Journal of New Media
DOI:10.32604/jnm.2020.010125
Article

Tech Science Press

Analysis and Prediction of New Media Information Dissemination of Police Microblog

Leyao Chen, Lei Hong[*] and Jiayin Liu

Jiangsu Police Institute, Nanjing, 210000, China
[*] Corresponding Author: Lei Hong. Email: honglei@jspi.cn

Received: 20 February 2020; Accepted: 20 February 2020

Abstract: This paper aims to analyze the microblog data published by the official account in a certain province of China, and finds out the rule of Weibo that is easier to be forwarded in the new police media perspective. In this paper, a new topic-based model is proposed. Firstly, the LDA topic clustering algorithm is used to extract the topic categories with forwarding heat from the microblogs with high forwarding numbers, then the Naive Bayesian algorithm is used to topic categories. The sample data is processed to predict the type of microblog forwarding. In order to evaluate this method, a large number of microblog online data is used to analysis. The experimental results show that the proposed method can accurately predict the forwarding of Weibo.

Keywords: Weibo prediction; LDA algorithm; naive Bayesian algorithm; Data mining

1 Introduction

With the rapid development of network technology, new media is no longer a distant and unfamiliar concept, and new media for public security and government affairs are gradually emerging. As the “second battlefield” of public security work, the new media of public security government has closely integrated new media technology with police work, and its importance is self-evident. In recent years, local public security organs have paid growing attention to the construction of new media. They have opened official Weibo and WeChat to adapt to the new development of the Internet era and the new expectations of people. This makes the masses more aware of the work of public security, and the development of the new media of the police also enables the public opinion database to be fully developed. The police can use the social network behaviors of the masses (browse, like, forward, comment, etc.) and transmission range of the message to judge the social problems that the masses care about.

2 The Status of Research

In recent years, a lot of work has been studied from different perspectives of Weibo communication, including account social influence, publishing the viewpoint features in Weibo content and the social characteristics of users’ groups. However, the method based on BP (back propagation) neural network and the method of predicting the microblog forwarding amount under the emergency event and using the SIM-LSTM model to predict the microblog forwarding is too complicated and redundant [1,3,4,6], and does not have certain intuitiveness. The disadvantages of oter methods are also same with above methods. Through the analysis of forwarding microblogs, we find that the highly forwarded microblogs have certain thematic features. According to our proposed LDA-based naive Bayes algorithm, it is more intuitive and efficient to predict whether Weibo will be forwarded. At present, there is no research on the forecast of the public security police microblog in China. The focus of the current research on new media policing is still on how to deal with public opinion, so our research has the following advantages:

Fill in the current research gap, and create a set of methods that are exclusively targeted at public security microblogs, and provide positive ideas for public security work to guide public opinion and promote communication with net friends.

This is a more concise and intuitive algorithm. Compared with the impact of the social network that has too emphasized Weibo on forwarding, the algorithm is more direct and easier to implement.

3 The Methods of Research

According to CNNIC's "Statistical Report on the Development of China's Internet Network", as of December 2018, the number of netizens nationwide reached 829 million. In the 2018 China Weibo user scale and usage, the data shows that the number of Weibo users in China was 337 million in 2018, an increase of 34.56 million compared with the end of 2017. The proportion of Weibo users in the total number of Internet users reached 42.3%. The increase in the number of user groups also means that there will be more and more problems and conditions in the process of Weibo communication [2,5,7,8]. And the main method of exploration is too complicated, and it is not easy to apply directly to the official microblog number of the public security. So we propose a research method based on the LDA topic clustering model.

Firstly, we will introduce the model and research methods:

3.1 Latent Dirichlet Allocation

LDA is a probabilistic topic model: Latent Dirichlet Allocation, which can give the topic of each documents set as a probability distribution, extract some potential topics by analyzing some documents and also bring about text clustering or categorization based on topics. At the same time, it is also a typical word bag model, that is a document composed of a group of words, and there is no order between words and words.

