Predicting user interaction behavior on government microblogs: a machine learning approach

Yuxuan Li¹, Jie Xiong²*, Zhiwei Tang¹

¹School of Public Affairs and Administration, University of Electronic Science and Technology of China, No. 2006, Xiyuan Avenue, West High-tech Zone, Chengdu, Sichuan, China
²Cyberspace Administration of Guangzhou, No.65, Xiaobei Road, Yuexiu District, Guangzhou, Guangdong, China

*E-mail: 948130449@qq.com

Abstract. Nowadays, the Chinese government usually publishes policies via social media, where everyone can respond to the messages. Exploring the interaction behaviors of millions of users can help the government collect the users’ opinions and make decisions. We use the machine learning algorithm Multilayer Perceptron Network model to predict the user’s interactive behaviors under their comments/replies. We found that three kinds of features are useful in predicting user interaction behaviors: the Doc2vec word vector features, the user attributes, and the user history posting information. The experiments show the effectiveness of the neural network-based model, which provides a way to optimize the formulation and implementation of public policies.

1. Introduction
In the Web 2.0 environment, coping with changes in governance objects and governance environment, the government actively adapts social media to publish information and communicate with the public. In this way, online social media has facilitated the informatization of some government functions and has improved government credibility and governance effectiveness. In China, Sina Weibo is the most popular social media used by the government. Statistically, there are 179.32 million government microblogs opened and certified on Sina Weibo. Unlike other social media, Weibo allows users to quickly create and share volumes of information, opinions, and emotions on a large scale.

Weibo's "instant sharing" feature allows users to comment on posts and participate in discussions while experiencing the event. The user interaction, e.g., comment and reply, in the posts about government affairs increases public participation. Predicting the interaction behavior of government microblog users can optimize the government’s management of online social media, thus benefit the vast number of users.

With the rapid development of artificial intelligence technology, the use of neural networks and other machine learning algorithms to study the behaviors of users on social platforms has received extensive attention and become a research hotspot in related fields such as sociology [1]. Some traditional machine learning algorithms, such as logistic regression, Bayesian network, support vector machines, predict user behaviors have achieved excellent results. However, with the rapid growth of data volume and the increasingly complicated structure of social networks, traditional machine...
learning algorithms encounter problems like performance degradation and weak robustness [2]. In this paper, the classification algorithm based on multilayer perceptron is mainly used to predict the user interaction behavior on government microblogs, and other traditional machine learning algorithms are supplemented for experimental comparison.

In terms of theory, this paper builds a theoretical framework based on the Affective Response Model, Social Network Theory and Social Learning Theory. Affective Response Model is an emotional classification principle in the ICT context composed of five pertinent affective concepts: location, time, special and general stimulation, object stimulation and behavioral stimulation, process and result [3]. Social Network Theory pointed out that people in social situations will think and act in a similar way because of the relationship between them[4]. Users' online participation in online social media is closely related to social networks, and the social relationships a user has will have an important impact on the availability of social support. Social Learning Theory holds that learning results from the interaction of individual cognition, environment and behavior in social environment. Observable behaviors are affected by the interaction of "personal factors, environmental factors and behavioral factors". [5].

Moreover in [6], the author verified the correctness of the hypothesis based on the above-mentioned by using the exponential random graph model (ERGM). They concluded that user interaction in online government microblogs is highly reciprocal, transitive and interest-based homophily rather than gender homophily. Thus, according to the above theory and existing research, we defined basic, microblog information, user information and similarity as characteristic components to predict user interaction behavior on government microblogs.

2. Literature review

We introduce three kinds of machine learning methods commonly used in prediction: logistic regression, Bayesian network, and support vector machine. Logistic regression is a relatively simple two-classification model [7] commonly used in data mining. The category of an unknown category object is determined by its attribute feature sequence. Bayesian Network is a probability graph model, which is also known as the belief network, belief network, and directed acyclic graph model[8]. Its core idea is to provide an effective reasoning step that can simplify system calculation and integrate multi-source information. Support Vector Machines, (SVM) is a two-class supervised learning model based on statistical learning theory [9]. It seeks to minimize structural risks to improve the generalization ability of learning machines to achieve empirical risks and minimize confidence ranges, thus achieving better statistical law effects even with fewer statistical samples. Some studies have shown that the artificial neural network, support vector machine, and decision tree models have reached 81.31%, 90.41%, and 86.61% prediction accuracy respectively, while the accuracy of the traditional logistic regression model is 76.97%[10]. In order to improve the accuracy of the prediction model and consider the interpretability of the model, we use a multilayer perceptrine machine to carry out the prediction research on user interaction behavior of government microblog. We make the experimental comparison with Bayesian network (BY), logical regression, and Support Vector Machine to illustrate our method's superiority.

