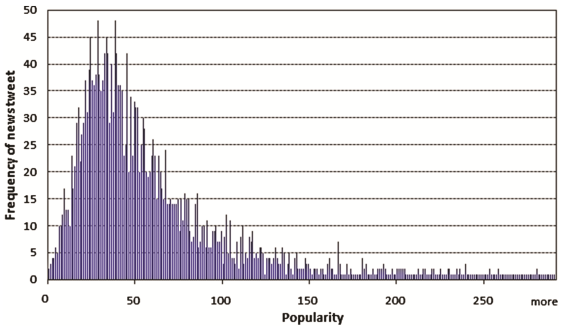

Fig. 1. Frequency of the popularity of news tweets.

To study the evolution of the popularity, Fig. 2 illustrates the correlation of the retweet times and the cumulative number of retweets for the four popularity levels: minimum (2 retweets), median (50 retweets), average (75 retweets), and maximum (2,865 retweets). In Fig. 2, the numbers of retweets for all popularity levels rapidly increase in the early stage after the tweets were posted (within 1,000–10,000 s, which is difference from other kinds of tweets that ranged from many hours to many days (Ma et al. 2013; Szabo and Huberman 2010) and then are almost stable. This makes the prediction model for news tweets more challenging. The time for the final retweet to occur for all levels ranged from 1.5 h to 6 days. In order to identify the prediction time with the highest performance of the prediction model, the time to reach 50 % of number of retweets for the median popularity (50 retweets) is adopted, which is 550 s. Note that the prediction time can be defined differently in which the time close to the time of final retweet yields higher accuracy but it might be too late for enhancing the popularity of

the news tweets, while the less time gains lower accuracy (e.g., most of tweets may receive similar number of retweets at 10 s after the tweets are posted) but have enough time to disseminate the tweets to the right target group or influential users. This temporal analysis is considered as an interesting open research question.

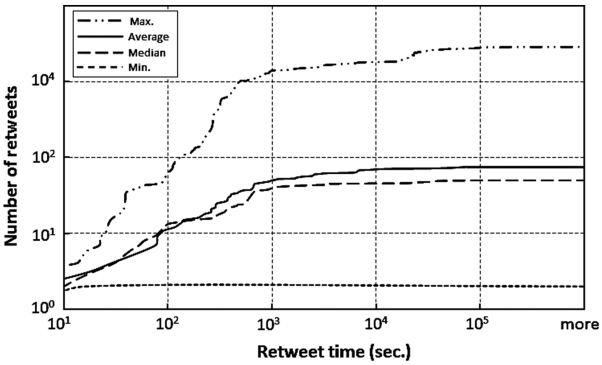

Fig. 2. Correlation of the retweet time and the cumulative number of retweets for the four popularity levels.

3.2 Impact of the Influential Users

The popularity of news tweets relies on many factors and an important factor is users. This Subsection, therefore, studies the impact of the influential retweeters on the popularity of news tweets. The retweeters are distinguished into two types including (*i*) *active user* and (*ii*) *popular user* as described following.

3.2.1 Active User

An *active user* is a person who has stable and frequent participation in the social network during a period of time (Fabián 2015; Yin and Zhang 2012). However, the activity rate of a user often fluctuates with time and it is costly to calculate the activity rate at any given time point. Thus, this study defines an active user based on user's participation during his/her Twitter's account lifetime. Let U be the set of users. For a user u U, his/her activity rate, denoted by $ActivityRate(u)$, is defined to specify the participation of user u and is calculated by counting the number of tweets plus the number of retweets posted and shared by the user throughout his/her Twitter's account lifetime, which could then be formally defined as follows:

$$ActivityRate(u) = \frac{T_u + RT_u}{accountTime(u)}, \tag{1}$$

where

– u is a retweeter,
– T_u is the number of tweets posted by user u,

– RT_u is the number of retweets shared by user u, and
– *accountTime* u is the Twitter's account lifetime of user u in day unit.

A user who participates in posting high number of tweets and retweeting many tweets during his/her account lifetime is considered as an active user. Note that the number of tweets is considered in order to avoid users who have high number of retweets without posting any tweet, whose tend to be spammers or bots (Messias et al. 2013).