The traditional way to judge the similarity of two documents is by looking at the number of words that appear together in two documents [9], such as TF-IDF, etc. This method does not take into account the semantic association behind the text and the polysemy of the word, possibly in two documents that are few or no words in common for each document, but two documents are similar. LDA can handle this situation well by extracting the potential topics of the document and quantifying the probability of the relevant subject words.

The subject refers to a list of words that are semantically related to the topic and their weights, that is, the vector of the conditional probability of each word under the topic. The closer the relationship is to the topic, the greater the conditional probability, and vice versa. In the topic model, a topic represents a concept, an aspect, expressed as a series of related words, and is the conditional probability of these words. More vividly speaking, the theme is a bucket with words with high probability of occurrence. These words have a strong correlation with this theme.

The model generation process is as follows:

Select a document d_i according to the prior probability $p^{(di)}$

Extract the subject distribution θ_i of the sample generation document d_i from the α of the Dirichlet distribution,

Extract the subject $z_{(i,j)}$ of the j_{th} word from the sample polynomial distribution θ_i.

Extract the sample from the Dirichlet distribution β to generate the word distribution $\phi_{z(i,j)}$ corresponding to the topic $z_{(i,j)}$. The word distribution $\phi_{z(i,j)}$ is generated by the Dirichlet distribution with the parameter β.

Collect samples from the polynomial distribution $\phi_{z(i,j)}$ of the word and finally generate the word $\omega_{(i,j)}$

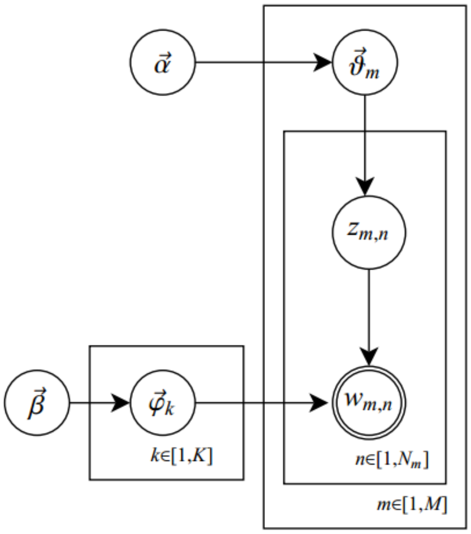

Figure 1: LDA model generation flow chart

3.2 Naive Bayesian Algorithm

Bayesian classification is a general term for a class of classification algorithms. These algorithms are based on Bayes' theorem, so they are collectively called Bayesian classification. The Naive Bayes Classifier is a simple and easy to understand classification method that looks Naive but works very well. The principle is Bayes' theorem, which obtains new information from the data and then updates the prior probability to obtain the posterior probability. It is like saying that we judge the quality of a person. For a stranger, the possibility of his quality is good or bad that is 50%. If he says that he has done a good thing, then this new information has increased the probability that we will judge him to be a good person. The advantage of Naive Bayes classification is that it is not afraid of "data impurities" and irrelevant variables. Its Naive is that it assumes that each feature attribute is irrelevant and independent. For the given item to be classified, the probability of occurrence of each category under the condition of occurrence of this item, and which is the largest, is considered to belong to which category.

Firstly give the Bayesian formula:

$$P(B_{\mathrm{i}} \mid A)=\frac{P(A \mid B_i)P(B_i)}{\sum_{j=1}^{n} P(A \mid B_j)P(B_j)}$$

The naive Bates calculation formula (the expression under multiple features) is:

$$P(C=\mathrm{c} \mid A_1=a_1 \ldots A_n=a_n)=\alpha P(C=c)\prod_{i=1}^{n} P(A_i \mid C=c)$$

We want to predict whether Weibo will be forwarded, actually something what we do and will do is to ask for the posterior probability of this matter. According to the Bayesian inference, the adjustment factor is based on the probability that the event has occurred, and the probability that the event may occur, and the ratio of the probability of detection. The posterior probability is obtained by adjusting the prior probability by this ratio. It also achieves the prediction we want.

