Predict Individual Retweet Behavior Based on Multi-feature

Chunjia Wang, Yongquan Fan*, Yajun Du and Zefen Sun
School of Computer and Software Engineering, Xihua University, Chengdu, China

*Corresponding author

Abstract. With the rapid development of social media, more and more people get information from social media. The retweet prediction task helps to study the process of information dissemination on social media. Previous methods take into account text features and ignore user’s sentiments and interests. The history tweets tweeted by the user not only reflect the user’s sentiments that the user tends to express on social media, but also shows the user’s interests. In this work, we used a deep learning method to extract features from user profiles, user history, user network, target tweet, then we used these features to predict whether a tweet will be retweeted by a user. Experimental results on dataset show that the proposed model can effectively predict the retweet behavior of users.

Keywords: Deep learning; Retweet behaviour; Prediction model; Social media.

1. Introduction
In recent years, online social media have grown tremendously. The online social media is a tool and platform for people to share life, works, and opinions with each other. There are Weibo, WeChat, QQ in China, and Twitter, Facebook in foreign countries. Compared with traditional social media, but also have high immediacy. The dissemination of information on online social media is spontaneous, and this process relies on the user’s retweet behavior. Previous statistics show that 1% of Twitter users generated 50% of content and controlled 25% of information dissemination [1]. It can be seen that retweeting behavior plays a direct role in the dissemination of information and is a decisive factor in the dissemination of information. The retweet prediction of the user has become a key task in social media analytics and the dissemination of information.

There have been much research works on the retweet prediction task. Zhang et al. [2] measured the retweeting behavior in the process of information dissemination, and studied the relationship between three attributes and the retweeting behavior based on three key attributes: user enthusiasm, user engine, and user duration. Liu et al. [3] proposed a generative graphical model that uses the heterogeneous information networks and node-related texts to mine topic-level influences, and further pointed out that using topic-level influences in predicting the retweeting behavior of user can greatly improve accuracy. Xu et al. [4] divided the influencing factors of users retweet behavior into social relationship-based characteristics, content-based features, and publisher-based features, and trained several classifiers to predict user’s behaviors(decision trees, SVM, logistic regression). The exclusion method compares the effectiveness of various features and shows that social relationship characteristics are more important than others. Zhang et al. [5] used a convolutional neural network model to predict the retweeting behavior and improved the performance of the model by establishing a feature matrix containing similar interests and concerns of users. Hu et al. [6] studied the propagation path of the image from a global perspective to study who will retweet the image and predict the entire propagation path of the image. Yuan et al. [7] considered that the relationship between users is not
static but dynamic, and proposed a model to predict repliers and retweeters given a particular tweet posted at a certain time in a microblog-based social network which considers elaborate user-level and tweet-level features to address these dynamics.

Previous works [8] have shown that the sentimental information is more easily retweeted by those who want to share these sentiments. For example, we know that the sentiments contained in text information are positive, and this tweet may be retweeted by users who have previously shared similar sentiments. In order to accurately analyze, research and predict a user’s retweet behavior, it is necessary to consider the role of user’s sentiments in this process.

In this work, we study the retweeting behavior of individual users under the joint influence of sentimental needs and cognitive needs. The user’s historical tweets not only contain his or her sentimental needs but also reflect his or her cognitive needs. In order to capture this information contained in the user's historical tweets, we propose a model based on deep learning to get this information.

The major contributions of this paper can be summarized as follows:
1. We propose a novel model based on deep learning to predict the retweeting behavior of individual users. The model takes into account the historical tweets of user, the structure of user network and users preferences.
2. We collect a large dataset of social media containing user profile information, historical tweets, and network information.
3. The relevant experiments prove that the proposed method can well predict the retweeting behavior of individual users.

2. Retweet Behavior Prediction
In this section, first, we give a formal definition of the problem of predicting the retweeting behavior of individual users. Next, we introduce how to learn the single tweet embedding and we introduce how to learn user history embedding. Then we introduce joint learning to get the user embedding. Finally, we detail the specific details of the prediction model presented in this work.

2.1. Problem Informal
We define a single-user retweet prediction problem as given a tweet for a user, then we predict whether the user will retweet the tweet. Here, we use tuples \((i, t, m, c)\) to indicate that the user \(i\) is retweeted the tweet \(m\) at the time \(t\) and added their own content \(c\). We use \(y_{i,m}\) indicates the result of the prediction model, where \(y_{i,m} = 0\) means that the user \(i\) do not retweet the tweet \(m\), \(y_{i,m} = 1\) means that the user \(i\) retweet the tweet \(m\). Then, the formal definition of a single user retweet behavior prediction problem is as follows:

\[ y_{i,m} = \text{predict} (i, m) \]  

(1)

where \(\text{predict}(\cdot)\) is the proposed model.