3.2.2 Popular User

A *popular user*, also called a celebrity, is a person who is recognized or followed by many other users on the network (Fabián 2015). Many recent researches evaluate the popularity of a user by concerning his/her number of followers and/or followees. A user who has high number of followers or high ratio of followers and followees is considered as a popular user. On the other hand, this research determines the popularity of a user by focusing on the number of *list*. In twitter, a user creates a list to organize and keep a closer eye on his/her followees into a group related to a specific topic such as celebrities, technical leaders, or news lists. Hence, a popular user, evaluated based on the list, is an expert or an influencer in a specific field considered based on the community point of view. For a user $u \quad U$, his/her popularity score, denoted by $PopularityScore(u)$, is defined to specify the recognition of a user in the Twitter's network. Intuitively, it is computed by counting the number of lists that user u is listed, which could be defined as follows:

$$PopularityScore(u) = L_u, \tag{2}$$

where

– u is a retweeter, and
– L_u is the number of lists that user u is listed.

Thus, a user who is listed in many lists is considered as a popular user.

Figure 3(a) shows the correlation between the *ActivityRate* of retweeters and the popularity of news tweets, while Fig. 3(b) shows the correlation between the *PopularityScore* of retweeters and the popularity of news tweets. The figures plot the averages of the

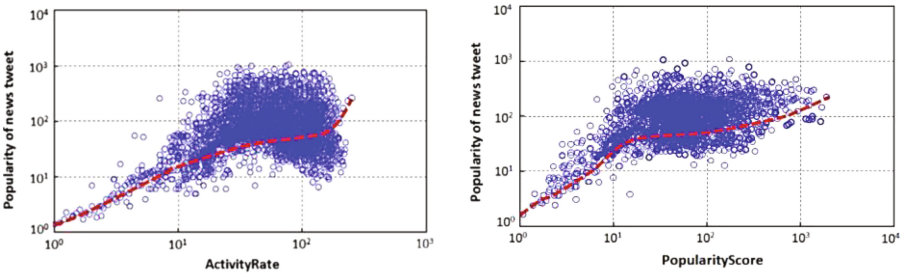

Fig. 3. Correlation between (a) *ActivityRate* and (b) *PopularityScore* and the popularity of news tweets.

ActivityRate and *PopularityScore* of retweeters for each news tweet against the popularity of the tweets retweeted by the retweeters, while the medians of the *ActivityRate* and the *PopularityScore* per bin are drawn in the dashed line. Obviously, highly popular tweets are mainly retweeted by users who have high activity rate and high *PopularityScore*. The average of the *ActivityRate* and the *PopularityScore* of retweeters are above the medians, indicating that there are many retweeters who have very high activity rate and concerned as popular users retweet these popular tweets. Therefore, a tweet retweeted by retweeters who have high *ActivityRate* and *PopularityScore* is likely to be a popular tweet. According to the linear relationships, the *ActivityRate* and the *PopularityScore* features are introduced for predicting the popularity of news tweets (Fig. 3).

4 The Prediction Model

This section presents the feature-based approach for predicting the popularity of the news tweets by means of a multiple linear regression. The features of the retweeters, the content, and the authors are described.

Table 1. Features for the popularity prediction

Name	Description
Retweeter-based features	
Number of Retweeters	Number of all retweeters who retweet a tweet
ActivityRate	Average of *ActivityRate* of all retweeters who retweet a tweet
TweetCountScore	Average of *TweetCountScore* of all retweeters who retweet a tweet
PopuralityScore	Average of *PopuralityScore* of all retweeters who retweet a tweet
FollowerRank	Average of *FollowerRank* of all retweeters who retweet a tweet
Content-based features	
Hashtag	Number of hashtags in a tweet
Mention	Number of usernames mentioned in a tweet (excluding "RT @username")
URL	Number of URLs found in a tweet
Media	Number of attached photos or videos
Tweet period	Period of tweet time (e.g., after midnight, morning, afternoon, evening)
Tweet day	Day of tweet's time (e.g., weekday or weekend)
Author-based features	
Follower	Number of followers of the author who posts a tweet
Followee	Number of followees of the author who posts a tweet
Status	Number of tweets and retweets made by the author who posts a tweet

4.1 Features

Table 1 summarizes the features for predicting the popularity of the news tweets, divided into three groups including: (*i*) *retweeter-based*, (*ii*) *content-based*, and (*iii*) *author-based features*.

