Sentiment Analysis of the Correlation between Regular Tweets and Retweets

Jundong Chen[1], He Li[2], Zeju Wu[2]

[1]Department of Math and Computer Science
Dickinson State University
ND 58601, USA
Email: Jundong.Chen@DickinsonState.edu

[2]College of Communication and Electrical Engineering
Qingdao University of Technology
Qingdao, China
Email: sdlclh@189.cn, wuzeju_qut@163.com

***Abstract*—In this paper, we study the influence from the sentiment of regular tweets on retweeting. We propose a method to calculate the sentiment score for each tweet and each Twitter user. This method enables us to place the tweets and retweets into the same time period to explore the sentiment factor. We adopt the correlation coefficient between the sentiment scores of regular tweets and those of retweets to measure the influence. We categorize the Twitter users in three different ways to investigate other three factors, which are the number of followers, betweenness centrality and the types of accounts. Community detection and machine learning are integrated into our approach. We find that the difference for correlation coefficients exists between different levels of the number of followers, and different types of users. Our method sheds a light on better predicting the dynamics of tweets diffusion.**

***Index Terms*—sentiment analysis, social network, machine learning, Twitter classification**

I. Introduction

Twitter is a popular platform for sharing information. The shared information spreads in a cascading way through followers and followers' followers, etc. Understanding tweets diffusion is very important for anyone who wants to spread their information in an efficient way, such as online advertisers to conduct targeted marketing campaigns, and policy makers to efficiently propagate their policies [1]. One of the most important mechanisms for information diffusion on Twitter is retweeting.

Researchers have been striving to discover the pattern of retweeting, and a step further to predict retweeting. Retweet prediction aims to solve the problem that which tweets are more likely to be retweeted by which Twitter users. Petrovic et al. [2] used a machine learning approach to predict which tweet is retweeted and when is retweeted. Luo et al. [3] developed a prediction model based on time series to better understand the underlying mechanism of retweeting behaviors. Bradlow et al. [4] designed a probabilistic model to forecast the total number of retweets based on a Bayesian approach. However, up to now, the approaches proposed only help us to understand retweeting and barely provide a precise prediction model.

To better predict retweeting, it is necessary to investigate and analyze various factors which impact retweeting. Boyd et al. [5] analyzed why people retweet and what people retweet by elaborating on some examples and case studies about retweets. Suh et al. [6] identified factors that are significantly associated with retweet rate. Those factors are some of the content features and contextual features. They also built a linear model to predict the retweet rate from those features.

Researchers also try to include sentiment as a factor for tweets diffusion. Ferrara et al. conducted a quantitative analysis of the effect of sentiment on information diffusion [7]. They attempted to answer some questions, such as whether positive tweets or negative ones spread faster, what types of emotions of tweets are more popular, etc. To the best of our knowledge, not much work has been done to investigate the sentiment as a factor for retweeting.

We develop an approach to study the impact from sentiment on retweeting. First we use a crawler to collect a dataset which comprises of tweet contents. We calculate the sentiment score for each tweet for each Twitter account. Then we find the correlation for the sentiment scores between original tweets and retweets. We categorize all the Twitter accounts in three different ways. The first way is based on the number of followers. The second way is based on the betweenness centrality. The third one is based on different types of accounts. The correlation coefficients are calculated under the three different ways of categorization.

The structure of this paper is as follows. We first discuss related work and introduce our work in the introduction section. Then, in Section II we explain the process of data collection and why it is impossible to analyze retweeting by first drawing a complete route for a tweet's propagation. Our methodology for the sentiment analysis is provided in Section III. We calculate and compare the correlation coefficients under different ways of categorization in Section IV. We conclude this paper with a discussion in Section V.

II. Data collection

A. Access to Twitter API

First, we obtain the API keys from Twitter's development website. The following code shows how to setup Twitter API. We use Tweepy which is a variant of Twitter API [8]. The real API keys are replaced by 'xxx's.

```
CONSUMER_KEY = 'xxx'
CONSUMER_SECRET = 'xxx'
ACCESS_TOKEN = 'xxx'
ACCESS_TOKEN_SECRET = 'xxx'
auth = tweepy.OAuthHandler(CONSUMER_KEY,
    CONSUMER_SECRET)
auth.set_access_token(ACCESS_TOKEN,
    ACCESS_TOKEN_SECRET)
api = tweepy.API(auth)
```

B. Twitter users' IDs collection

Twitter users' IDs are collected by snow ball sampling technique [9]. Starting from the first user which is called a 'seed' identified by its screen name, we can find the list of its followers. The Python code snippet is shown as below.

```
user = api.get_user(screen_name)
```

Then we find the followers for each of the first Twitter user's followers. This is the 1st depth. We further depths until we collect a certain amount of Twitter users' information. The information for each user is stored in a json file, which includes user id, screen name, followers list, etc. In this process, we adopt breath-first search algorithm.