2.2. Method
The deep learning method has been proven to have the ability to automatically learn the best features in industry and academia, and in many tasks, deep learning methods have shown good performance, such as natural language processing, computer vision, image classification, etc. Our model uses deep learning to automatically learn related features, which avoids the problem of performing a large number of artificially designed features and ineffective features such as invalid features.

2.2.1. Users history tweets embedding. We use a list to represent a collection of tweets that the user has posted in the past \(L_i = [m_1, m_2, m_3, \ldots, m_n]\), which \(m_1, m_2, m_3, \ldots, m_n\) respectively represent a tweet posted by the user, \(n\) indicating the total number of tweets posted in the \(i\) user’s history. In this work, we first learn single tweet embedding and then learn the user history embedding.
a) A single tweet embedding: The texts contained in the tweet are sentiments and opinions the user wanted to express. The text can be thought of as a sequence of words. We use the word vector technique Word2Vec [9] to perform a vector representation of each word in the text. Next, we use BiLSTM [10] to progress the sequence of word-vector to obtain tweet embedding.

b) User’s history tweets embedding: Historical tweets published by users can be arranged into a tweet sequence in the order of publication time. In this sequence, front and back tweets may be related. This relationship is uncertain, and we do not know the length of the interval. For example, a user A tweets a tweet about event B at time $T_1$, and then the user A tweets a tweet about event B at time $T_2 (T_2 > T_1)$, but the contents of the two tweets are not the same. Therefore, we use BiLSTM to process the single tweet vector sequence learned in the previous step and to obtain the vector of the user’s history tweet. BiLSTM can learn both short-range dependencies in sequences and long-distance dependencies. Fig. 1 shows the process of the overall vector representation of the historical tweet.

![Figure 1. The framework of learning user’s history embedding.](image)

2.2.2. User embedding: The user profile implies the user’s interests and affects the user’s behavior to some extent. Therefore, according to some previous studies, we take gender, whether authentication, topic distribution, and age. To capture semantic features, we learn user embedding by predicting the real author of the historical tweets [11]. To this end, this goal is trained by the loss function $L_u$ below:

$$ L_u = \sum_{i=1}^{k} E_{\theta_i} - P_{\text{true}}(u_i) \log \sigma(-u_i^T M_i) $$

where $M_i$ represents the user's historical tweets, $u$ and $u_b$ represent the real user and the randomly selected user in the social network, $k$ is the number of tweets.

2.2.3. Network embedding. The user network structure of the online social media reflects the relationship between the user and role of the user in social media. The network structure information implies the influence of surrounding users on the network on the target users. We can capture these impact information by the network structure embedding. We use the network embedding technology to embed the user’s social network. Users’ social network is relatively static and does not change for a period of time. There are many state-of-the-art methods in network embedding. The main focus of our work is not here. We directly choose one of the methods instead of designing a network embedding method.

2.3. Fusion Prediction Layer
In the previous sections, we introduced how to encode user's historical tweets into vectors, how to encode some features of user profile into vectors, how to calculate the similarity between the user and the target tweet, and how to embed the user's network. Next, in this section, we describe how to combine the features extracted in the previous sections to predict the user's retweet behavior.
where \( W \) is the parameter matrix and \( b \) is the offset, \( M_i \) is the representation of the target tweet, \( u_i \) is the representation of the target user, \( N_{ui} \) is the user’s result of network embedding, and \( T_p \) is the target tweet embedding.

We splicing the vectors encoded in the previous sections, and classifying the spliced vectors, which is the prediction result of the retweeting behavior of users.

2.4. Optimization

The essence of the user's retweet behavior prediction model in our work is classification. We use the loss function of the typical classification model as the optimization goal of the training process.

\[
Cost(y_i, \hat{y}_i) = -\hat{y}_i \cdot \log(\hat{y}_i) - (1 - y_i) \log(1 - \hat{y}_i)
\]  

where \( y_i \) and \( \hat{y}_i \) represent the predicted user's retweet behavior and the actual user's retweet behavior, respectively. To optimize goal, we used the Stochastic Gradient Descent (SGD) and the AdaDelta update rule. In the Adadelta update rule, the learning rate does not need to be set.

3. Experiment

3.1. Dataset

In order to evaluate the method in this work, we randomly select ten users as the seed users among the retweeted users of each original tweet, and then crawl ten tweets recently released by these users before retweet the target tweets. In order to more accurately consider the characteristics of the user network, we do not do anything with the user’s attention relationship in the original dataset. The resulting dataset information is shown in Tab. 1. We divide the final dataset into a training set and a test set, and 80% of the data is used for training, and the remaining 20% is used to test the training results.