(i) ***Retweeter-based features*** describe the characteristics of retweeters.
- *Number of Retweeters*: Number of all retweeters who retweet a tweet.
- *ActivityRate*: This feature finds the average of *ActivityRate* of retweeters who retweet a tweet.
- *TweetCountScore*: By applying the *TweetCountScore* (Noro et al. 2012), this feature calculates the average of *TweetCountScore* of all retweeters. The *TweetCountScore* counts the number of tweets plus the number of retweets shared by a retweeter u, and is divided by the *TweetCountScore* of another retweeter u' who post the maximum number of tweets and retweets, which can be computed as follows:

$$TweetCountScore(u) = \frac{T_u + RT_u}{\max(TweetCountScore(u'))}.$$

A retweeter who posts and retweets many tweets is considered as an active user.
- *PopularityScore*: This feature computes the average of *PopularityScore* of all retweeters who retweet a tweet.
- *FollowerRank*: By applying the *FollowerRank* (Nagmoti et al. 2010), this feature finds the average of *FollowerRank* of all retweeters who retweet a tweet. The *FollowerRank* evaluates the ratio of followers and followees of a retweeter u, which can be evaluated as follows:

$$FollowerRank(u) = \frac{\#followers(u)}{\#followers(u) + \#followees(u)}.$$

A retweeter who has high number of followers and less number of followees is concerned as a popular user.

(ii) ***Content-based features*** explain the content attributes of a tweet.
- *Hashtag*: Number of hashtags found in a tweet. In Twitter, a hashtag is frequently used to identify a topical keyword (e.g., #mufc identifies that the tweet is classified in mufc topic).
- *Mention*: Number of usernames mentioned in a tweet. Usually, an ampersand symbol is used for mentioning to another user (e.g., @POTUS refers to the President Barack Obama).
- *URL*: Number of URLs found in a tweet.
- *Media*: Number of photos or videos attached in a tweet.
- *Tweet period*: Period of tweet's time in day, which can be divided into four slots including *after midnight*, *morning*, *afternoon*, and *evening*.
- *Tweet day*: Day of tweet's time, which can be classified into two groups including *weekday* and *weekend*.

(iii) ***Author-based features*** identify social characteristics of an author who posts a tweet.
- *Follower*: Number of followers of the author who posts a tweet.
- *Followee*: Number of followees of the author who posts a tweet.

- *Status*: Number of tweets and retweets made by the author who posts a tweet. Intuitively, it specifies the participations of the author in Twitter's network.

4.2 Modeling the Prediction of Popular Tweets

Based on the features formally defined in the previous section, the prediction model has been developed using a multiple linear regression to predict the popularity of news tweets. In previous researches (Suh et al. 2010; Hong et al. 2011; Ma et al. 2013), the prediction problem was relaxed by predicting the *range* of the popularity. In contrast, this research aims at predicting the *exact value* of the popularity which can be achieved using the multiple linear regression. Several key assumptions of the multiple linear regression analysis were investigated and summarized (Due to limitation of space, the statistical results are omitted). Firstly, the linear relationships between the independent and dependent variables were examined using scatterplots and the results show that all retweeter-based features, author-based features, and media feature have different level of strong linear relationship, while most of the content-based features present a roughly linear relationship. Secondly, the normal distribution was investigated using a histogram. The results show that the data is normally distributed. Thirdly, there is no serious multicollinearity found in the dataset and the highest collinearity is *mention* and *PopularityScore* features with an acceptable value of 0.48. In addition, the homoscedasticity was investigated by creating a scatterplot of the residuals against the independent variables. The distributions of residuals show that there is no serious heteroscedasticity. Hence, the abovementioned features were applied to the regression model for predicting the popularity of news tweets, while the prediction time was set to 550 s (time to reach 50 % of the median popularity). Let NT be the set of news tweets, the popularity of news tweet $nt \in NT$, denoted by $P(nt)$ could then be calculated as follows:

$$P(nt) = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \cdots + \beta_i x_i,$$

where

- β_0 is the intercept term,
- $\beta_1, \beta_2, \ldots, \beta_i$ are the regression coefficient of features $x_1, x_2, \ldots, x_i$ respectively.