3. Method

In this section, the multilayer perceptron model and Doc2vec feature training method are used to predict users' interactive behaviors, represented by comments/replies in the online comments of government microblogs. Figure 1 shows the overall flow of the proposed user interaction behavior prediction scheme.
Figure 1. Flowchart of user interaction behavior prediction

In the collected data set, the overall process of predicting the interactive behavior of users who comment on government microblogs online, represented by comments/replies, is as follows. Preprocessing the original data set, including necessary data cleanings and text processing operations such as Chinese word segmentation and part of speech tagging, for subsequent feature extraction processing; In the feature extraction stage, a series of text features required by the prediction task are extracted according to the previous feature analysis. Furthermore, The Doc2vec method is used to extract text word vector features of government microblog posting information and user history posting information. We select "whether the user comments on a microblog" as the binary label. Next, we use the multilayer perceptron to train the prediction task model and adopt the gradient descent learning strategy to obtain the optimal parameters. Finally, we predict the user interaction behavior in the standard test set of the prediction task and use three machine learning methods, Bayesian network, logistic regression, and support vector machine for comparative experiments, including verifying the experimental results in different test set proportions.

3.1 Data set

In this paper, the representative and influential 10 government microblog accounts on Sina Weibo are selected as the main research objects. We crawl the microblog information (e.g., microblog ID, content, time, number of comments, number of forwards), microblog comment information (comment ID, content, time) and comment users (e.g., user ID, gender, region, number of fans, number of concerns, authentication) published by these accounts within a certain period as research samples. In order to ensure the validity of the data, we preprocessed the data: set the filtering conditions for the Weibo content: "Weibo content is original content, non-reposted content, and its number of comments is required to be greater than 1", user information filtering conditions It is the comment user's request that "authentication, gender, historical Weibo, tags and other information exists", and the filter condition of the comment content data is "the comment content requires original, non-retweet content". The processed data is shown in Table 1.

| Account             | Number of published Weibo | Number of comments |
|---------------------|---------------------------|--------------------|
| Safe Beijing        | 159                       | 5967               |
| Anhui Police Online | W172                      | 1055               |
| Chengdu Publish     | 169                       | 2202               |
| Shenzhen Traffic Police | 141                  | 1596               |
| Safe Luoyang        | 139                       | 1138               |
| Nanjing Publish     | 179                       | 4293               |
| Shanghai Publish    | 174                       | 1464               |
| Shandong Court      | 190                       | 1108               |
| Tianjin Traffic Police | 152                 | 1511               |
| Hangzhou Publish    | 187                       | 1582               |
| **Sum**             | **1662**                  | **21916**          |
In user interaction behavior prediction, we aim to predict whether the user comments on the government microblog, i.e., the model should output a binary result. First of all, the corresponding comment and non-comment data sets should be constructed in a general way: traversing the entire data set, crawling the user ID and microblog ID, and using mathematical statistics method to calculate the number of comments of users with a specific ID under the corresponding ID microblog. A total of 21916 pieces of data were collected. Then 1000 users and 100 microblogs were randomly selected to exclude the matched ones, and 21916 samples without comments were treated as negative samples.

3.2 Multilayer perceptron
Natural language processing (NLP) is a theory-motivated range of computational techniques for the automatic analysis and representation of human language [11]. From the perspective of information processing, the multilayer perceptron establishes a simple model by abstracting the human brain's neuron network and forms different networks in different connection methods. The network consists of a large number of connected nodes/neurons. Each node/neuron represents a specific output function, called an excitation function. The network output will be affected by the connection method, weight value, and excitation function of the network. The network itself is usually an approximation of an algorithm or function in nature, or it may be an expression of a logical strategy. The neural network model of a multilayer perceptron contains a three-layer structure. In addition to the most basic input and output layers, there can be multiple hidden layers in the middle. The simplest MLP contains only one hidden layer, that is, a three-layer structure.