C. Tweets collection

We collect tweets for each of the Twitter users we gathered. The code snippet is an example for retrieving tweets for a certain user.

```
tweets = api.user_timeline(screen_name =
    screen_name,count=200)
```

To call $user\_timeline$ method of API, we specify the screen name of a user and decide the count of tweets. In fact, 200 is the maximum number of tweets we can obtain for one user per request through Twitter API.

D. Tweets propagation

To analyze or predict retweeting, it would be better if we build a complete route for a tweet which is retweeted by the Twitter users.

Twitter API doesn't provide the information of intermediary retweet, but only the original tweets. So we cannot find a chain to build a path of retweet propagation. Retweet time, and retweeters' IDs are the only useful information for building propagation path, which are contained in the original tweet. Moreover, Twitter API setting a limitation that only up to 100 retweeters' IDs available makes it even more impossible to build propagation path.

E. Impact factors on the path of retweets

Since it is impossible to build a path for a tweet's propagation, we can only find the factors which impact the propagation of tweets. In this paper, we only focus on the sentimental factor. Here, we try to answer the questions: Is it more likely that the people in positive mood retweet a positive tweet or a negative tweet? Do the different types of Twitter users behave differently regarding this question?

III. Sentiment Analysis

Sentiment analysis is an effective tool to discover the sentiment embodied by social medias, such as Twitter, Facebook, etc. There are a number of sentiment analysis methods which can capture content sentiment for short informal text. To calculate the sentiment score for each tweet in our dataset, we use AFINN, along with a list of English stop words so that we don't count frequent common words like "a" or "the". We use AFINN because it is more focused on the language used in microblogging platforms [10].

After collecting the tweets, we first check if the tweet is a pure URL. If it is just an URL, then we assign 0 as its sentiment score because we assume that URLs do not carry any positive or negative sentiment. Then we tokenize the tweets by removing special characters, such as "?", "!","(", ")", "{", "}", etc. We categorize all the tweets for each user account into two categories, regular tweets and retweets by checking if a tweet contains "RT@".

After tokenizing the tweets, each tweet is separated into a series of tokens, which are essentially some words with sentiment scores in an AFINN file. In our work, we use AFINN-111 which is the newest version with 2,477 words and phrases [11].

We calculate the sentiment score for each tweet by summing up the scores for all tokens in each tweet. Suppose we analyze n tweets for a Twitter user, S_{Reg_i} denotes the total sentiment score for all the regular tweets in the ith day and $S_{Reg_i}(t)$ is the sentiment score for one tweet t in S_{Reg_i}. We use S_{Rei} to represent the total sentiment score for all the retweets in the ith day and $S_{Rei}(t)$ is the sentiment score for a tweet t in S_{Rei}. We have the following equations for these variables regarding sentiment scores.

$$S_{Reg_i} = \sum_t S_{Reg_i}(t) \tag{1}$$

$$S_{Rei} = \sum_t S_{Rei}(t) \tag{2}$$

Then, we find the mean of the sentiment scores over all the days for each user by the equations below.

$$\overline{S}_{Reg_i} = \frac{1}{n}\sum_{i=1}^{n} S_{Reg_i} \tag{3}$$

$$\overline{S}_{Rei} = \frac{1}{n}\sum_{i=1}^{n} S_{Rei} \quad (4)$$

In Equation 1 and Equation 2, we calculate the total sentiment scores for regular tweets and retweets which are posted in the same day. This is because people change their moods from day to day and we want to put the regular tweets and retweets into the same time frame.

IV. Correlation between the sentiment scores of regular tweets and those of retweets

In the previous section, we have prepared the sentiment scores for the regular tweets and retweets. In this section, we categorize the users in 3 different ways, which are based on levels of number of followers, levels of betweenness centrality, and types of users. With each categorization, we investigate the correlation between the sentiment scores of regular tweets and those of retweets.