5 Experiment

This section presents an experimental setup including the evaluation metrics and the popularity predicting methods used to compare the experimental results. In addition, the experimental results are reported and discussed.

5.1 Evaluation Metrics and Methods Compared

Three standard performance metrics are applied as described below:

- *Root Mean Squared Error* (*RMSE*) is the square root of the average squared distance of a data point from the fitted line. However, one important limitation of squared errors is that it often places emphasis on the effect of outliers.

- *Mean Absolute Error* (*MAE*) is a quantity used to measure how close forecasts or predictions are to the eventual outcomes.
- *Adjusted* R^2 is the most widely used and reported measure of error and goodness-of-fit of a regression model. Based on R^2, *adjusted* R^2 calculates the proportion of the variation in the dependent variable accounted by the explanatory variables.

The performance metrics compare the efficiency of the methods that applied different features as depicted in Table 2.

Table 2. Evaluation methods.

Features	Method							
	Baseline	I	II	III	IV	V	VI	VII
Retweeter-based features								
Number of Retweeters		√						
ActivityRate			√				√	
TweetCountScore				√				√
PopularityScore					√		√	
FollowerRank						√		√
Content-based features	√	√	√	√	√	√	√	√
Author-based features	√	√	√	√	√	√	√	√

In Table 2, Baseline method evaluates the popularity of news tweets based on the *content-based* and the *author-based features* only, while other methods apply different *retweeter-based features* together with the *content-based* and the *author-based features*. Method I concerns that all retweeters have identical influence on the popularity and then uses the *number of retweeters* as a retweeter-based feature. By concerning the influence of active retweeters, Method II and Method III apply *ActivityRate* and *TweetCountScore*, respectively. By focusing on the influence of popular users, Method IV and Method V adapt *PopularityScore* and *FollowerRank*, respectively. Method VI determines the popularity using the two proposed measures, *ActivityRate* and *PopularityScore*, while Method VII applies *TweetCountScore* and *FollowerRank*.

5.2 Experimental Results and Discussion

In this Subsection, the experimental results are reported and discussed. Table 3 summarizes the accuracy of Baseline and the seven methods for predicting the popularity.

Table 3 shows that the accuracy of Baseline method was significantly lower than that of other methods which implies that the popularity prediction built based on the content-based and author-based features yields high error and needs to be improved using other effective features. By applying a retweeter-based feature, the accuracies of all the seven methods were increased. Obviously, the accuracy of Method VI was significantly higher than that of other methods with small values of RSME and MAE (6.304 and 3.833, respectively) suggesting that the *ActivityRate* and *PopularityScore* are considered as important and effective features in order to enhance the performance of the prediction model. Notice that the values of RSME and MAE might be larger than

Table 3. Accuracy of baseline and the seven compared methods.

Method	RMSE	MAE	Adjusted R^2
Baseline method	27.617	23.763	0.453
Method I (number of retweeters)	19.336	13.621	0.551
Method II (*ActivityRate*)	8.103	4.174	0.745
Method III (*TweetCountScore*)	15.801	9.135	0.673
Method IV (*PopularityScore*)	9.657	5.209	0.710
Method V (*FollowerRank*)	13.285	8.478	0.655
Method VI (*ActivityRate* and *PopularityScore*)	6.304	3.833	0.815
Method VII (*TweetCountScore* and *FollowerRank*)	15.021	9.013	0.705

that of the other related works because this experiment predicted the popularity as an exact value, while the other works estimated the popularity in term of a popularity range (e.g., popular or not popular).