The multilayer perceptron is fully connected between layers, i.e., any neuron in the previous layer is connected to all neurons in the next layer. The bottom layer is the input layer; the middle is the hidden layer; the top layer is the output layer. The input layer feeds the model with the extracted N-dimensional eigenvector. The hidden layer and the input layer are fully connected. Assuming that the input layer is represented by a vector \( x \), the output of the hidden layer is \( f(w_1x + b_1) \), \( w_1 \) is the weight, \( b_1 \) is the offset activation function. Sigmoid function, Tanh function, Softsign function, and ReLU function are commonly used as the activation function. The training process is shown in Figure 2.

![Figure 2. Multilayer perceptron training flowchart](image)

3.3 Feature extraction and training
In the feature extraction stage, the word vector features of government microblog posts and user history posts extracted by the doc2vec method are taken as basic characteristic components, and we obtain four types of characteristic components. The microblog information characteristic components, user information and similarity have 13 feature parameters in total. In the subsequent experiments, the
network structure characteristic components are used as the basic components, and other characteristic components are combined to form seven feature combinations to verify the prediction effect. The specific statistics are shown in Table 2.

**Table 2. Summary of different kinds of features**

| Characteristic Components | Characteristics | Interpretation |
|---------------------------|-----------------|----------------|
| Basic                     | Doc2vec-1       | Word Vector Features of Government Microblog Post Information |
|                           | Doc2vec-2       | Word Vector Features of User's Personal Historical Posting Information |
| Microblog information (tweet) | Information emotion score | [0-1]. The emotion dictionary is used to score the emotion of the target microblog information. A higher score indicates positive emotion and a lower score indicates negative emotion. |
|                           | Number of information comments | Logarithm of Government Microblog Comments. (ln) |
|                           | Information comment length | Logarithm of User Comment Length under Government Microblog (ln) |
|                           | Information content richness | [0-1]. The larger the number of "emoticons, @, external links, pictures, videos" in microblog information, the closer it is to "1" |
|                           | Number of Messages Thumbs ups | Logarithm of the Number of Government Microblog Thumbs ups (ln) |
| User information (user)   | Historical emotion | [0-1]. Average emotional scores of microblog messages sent by users' personal homepages |
|                           | Activity        | [0-1]. If the time interval between the two microblog posts on the user's homepage is less than or equal to 24 hours, the active user of the user is considered as "1"; if it is more than 24 hours, the user is considered as inactive user "0" |
|                           | Number of posts posted on Weibo | Logarithm of the number of posts posted on the personal homepage of government microblog users (ln) |
|                           | Certification or not | [0,1], "1" means passed the certification and "0" means failed the certification. |
|                           | Gender          | [0,1]. The gender displayed on the user's personal page, where "1" represents male and "0" represents female. |
|                           | Number of fans  | Logarithm of the Number of fans of users who made comments (ln) |
| Similarity (sim)          | Attribute Similarity between Users and Microblogs | The tf-idf method is adopted to splice the user's personal account information (province, city, gender, label, etc.) as attribute content and carry out numerical processing to obtain attribute feature vectors about the user, thus calculating the user similarity by cosine similarity. |
|                           | Interest Similarity between Users and Microblog Information | The tf-idf method is adopted to digitize the user's historical microblog information content and the information content published by the microblog platform to obtain the attribute feature vectors of the user's post information and the microblog platform tweet information, thus calculating the similarity by cosine similarity. |

In order to speed up the training and improve the correct rate, the standard deviation StandardScaler is selected to normalize so that the data presents a standard normal distribution of the zero mean. After the normalized feature vector is obtained, it is input into a neural network containing four hidden layers, and the number of neurons in the hidden layers is 10, with a learning rate of 0.0001. The Adam method is also used to solve the problem, with the maximum number of iterations set to 100,000.