A. Correlation coefficients for different Levels of $followers\_count$

In the dataset we have collected, there are $5,922,804$ tweets for totally $2,403$ Twitter users. We divide all users into 3 groups according to different range of $followers\_count$. The $followers\_count$ is a field of user object in Twitter API, which represents the number of followers of a Twitter account. Each group encompasses 801 users. The low range of $followers\_count$ is from 1 to 916, medium range from 917 to $13,648$, and high range from $13,694$ to $73,347,885$. As shown in Fig. 1, the $followers\_count$ follows a long-tailed distribution, which means most of the users have a small number of followers. That explains why the low range just falls into the first interval of the histogram in Fig. 1, but includes $1/3$ of all the users.

After checking the dataset, we find that only 8.4% of all the users have negative average sentiment scores for regular tweets, and 9.0% for retweets. That means most of the users have positive sentiment scores for their average one-day tweets.

In Table I, we find that the Pearson correlation coefficient for the low level of $followers\_count$ is much lower than the coefficients at medium and high levels. Hence, the users with a low-level $followers\_count$ are less likely to retweet a positive sentiment while posting positive tweets at the same day, compared with the users with medium-level or high-level $followers\_count$. The coefficients for the medium level and high level are at the same order of magnitude (10^{-1}), which are higher than the order of magnitude (10^{-3}) for low-level $followers\_count$. This result implies that the Twitter users with relatively high number of followers tend to retweet positively while tweet positively.

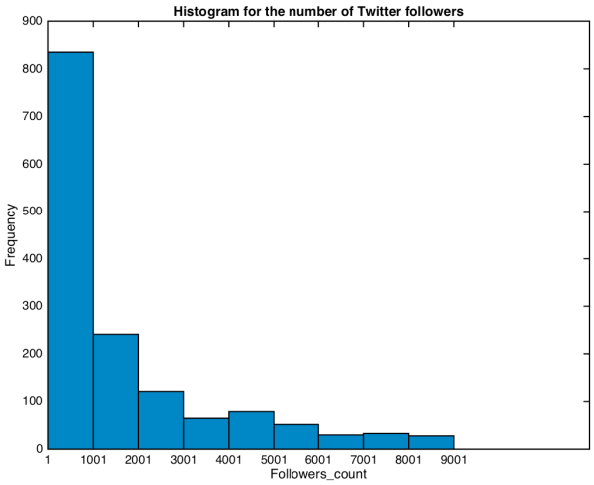

Fig. 1. The number of Twitter followers exhibits long-tailed distribution.

TABLE I. Correlation coefficients of the sentiment scores between regular tweets and retweets for different levels of $followers\_count$

Range of $followers\_count$	**Pearson correlation coefficient between** S_{Reg} **and** S_{Re}
[1, 916]	0.0058
[917, 13648]	0.2417
[13694, 73347885]	0.1341

B. Correlation coefficients for different levels of betweenness centrality

We build up a network graph for the Twitter users in our dataset. The network is a directed graph which consists of users and relationships [12]. Nodes in the network represent the users, and edges correspond to relationships. For example, if there is an edge from node A to node B, then it means user A follows user B.

Network communities allows us to discover the nodes which interact frequently and the relations between each other [13]. By using community detection, Twitter users in the network are divided into different communities as shown in Fig. 2. Different colors represent different communities. In this figure, we only show a representative graph with $1,871$ nodes whose degrees are larger than 80 because it is difficult to display the communities for all the $279,263$ nodes we have collected. We only collect tweets for $2,403$ users in all the $279,263$ users. Other $276,860$ users just appear as the followers for the $2,403$ users. It is hard to display the communities apart from each other because they are intertwined. However, we still find that the community with green color is apart from other communities.

In graph theory, ***betweenness centrality*** measures how

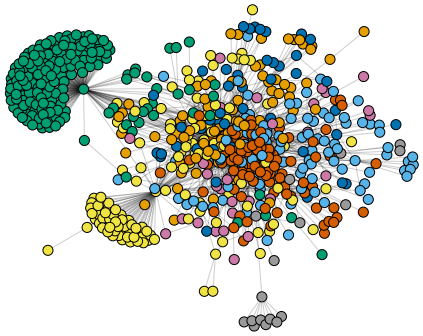

Fig. 2. This figure shows the Twitter network with 1,871 nodes and 14,680 edges, and communities being detected.

often a node appears on shortest paths between nodes in the network. Some nodes serve as the hubs of communities. It is certain that these nodes have high centrality. There are some other nodes which are not hubs of communities, but still they have high centrality because they bridge the different communities.