By focusing on the influence of active user, the accuracy of Method II was higher than that of Method III indicating that the *ActivityRate* is an effective measure for determining active users. An important reason is that the participation of a retweeter during his/her account lifetime is considered by the *ActivityRate*, while the account lifetime is ignored by the *TweetCountScore*. Thus, an active retweeter measured by the *TweetCountScore* may be very active in a specific period and vary vastly during his/her account lifetime that causes a higher error of the predicting result. By considering the influence of popular user as a retweeter-based feature, the accuracy of Method IV was higher than that of Method V suggesting that the *PopularityScore* is a potential measure for evaluating the influence of popular users. One reason is that the *PopularityScore* concerns only the expertise or the popularity of a user based on the other users' point of view but the *FollowerRank* takes account of not only the number of followers but also the number of followees (or friends). Hence, a retweeter who has large number of followers and knows or follows a lot of friends is not considered as a popular user when evaluated by the *TweetCountScore*.

In addition, the prediction model of Method VI generated based on the multiple linear regression resulted that the retweeter-based features are strongly predictive of the popularity when comparing to the content-based and author-based features. The *ActivityRate* and *PopularityScore* obtained high importance with the significant ($p < 0.000$) and positive coefficient (β) values (1.437 and 1.105, respectively). According to the regression coefficient values, active users are considered as influential users who affect the popularity of news tweets, while popular users have lower influence. The result is intuitive since active users frequently participate in posting and sharing vast amount of social media content, as well as retweeting news tweets. An interesting issue is that how popular users affects the popularity. To investigate this issue, the retweet network of news tweets retweeted by top ten popular users (retweeters who obtain highest values of *PopularityScore*) was analyzed. The result revealed that the popular users mainly retweeted a tweet in the early state (about 10–20 min after a tweet is posted) and before many active users. By using Twitter REST API, the relationship between the popular users and the active users was examined. The result shown that there was a number of active users who follow the popular users retweeted the same tweet after the popular

users, while such active users do not follow the news source. The active users who later retweeted the same tweet may receive the news tweets from the popular users. Thus, the popular users also play important role in disseminating news tweets and help increase the popularity. Accordingly, a tweet retweeted by many active retweeters who have high *ActivityRate* and many popular users who gain high *PopularityScore* is likely to receive high number of retweets and considered as a popular tweet.

Furthermore, most of the author-based features obtained higher coefficient values than that of the content-based features which similar to the results reported in the previous researches (Hong et al. 2011; Ma et al. 2013). Interestingly, the media feature (a content-based feature) had regression coefficient (0.453) close to the author-based features. One possible reason is that the photos or videos attached in a tweet give informative data and attract large number of retweeters.

6 Conclusions and Future Work

This paper presents *retweeter-based features* for predicting the popularity of news tweets using a multiple linear regression model. In order to understand how news tweets are disseminated over the social network, the characteristics and the evolution of popular news tweets are analyzed. The result reveals that the popularity of news tweets is dramatically increased in a few minutes after the tweets are posted and then almost stable. By analyzing the impact of the two types of retweeters, *active users* and *popular users*, the results of the analysis shows strong linear relationships and suggests that a tweet retweeted by very active and most popular users is likely to receive large number of retweets. The *ActivityRate* and the *PopularityScore* are proposed for measuring the influence of retweeters and used as retweeter-based features for modeling the prediction of the popularity of news tweets. The experimental results demonstrate that the application of the *ActivityRate* and the *PopularityScore* (method VI) enhances the performance of the prediction model with high accuracy when comparing to other retweeter-based features as well as the content-based and the author-based features. The result of this research can be applied in online news applications to help promoting news content and enhancing a news recommendation system.

The future work, therefore, includes the development of social media analytic such as breaking news detection application and news recommendation system. For example, a news article retweeted by popular users in the early stage should be recommended to other similar popular users or to active users as many as possible in order to increase the number of views of the news article. In addition, an intensive research on the temporal analysis is required in order to identify the prediction time and investigate more time evolving features for the prediction model.