In this work, we removed the stop words, low-frequency words and special symbols in the text.

| Table 1. Extended Dataset Statistics. |
|--------------------------------------|
| User Relationships Original Retweet |
| 2,978,950 308,489,739 300,000 29,789,500 |

3.2. Baselines

This article will predict user forwarding behavior as a classification problem. This article uses the following methods as a baseline method to evaluate the performance of our proposed method:

- **Random**: The problem of predicting users forwarding tweets is considered a two-category problem. Predict whether the user forwards a tweet and randomly determines whether the user forwards it.
- **Experience**: The user's forwarding rate is taken as the probability of determining whether the user forwards a tweet.
- **SVM**: In characterizing the tweet vector, we use the average word vector to represent the target tweet vector and the user history tweet vector.
- **LSTM**: When embedding is performed on user history tweets, LSTM is used to extract tweet features.
- **SA-LSTM**: This method is the prediction method proposed in this paper. We use BiLSTM to process vector learning of target tweets and vector learning of user history tweets.

3.3. Comparison and Analysis of Experimental Results

Tab. 2 shows the comparison of the proposed method with the baseline method on our data set. We used three evaluation indicators of accuracy, recall, and F1. Among them, the result of Random is the worst. In our opinion, Random does not use any information of users and tweets, but only randomly
determines whether users are retweet. In this way, the randomness is too high to accurately determine the retweet behavior. The ratio of users' retweets is used to predict the probability of users' retweeting behaviors. This method does not take into account some information of users and tweets, resulting in too one-sided results. Support vector machines (SVM) do not work well because they use only textual information and do not take into account the fact that users do not lonely in the network. The proposed method achieves good results, because the proposed method not only considers the text information that affects the retweet behavior of users, but also considers the influence of users from the surrounding users in the network.

4. Conclusion
In this paper, we propose a deep neural network to embed user history, and we use the user history vector to learn user embedding. To gain information about the user network, we use network embedding to embed user node. And parameters are determined flexibly through training. To end, we use user history embedding, user embedding, target tweet embedding and the results of network embedding to predict whether the user will retweet the target tweet. We collected a large number of tweets from Weibo. And we compared the proposed method with the baseline method in this dataset. The experimental results show that the proposed method performs better than these baseline method.

| model     | P(precision) | R(Recall) | F1-Score |
|-----------|--------------|-----------|----------|
| Random    | 0.255        | 0.502     | 0.338    |
| Experience| 0.404        | 0.544     | 0.464    |
| SVM       | 0.367        | 0.632     | 0.464    |
| LSTM      | 0.610        | 0.509     | 0.555    |
| SA-LSTM   | 0.711        | 0.682     | 0.696    |

Acknowledgement
This work is partially supported by the National Natural Science Foundation (Nos.61472329, 61802316, 61872298, 61532009 and 61602389), Sichuan Science and Technology Program (Grant No. 2018GZ0096).

References
[1] X. Tang, Q. Miao, Y. N. Quan, J. Tang, K. Deng. "Predicting individual retweet behavior by user similarity: A multi-task learning approach", Knowledge-Based Systems, 2015.
[2] X. Zhang, D. D. Han, R. Q. Yang and Z. Q. Zhang, Users participation and social influence during information spreading on twitter, PLOS ONE, 12(2017).
[3] L. Liu, J. Tang, J. Han, M. Jiang and S. Yang, Mining topic-level influence in heterogeneous networks, Proceedings of the 19th ACM international conference on Information and knowledge management, ACM, 2010, pp. 199-208, doi:10.1145/1871437.1871467.
[4] Z. H. Xu and Q. Yang, Analyzing user retweet behavior on twitter, Proceedings of the 2012 International Conference on Advances in Social Networks Analysis and Mining (ASONAM 2012), IEEE Computer Society, 2012, pp. 46-50, doi:10.1109/asonam.2012.18.
[5] Q. Zhang, Y. Y. Gong, J. D. Wu, H. R. Huang and X. J. Huang, Retweet prediction with attention-based deep neural network, Proceedings of the 25th ACM International on Conference on Information and Knowledge Management, ACM, 2016, pp. 75-84, doi:10.1145/2983323.2983809.
[6] W. Hu et al, Who will share my image?: Predicting the content diffusion path in online social networks, Proceedings of the eleventh ACM international conference on web search and data mining (WSDM 18), 2018, pp. 252–260, doi: 10.1145/3159652.3159705.
[7] N. J. Yuan et al, “Who will reply to/retweet this tweet?: The dynamics of intimacy from online social interactions,” Proceedings of the Ninth ACM International Conference on Web Search and Data Mining, ACM, 2016, pp. 3-12, doi:10.1145/2835776.2835800.

[8] E. B. Blankenship et al, Sentiment, Contents, and Retweets: A Study of Two Vaccine-Related Twitter Datasets, The Permanente journal, 22(2018), 138.

[9] T. Mikolov, K. Chen, G. Corrado and J. Dean, Efficient estimation of word representations in vector space, arXiv preprint arXiv:1301.3781, 2013.

[10] S. Hochreiter and J. Schmidhuber, Long short-term memory, Neural computation, 9(1997), 1735-1780.

[11] G. Chen, Q. Kong, N. Xu, and W. Mao, NPP: A neural popularity prediction model for social media content. Neurocomputing, 333(2019), 221-230.