Through the screening of the acquired sample data, the sample with high forwarding number is found as the input data of the LDA model. The model aims to output the topic features of the highly forwarded text, and as the input of the naive Bayesian algorithm, the final prediction result is output. The research method flow is shown in the following figure:

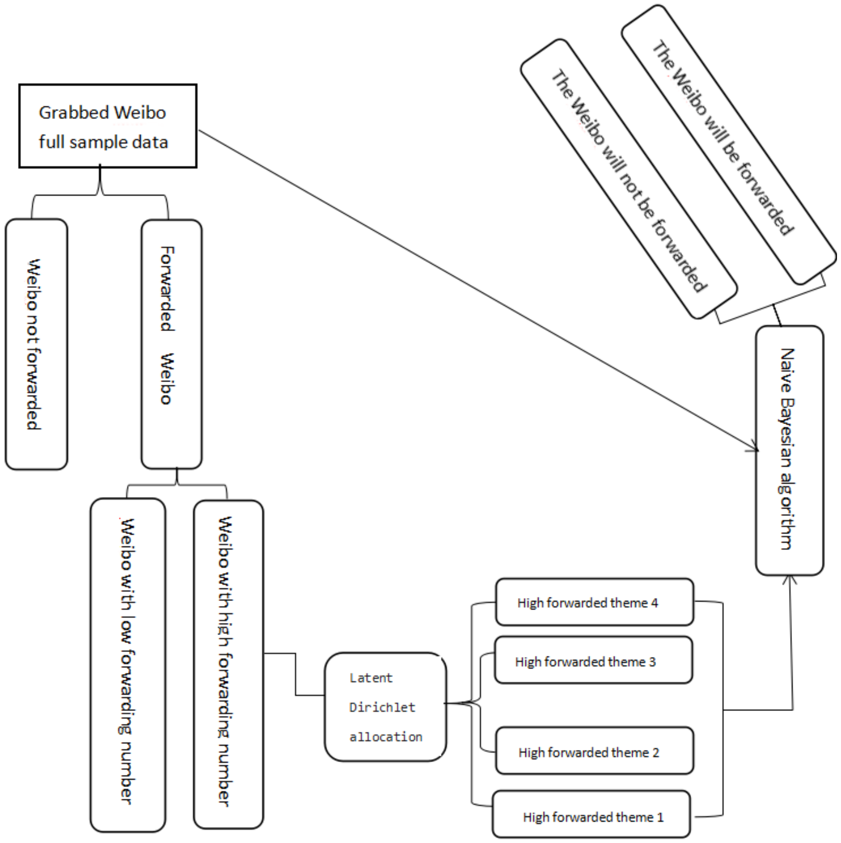
Grabbed Weibo full sample data
Weibo not forwarded
Forwarded Weibo
Weibo with low forwarding number
Weibo with high forwarding number
Latent Dirichlet allocation
High forwarded theme 4
High forwarded theme 3
High forwarded theme 2
High forwarded theme 1
Naive Bayesian algorithm
The Weibo will not be forwarded
The Weibo will be forwarded

Figure 2: Research method flow chart

Performing the first cleaning in the captured microblog full sample data, and perform the first data cleaning according to whether the repost_num forwarding value is zero. Then, according to our proposed forwarding number not less than 1.5 times the average number, the second data cleaning is performed. The data of the 22w+ strips of the microblog full sample is filtered into the high forwarding number data of 10w+ strips, and is left for further processing.

Inputting the high forwarding number data obtained in the first step into the LDA topic model, train the theme distribution of the text data through machine learning, set the number of displayed words for each topic, and leave the obtained subject as the simplest in the next link. Bayesian theme classification features.

The topics obtained in the second step are listed as the characteristics of the naive Bayesian algorithm, and the captured microblog full sample data is divided into a training set and a test set, and this step will be from the test set. It is concluded which microblogs will be forwarded and which will not, leaving for the fourth verification.

Writing a verification script to check whether the repost_num forwarding value of the microblog that is predicted to be forwarded is zero, and calculate the recall rate and the precision rate. Finally, the prediction rate of the research algorithm is tested according to the two criteria.