References

Bao, P., Shen, H.-W., Huang, J., Cheng, X.-Q.: Popularity prediction in microblogging network: a case study on Sina Weibo. In: WWW 2013 (2013)

Bigonha, C.A., Cardoso, T.N., Moro, M.M., Goncalves, M.A., Almeida, V.: Sentiment-based influence detection on Twitter. J. Braz. Comp. Soc. **18**(3), 169–183 (2012)

Cappelletti, R., Sastry, N.: IARank: ranking users on Twitter in near real-time, based on their information amplification potential. In: Social Informatics 2012, Washington, DC, USA, pp. 70–77 (2012)
Fabián, R.: Measuring user influence on Twitter: a survey (2015). arXiv:1508.07951
Gayo-Avello, D.: Nepotistic relationships in Twitter and their impact on rank prestige algorithms. Inf. Process. Manag. **49**(6), 1250–1280 (2013)
Goel, S., Watts, D.J., Goldstein, D.G.: The structure of online diffusion networks. In: EC 2012, New York, USA (2012)
Hajian, B., White, T.: Modelling influence in a social network: metrics and evaluation. In: PASSAT/SocialCom 2011, Boston, MA, USA (2011)
Hong, L., Dan, O., Davison, B.D.: Predicting popular messages in twitter. In: WWW 2011 (2011)
Khrabrov, A., Cybenko, G.: Discovering influence in communication networks using dynamic graph analysis. In: PASSAT 2010, Minneapolis, Minnesota, USA (2010)
Kwak, H., Lee, C., Park, H., Moon, S.: What is Twitter, a social network or a news media? In: WWW 2010, New York, USA (2010)
Lee, C., Kwak, H., Park, H., Moon, S.B.: Finding influentials based on the temporal order of information adoption in Twitter. In: WWW 2010, Raleigh, North Carolina, USA (2010)
Ma, Z., Sun, A., Cong, G.: On predicting the popularity of newly emerging hashtags in Twitter. J. Am. Soc. Inf. Sci. Technol. **64**(7), 641399–641410 (2013)
Messias, J., Schmidt, L., Oliveira, R., Benevenuto, F.: You followed my bot! Transforming robots into influential users in Twitter. First Monday **18**(7) (2013)
Nagmoti, R., Teredesai, A., Cock, M.D.: Ranking approaches for microblog search. In: WI 2010, Toronto, Canada (2010)
Neves, A., Vieira, R., Mourao, F., Rocha, L.: Quantifying complementarity among strategies for influeners' detection on Twitter. In: ICCS 2015 (2015)
Noro, T., Ru, F., Xiao, F., Tokuda, T.: Twitter user rank using keyword search. In: 22nd European-Japanese Conference on Information Modelling, pp. 31–48. IOS Press, Prague (2012)
Petrovic, S., Osborne, M., Lavrenko, V.: RT to Win! Predicting message propagation in Twitter. In: ICWSM 2011 (2011)
Romero, D.M., Galuba, W., Asur, S., Huberman, B.A.: Influence and passivity in social media. In: Gunopulos, D., Hofmann, T., Malerba, D., Vazirgiannis, M. (eds.) ECML PKDD 2011, Part III. LNCS, vol. 6913, pp. 18–33. Springer, Heidelberg (2011)
Srinivasan, M.S., Srinivasa, S., Thulasidasan, S.: Exploring celebrity dynamics on Twitter. In: I-CARE 2013, Hyderabad, India (2013)
Suh, B., Hong, L., Pirolli, P., Chi, E.H.: Want to be retweeted? Large scale analytics on factors impacting retweet in Twitter network. In: SOCIALCOM 2010 (2010)
Szabo, G., Huberman, B.A.: Predicting the popularity of online content. Commun. ACM **53**(8), 80–88 (2010)
Yin, Z., Zhang, Y.: Measuring pair-wise social influence in microblog. In: ASE/IEEE International conference on Social Computing and 2012 ASE/IEEE International Conference on Privacy, Security, Risk and Trust (2012)
Yuan, J., Li, L., Huang, L.L.: Topology-based algorithm for users' influence on specific topics in micro-blog. J. Inf. Comput. Sci. **10**(8), 2247–2259 (2013)
Zaman, T., Fox, E.B., Bradlow, E.T.: A Bayesian Approach For Predicting The Popularity Of Tweets. MIT, Cambridge (2014)