The loss function is used to check whether the training process terminates. Its mathematical expression is:

\[ L(\hat{y}, y) = -(y \log \hat{y} + (1 - y) \log (1 - \hat{y})) \]  

(1)

Where \( y \) is the true value of the sample and \( \hat{y} \) is the predicted value of the sample. The ideal result is when \( y=1, \hat{y} \) is close to 1, \( L(\hat{y}, y) \) is close to 0 and when \( y=0, \hat{y} \) is close to 0, and \( L(\hat{y}, y) \) is close to 0. Finally, the average value of the loss function of all training data set samples are the cost function of the training set, and the mathematical formulation of the cost function is written as follows:

\[ \text{Average Loss} = \frac{1}{n} \sum_{i=1}^{n} L(\hat{y}_i, y_i) \]
\[ J(w,b) = \frac{1}{m} \sum_{i=1}^{m} L(\hat{y}^{(i)}, y^{(i)}) = -\frac{1}{m} \sum_{i=1}^{m} [y^{(i)} \log \hat{y}^{(i)} + (1 - y^{(i)}) \log (1 - \hat{y}^{(i)})] \quad (2) \]

Our goal is to calculate the best values (matrices) of W and B through iterations and minimize the cost function to zero. After achieving the best training results, we obtain the best parameters of the prediction model.

4. Results and discussion

The accuracy, precision, sensitivity (recall), and F1 score are used as evaluation metrics, and we choose the accuracy as the main evaluation metric. The comparison methods include the Bayesian network, logistic regression, and support vector machine.

4.1 Prediction results

Figure 4.2 is a prediction result of the interactive behavior prediction task using the multi-layer perceptron. The x-axis represents the proportion of the test data set to the total data set, and the y-axis represents the accuracy rate of the prediction result, i.e., the proportion of the test samples that predict the correct emotional tendency of the user to the total test samples. The curves of different colors represent the multilayer perceptron (MLP) prediction results that are trained by adding different characteristic components to the basic characteristic component.

**Figure 3.** Prediction Result Diagram of User Interaction Behavior of Multilayer Perceptron

The classification accuracy of the basic characteristic component is 73.2% under the test set ratio of 10%. With the increase of the proportion of test sets, the classification accuracy of all kinds of feature combinations fluctuates slightly. The prediction accuracy reaches the highest point when there are some kinds of feature combinations in the proportion of 20%, 30%, and 40% of test sets, respectively. However, the peak of the accuracy appears in the full-feature prediction experiment with the proportion of test sets being 10%, and the accuracy of the prediction results generally increases with the increase of the feature combination.

The details are as follows: the test set accounts for 10%, and the classification accuracy rate of all characteristic components is the highest, reaching 90.1%; The classification accuracy rate is 89.2% when the test set accounts for 30% and includes basic characteristic component, user characteristic component and microblog information characteristic component, which shows that the features at user level and microblog information level play a significant role in realizing prediction. On the other hand, the classification accuracy of all feature combinations has no significant decline with the increase of the test set proportion, which can be considered as the model can fit the distribution of the entire data set with fewer data.

4.2 Experimental comparison and result analysis

Figure 4-6 are prediction results using three machine learning algorithms of Bayesian network, logistic regression, and support vector machine. The x-axis represents the proportion of the test data set to the total data set, and the y-axis represents the correct rate of the prediction result, that is, the proportion of the test samples that predicts the correct emotional tendency of the user to the total test samples. The curves of different colors represent the prediction results that are trained with different feature combinations.
Figure 4. The user interaction behavior prediction result on Bayesian network

Figure 5. The user interaction behavior prediction result on logistic regression

Figure 6. The user interaction behavior prediction result on support vector machine

Table 3 summarizes the prediction accuracy of different methods under different test set ratios. Table 4 summarizes the performance comparison of different methods. According to the chart, the accuracy rate of the three machine learning control models is lower than that of the multi-layer perceptron model in predicting the user's interaction behavior. Compared with other comparison methods, the support vector machine has the highest accuracy rate, which reaches a peak of 81.3% when the proportion of test sets is 30%. Moreover, different combinations of characteristic components will have a more significant impact on the accuracy rate. Support vector machine, logistic regression model, and Bayesian network achieve the best prediction results when all features are included. Besides, the prediction results of the three models have not changed significantly under different data proportions, which proves that the models can fit the distribution of the whole data set with some data sets without obvious overfitting.

| Test Set Ratios | MLP   | BY    | LR    | SVM   |
|-----------------|-------|-------|-------|-------|
| 10%             | 90.1% | 68.4% | 73.9% | 80.7% |
| 20%             | 89.1% | 67.6% | 74.3% | 81.1% |
| 30%             | 88.4% | 66.7% | 75.4% | 81.3% |
| 40%             | 86.7% | 67.2% | 74.3% | 80.3% |

Table 4. Summary of performance data for different methods
The MLP method outperforms other traditional machine learning methods by 8.8%-21.7% in terms of accuracy. In terms of recall, the multi-layer perceptron model is 9.6%-40.7% higher than other methods, which shows that the multi-layer perceptron model achieves the best prediction effect in the prediction coverage rate. In terms of the F1 score, the multi-layer perceptron model is 9.8%-33.0% higher than other methods, which sufficiently proves the effectiveness of the multilayer perceptron method in this prediction task.