The betweenness centrality b_i for node i is given by the formula below [14], where $n^i_{s,t}$ is the number of paths from s to t that pass through i and $n_{s,t}$ is the total number of paths from s to t.

$$b_i = \sum_{s,t} \frac{n^i_{s,t}}{n_{s,t}} \tag{5}$$

According to different ranges of betweenness centrality, we divide the Twitter users into three groups. In each group, we calculate the correlation coefficients between the sentiment scores of regular tweets and those of retweets, which are shown in Table II. From this table, we find that the Twitter users with the betweenness centrality in the range $[1.87 \times 10^{-5}, 7.91 \times 10^{-5}]$ have a higher Pearson correlation coefficient than the other two groups. It is easy to understand that the correlation coefficient in range $[1.87 \times 10^{-5}, 7.91 \times 10^{-5}]$ is higher than that in range $[0, 1.86 \times 10^{-5}]$ because $[1.87 \times 10^{-5}, 7.91 \times 10^{-5}]$ is a higher range of betweenness centrality than $[0, 1.86 \times 10^{-5}]$. It seems strange that the Pearson correlation coefficient in the range $[7.92 \times 10^{-5}, 1.19 \times 10^{-2}]$ is the lowest in all the 3 groups. Actually, the fact that the nodes bridge the communities explains this phenomenon that seems strange. Although the nodes that bridge the communities have high betweenness centrality, their degrees are low. Therefore, the nodes which bridge communities but have low degrees lead to a low correlation coefficient in the high range of betweenness centrality.

TABLE II. CORRELATION COEFFICIENTS FOR THE SENTIMENT SCORES BETWEEN REGULAR TWEETS AND RETWEETS FOR DIFFERENT LEVELS OF BETWEENNESS CENTRALITY

Range of betweenness centrality	**Pearson correlation coefficient between S_{Reg} and S_{Re}**
[0, 1.86×10^{-5}]	0.2229
[1.87×10^{-5}, 7.91×10^{-5}]	0.3019
[7.92×10^{-5}, 1.19×10^{-2}]	0.0048

C. Correlation coefficients for different categories of Twitter users

Twitter users can be classified into different categories. Machine learning provides an effective tool to classify each Twitter user into a proper category based on the tweets they post [15]. In this section, we investigate the correlation coefficients in different categories of Twitter users.

First, we build a categories map. In our experiment, we choose 8 typical categories, which include "Tech", "Business & CEOs", "Entertainment", "Science", "Fashion, Travel & Lifestyle", "Sports", "Music", and "Politics". These categories are listed in the first column of Table III.

We use the dataset presented by Deshpande on PyCon France 2016 [16] as the training data. Each user is manually assigned to a category by checking the tweets. There are 200 tweets for each user. This training data is used to classify the Twitter users in our collected data which is the test data.

The training data is used to train our model. The model used in our experiment is a type of topic model, which is LDA (Latent Dirichlet Allocation) [17]. The LDA is one of the most popular topic modeling algorithms. It is a generative model with an assumption that each document is a mixture of a set topics, and each topic is a mixture of words.

To make a better LDA model with only the best topics, we transfer from a HDP (Hierarchical Dirichlet Process) model. HDP model is fully unsupervised. It helps determine the ideal number of topics by using posterior inference. It is shown by Deshpande [16] that the LDA model transferred from HDP model performs better than other models.

After the model is built and trained, we use it to classify the users in our dataset into appropriate categories. The correlation coefficients calculated on our dataset for the 8 different categories are shown in Table III. From the table, we find that "Sports" category has the highest correlation coefficient which means that in this category the users tend to retweet messages with positive sentiment while posting positive tweets. The "Fashion, Travel & Lifestyle" category has a negative correlation coefficient which means that the users in this category tend to retweet messages with lower sentiment scores while posting tweets with higher sentiment scores.

TABLE III. CORRELATION COEFFICIENTS OF THE SENTIMENT SCORES BETWEEN REGULAR TWEETS AND RETWEETS FOR DIFFERENT CATEGORIES OF TWITTER USERS

Categories	**Pearson correlation coefficient between S_{Reg} and S_{Re}**
Tech	0.0028
Business & CEOs	0.0620
Entertainment	0.1454
Science	0.1638
Fashion, Travel & Lifestyle	-0.0454
Sports	0.3464
Music	0.2122
Politics	0.1108

V. CONCLUSIONS

In this paper, we conduct sentiment analysis for $5,922,804$ tweets for $2,403$ Twitter users. We evaluate the sentiment scores for regular tweets and retweets for each user through a method with AFINN. Then, we categorize all the users in 3 different ways, which are based on $followers\_count$, betweenness centrality, and Twitter user classification. Finally, we calculate and compare the correlation coefficients between the sentiment scores of tweets and those of retweets for the categories for each way of categorization.