4 The Experimental Process

4.1 Data Acquisition and Data Mining

4.1.1 Data Acquisition and Cleaning

First of all, we developed a crawler program based on the scrapy framework to capture the data needed for the experiment. Each official Weibo account has a uniquely specified uid, so we can automatically fill it every time according to this rule. Take the url and implement the automatic jump page according to the url rule. We chose to use cookies to keep each of the next sessions, and to bypass the anti-crawl mechanism by controlling the interval at which data was fetched. The crawler program uses the principle of depth first to climb down the historical records on each official microblog number according to the chronological order, and input the microblog uid of thirteen fixed accounts into the queue program, and the queue loops and then crawls. A total of 266,266 pieces of Weibo information, 180,203 pieces of Weibo comment data and related account fans and other account information were obtained. For the storage of data, we use the non-relational mongodb database. Its light and rich functions provide great help for the next text analysis work. The crawled data is saved in the database in json format.

The first step of cleaning: use the json template that comes with python to process the original data, convert the json format into a dictionary and then operate it. We save the microblog data with the repost_num value greater than 1 to another file.

The second step of cleaning: through observation, we will study and analyze the content (microblog content) and repost_num (forwarding number) in the acquired data. Here we first calculate the average number of forwarding numbers in all the obtained microblog data. After calculating the full sample average, it is empirically possible to try to delineate that the high forwarding threshold is greater than the average and not less than 1.5 times the average. On this basis, the first step of data processing is implemented, and a high-forward microblog data sample is obtained. After calculation, the average sample size is 24, and the threshold is 36. Then, the statement is used to save the microblog data of repost_num not less than 36 to another file.

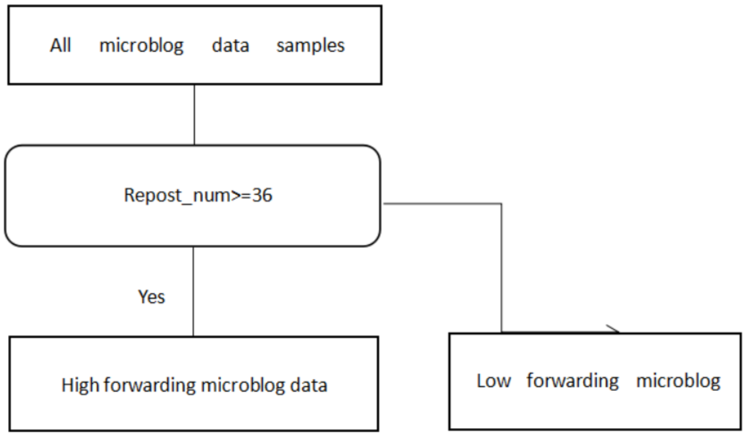

Figure 3: Data cleaning diagram

4.2 Data Mining

In the LDA topic model processing stage, the cleaned microblog text data is first processed with jieba Chinese word segmentation to obtain the input corpus of the LDA model. At this point, the corpus should be cleaned again, and the special symbols (.[-9], -., !~*) and the emojis that come with WeChat (such as [doge], [flowers], [microphones]) are removed). After reading the corpus data, the LDA model converts it into a dictionary, and calculates the text vector after counting the number of words. Then the document TF-IDF is calculated, and the LDA model is established and the training phase is entered. After several trainings, the selected stop words are selected and included in the stop word library, and the model is imported to continue training. After the training, you can specify the output theme and the keywords of each topic as needed, and because the microblog text belongs to short text, the topic classification is relatively difficult, it takes time to debug the theme, the number of keywords, and the stop thesaurus. Keep up-to-date updates to find a balance that yields accurate and common-sense subject classifications. The output of the LDA model is as follows:

In fact, we obtained the probability distribution of five topics in the experiment. Here is an example to illustrate the first topic probability distribution:

Top 5 topics in the 968th document:

Table 1: Results of the experiment

Word	Word vector probability
Word 0	0.51061165
Word 1	0.01250591
Word 2	0.01250591
Word 3	0.01250591
Word 4	0.45187062
Topic	Probability distributions
Topic 1	0.51061165
Topic 2	0.01250591
Topic 3	0.01250591
Topic 4	0.45187062
Topic 5	0.01250591

From the overall topic distribution, the word distribution of all topics generated by the sample data is as follows:

Table 2: Topics generated by the sample data

Topic	Word 1	Word 2	Word 3	Word 4	Word 5
Topic 1	high speed	prevention	means	police	internet
Topic 2	remind	prompt	fraud	SMS	mobile phone
Topic 3	notice	citywide	high temperature	improve	driver
Topic 4	college entrance examination	candidate	cheater	remind	receive
Topic 5	peace	police	kids	110	notice

After obtaining the high forwarding theme through the LDA topic model, enter the category feature of the classification in the naive Bayesian algorithm of the next step. Here, the full sample data of the microblogs captured is divided into a training set and a test set according to a ratio of 6:4. We add an index "subject category +I" to each microblog in the algorithm (where i is a number, that is, an annotation A Weibo is the first item of this topic, and then according to the algorithm to get the subject category to which each Weibo belongs. First, use the training set to carry out the training phase of machine learning, and then use the test set to conduct related tests to calculate whether the categories of these microblogs belong to the high forwarding category. If they belong, they will be forwarded, otherwise they will not be forwarded and all algorithms will be processed. After that, the output data is retained and left for verification.

The output data of Naive Bayes algorithm is as follows:

Table 3: The output data of Naive Bayes algorithm

	Total	The number of Weibo judged not to be forwarded	The number of Weibo judged to be forwarded
Forward topic training set results	10000	90	9910
Forward topic test set results	4980	110	4870
Other topic training set results	10000	9960	40
Other topic test set results	4980	4930	50

The microblogs predicted to be forwarded in the test set are extracted, compared with the forwarded microblogs obtained by the first round of data cleaning, and finally judged by the predicted and verified recall and precision.

5 Analysis of Verification Results

For the results obtained, we will verify that the results obtained by the above experimental model are correctly predicted by the precision and the recall rate. The precision rate refers to the proportion of microblogs that are correctly predicted in a class as the proportion of such microblogs. For example, the microblogs being forwarded have an accuracy rate of a/(a+c), and the recall rate is one type of the correctly predicted Weibo accounts for the proportion of all Weibo in this category. The following results show that the model has a recall rate of 0.61 and a precision of 0.64.

It can be seen from the above table that the model has a high prediction rate, and it can be seen from the predicted results that the microblogging of the life warnings issued by the public security official microblog (pufa push, fraud prevention, security common sense push, etc.) is easier. It is forwarded and shared by users, which is also a true reflection of the real life of the masses. The model can accurately predict the microblogs with wide spread and easy to cause sensation, which has a positive effect on guiding the public security departments to better build the police microblog number.

6 Summary

This paper firstly collects a large amount of online microblog data from Weibo, simulating the actual use environment of netizens. Then through the data analysis and the simplicity and efficiency of the algorithm, it is found that the microblog text topic category is an important factor affecting its microblog forwarding rate. Based on this, we propose an experimental method to predict the forwarding behavior of Weibo. The method is based on the LDA model and is modeled using the Naïve Bayes algorithm for prediction. Experiments show that there are two popular forwarding themes in public security police microblog: social hotspot case notification and life safety. From the final recall and precision of the model, this experimental method has certain accurate prediction ability. Through the predictions of the model, the life warning class (preventing fraud, etc.) is the most popular type of microblog tweets that can be forwarded by users. It can be seen from the displayed topic category keywords that the user forwards relevant content before and after the college entrance examination.