In addition, we found that three kinds of features are useful in predicting user interaction behaviors the Doc2vec word vector features, the user attributes, and the user history posting information. This finding verifies our research hypothesis based on the Active Response Model, Social Network Theory and Social Learning Theory, which is consistent with the research result of[1]. This means that the user's personal characteristics can be well predicted user interaction behavior on government microblogs. Our study provides important prediction model for public sector managers when they operating an enterprise or government social media platform, they can refer to the research conclusions of this article to predict the user's interactive popularity of the tweet information published by the platform, thereby improving management efficiency.

Furthermore, facing the analysis of social network users' emotions and behaviors with larger data scale and more complicated relationships, using more extracted data features and automatically learned text features and adopting deep neural network models and other methods will achieve better results. In addition, due to the weak interpretability of neural network parameters, more optimized learning strategies need to be further proposed for feature selection, extension and parameter learning of feature matrix in the later stage to improve the effectiveness of prediction methods.

5. Conclusion
In this paper, we use the Multilayer Perceptron Network model to predict user behavior in the comment section of the government microblog. We combine Doc2vec word vector features with microblog content features, user attribute features, and similarity features as the input to learn a multilayer perceptron model. Experiments show that our method outperforms traditional machine learning models like Bayesian Network, Logical Regression, and Support Vector Machine.

References
[1] Li Deng and Dong Yu. 2014. Deep Learning: Methods and Applications. Foundations and Trends® in Signal Processing. 18.7, 1527-1554. DOI=http://dx.doi.org/10.1561/2000000039
[2] Liu B. 2015. Sentiment analysis: Mining opinions, Sentiments, and emotions. Cambridge: Cambridge University Press.
[3] Ping Zhang. The affective response model: a theoretical framework of affective concepts and their relationship in the ICT context. MIS Quarterly. 37.1, 247-274.DOI= https://doi.org/10.25300/misq/2013/37.1.11
[4] Easley D, Kleinberg J. Networks, Crowds, and Markets: Reasoning About a Highly Connected World[M]. Cambridge University Press, 2010 :25-56.
[5] Bandura A. Social cognitive theory[J]. Handbook of Social Psychological Theories. 2011:349-73. DOI= http://dx.doi.org/10.4135/9781446249215.n18
[6] Jie Xiong, Xiaodong Feng, Ziwei Tang. 2020. Understanding User to User Interaction on
Government Microblogs: An Exponential Random Graph Model with the Homophily and Emotional Effect. Information Processing and Management. 57, 2, (Mar. 2020) DOI=https://doi.org/10.1016/j.ipm.2020.102229

[7] Dou, Huili, Wang, Guohua, and Guo, Min. 2012. Algorithm of Traffic State Probability Forecast Based on Logistic Regression. In International Conference on Electronics, Information and Communication Engineering (EICE 2012). Ed. Garry Lee. ASME Press, 99-104. DOI=https://doi.org/10.1115/1.859971.paper99

[8] Ephraim Nissan, Peter Reimann and Hans Spada (eds). 1995 Learning in Humans and Machines: Towards an Interdisciplinary Learning Science. Cambridge: Pergamon Press., 142-44 DOI=https://doi.org/10.1075/pc.5.2.12nis

[9] B. Richhariya, M. Tanveer. 2020. A reduced universum twin support vector machine for class imbalance learning”. Pattern Recognition, 102-119. DOI=https://doi.org/10.1016/j.patcog.2019.107150

[10] Dursun Delen, Leman Tomak, Kazim Topuz and Enes Eryarsoy. 2017. Investigating injury severity risk factors in automobile crashes with predictive analytics and sensitivity analysis methods. Journal of Transport & Health, 118-131, DOI=https://doi.org/10.1016/j.jth.2017.01.009

[11] E. Cambria and B. White. 2014. Jumping NLP Curves: A Review of Natural Language Processing Research. In IEEE Computational Intelligence Magazine, 9, 2(May 2014), 48-57. DOI=https://10.1109/MCI.2014.2307227