Through comparing the correlation coefficients at different levels of the $followers\_count$, we find that Twitter users at a relatively high level of $followers\_count$ tend to retweet positively while posting positive tweets. Our comparative analysis at different ranges of betweenness centrality shows that first as the betweenness centrality increases the correlation coefficient increases accordingly. However, the correlation coefficient in the highest range drops significantly, which implies that the degree of a user in the network exerts more influence than betweenness centrality does on the sentiment of retweets given the sentiment of tweets. We also find that the correlation coefficients for sentiment scores differentiate from one type of Twitter user to another. More detailedly, in our dataset the "Sports" category has the highest correlation and "Fashion, Travel & Lifestyle" category even exhibits negative correlation.

As future work, we plan to integrate this approach into a model to better predict the dynamics of retweeting.

ACKNOWLEDGMENT

This work was supported by the Shandong Provincial Natural Science Foundation of China [grant No. ZR2015FQ013] and the National Natural Science Foundation of China [grant No. 61501278].

REFERENCES

[1] M. M. Anwar, J. Li, and C. Liu, *Predicting the Spread of a New Tweet in Twitter*. Cham: Springer International Publishing, 2015, pp. 104–116.

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

[3] Z. Luo, Y. Wang, and X. Wu, *Predicting Retweeting Behavior Based on Autoregressive Moving Average Model*. Berlin, Heidelberg: Springer Berlin Heidelberg, 2012, pp. 777–782.

[4] E. T. Bradlow, E. B. Fox, and T. Zaman, "A bayesian approach for predicting the popularity of tweets," *The Annals of Applied Statistics*, vol. 8, no. 3, pp. 1583–1611, 2014.

[5] D. Boyd, S. Golder, and G. Lotan, "Tweet, tweet, retweet: Conversational aspects of retweeting on twitter," in *Proceedings of the 2010 43rd Hawaii International Conference on System Sciences*, ser. HICSS '10. Washington, DC, USA: IEEE Computer Society, 2010, pp. 1–10.

[6] B. Suh, L. Hong, P. Pirolli, and E. H. Chi, "Want to be retweeted? large scale analytics on factors impacting retweet in twitter network," in *Proceedings of the 2010 IEEE Second International Conference on Social Computing*, ser. SOCIALCOM '10. Washington, DC, USA: IEEE Computer Society, 2010, pp. 177–184.

[7] E. Ferrara and Z. Yang, "Quantifying the effect of sentiment on information diffusion in social media," *PeerJ*, 2015.

[8] R. Zafarani, M. A. Abbasi, and H. Liu, *Social Media Mining: An Introduction*. Cambridge University Press, 2014.

[9] M. Kay, "Generating a network graph of twitter followers using python and networkx," http://mark-kay.net/2014/08/15/network-graph-of-twitter-followers/, August 15, 2014.

[10] A. D'Andrea, F. Ferri, P. Grifoni, and T. Guzzo, "Approaches, Tools and Applications for Sentiment Analysis Implementation," *International Journal of Computer Applications*, vol. 125, no. 3, pp. 26–33, Sep. 2015.

[11] F. Å. Nielsen, "Afinn," Richard Petersens Plads, Building 321, DK-2800 Kgs. Lyngby, mar 2011. [Online]. Available: http://www2.imm.dtu.dk/pubdb/p.php?6010

[12] S. Aparicio, J. Villazón-Terrazas, and G. Álvarez, "A model for scale-free networks: Application to twitter," *Entropy*, vol. 17, pp. 5848–5867, 2015.

[13] J. Yang, J. J. McAuley, and J. Leskovec, "Community detection in networks with node attributes," *IEEE 13th International Conference on Data Mining*, pp. 1151–1156, 2013.

[14] M. Newman, *Networks: An Introduction*. New York, NY, USA: Oxford University Press, Inc., 2010.

[15] M. Pennacchiotti and A.-M. Popescu, "A machine learning approach to twitter user classification," in *ICWSM*, 2011.

[16] D. Deshpande, "Twitter user classification with gensim and scikit-learn," 2016.

[17] D. M. Blei, A. Y. Ng, M. I. Jordan, and J. Lafferty, "Latent dirichlet allocation," *Journal of Machine Learning Research*, vol. 3, p. 2003, 2003.