It is the era of rapid development of the Internet, but this development is a double-edged sword. Once the technology is mastered by the deaf, it is a tool for committing crime. Today, information crimes are intensifying. The development of telecom fraud is very rapid. It is so big that the “Xu Yuyu” incident is a sensation in the whole country. The public opinion is unprecedented, and the young people in the countryside are defrauded of how much savings they have. The public’s awareness of prevention cannot be cultivated overnight. Public security organs should make good use of the favorable conditions for the development of new media policing, actively guide the masses to pay attention to their own property security issues, and inform the public about the relevant cases. I also hope that the new media of the public security microblog can make a good voice of public security, hold the “second battlefield” of public security work under the new era and new situation, make full use of data empowerment, accurately issue every microblog, and be considerate of everything the masses care about.

Acknowledgement: Thanks to the teachers from the network security department of Jiangsu Police Institute for their guidance in the process of completing this article. This study is supported by Jiangsu Province University Students Practice Innovation and Entrepreneurship Training Program Project, Project Number: 201910329031Y, Project Name: Research on the influence of new media platform of Public Security Colleges under the background of big data. “Research on the reform and innovation of network public opinion teaching in public security colleges and universities from the perspective of overall national security” (Project No. C-B/2020/ 01/27). Jiangsu Province modern education technology research project “Research on the innovation of public security network public opinion teaching mode based on modern information technology” (Project No. 2017-R-59195). The key teaching reform project of Jiangsu Police Institute “Research on the reconstruction of online and offline hybrid” golden course “teaching system of Internet information inspection course” (Project No. 2019A30).

Funding Statement: This study is supported by Jiangsu Province University Students Practice Innovation and Entrepreneurship Training Program Project, Project Number: 201910329031Y, Project Name: Research on the influence of new media platform of Public Security Colleges under the background of big data. “Research on the reform and innovation of network public opinion teaching in public security colleges and universities from the perspective of overall national security” (Project No. C-B/2020/01/27). Jiangsu Province modern education technology research project “Research on the innovation of public security network public opinion teaching mode based on modern information technology” (Project No. 2017-R-59195). The key teaching reform project of Jiangsu Police Institute “Research on the reconstruction of online and offline hybrid” golden course” teaching system of Internet information inspection course (Project No. 2019A30).

Conflicts of Interest: The authors declare that they have no conflicts of interest to report regarding the present study.

References

[1] J. Zhang, J. Yin and C. T. Hu, “Prediction of Weibo forwarding based on active fan forwarding influence model,” *Journal of Jiangsu Police Officer Academy*, vol. 34, no. 1, pp. 116–121, 2019.

[2] G. Z. Luo. “Research on perspective evolution model and information forwarding prediction in social networks “, Beijing Jiaotong University, 2018.

[3] S. Q. Lai. “Research on factors affecting user microblog information forwarding”, *Library Work and Study*, no. 8, pp. 31-37, 2015.

[4] Q. Deng, Y. F. Ma, Y. Liu and H. Zhang,“Prediction of microblog forwarding based on BP neural network”, *Journal of Tsinghua University(Science and Technology)*, vol. 55, no. 12, pp. 1342-1347, 2015.

[5] S. K. Mu, L. Q. Zhang and C. F. Teng, “Prediction of Weibo forwarding behavior based on recurrent neural network”, *Computer Systems*, vol. 28, no. 8, pp. 155-161, 2019.

[6] P. Guan, Y. F. Wang and Z. Fu, “Analysis of the extraction of scientific literature subjects based on LDA topic model under different corpus”, *Library and Information Service*, vol. 60, no. 2, pp. 112-121, 2016.

[7] H. D. Zhao, G. Liu, C. Shi and B. Wu, “Prediction of Weibo forwarding based on forward propagation process”, *Chinese Journal of Electronics*, vol. 44, no. 12, pp. 2989-2996, 2016.

[8] Y. Li, Y. H. Chen and T. Liu, “A review of microblog information propagation prediction research”, *Journal of Software*, vol. 27, no. 2, pp. 247-263, 2016.

[9] D. M. Bleo, A. Y. Ng and M. I. Jordan, “Latent dirichlet allocation”, *Journal of Machine Learning Research*, no. 3, pp. 993-1022, 2003.