Analysis of Information Dissemination Orientation of New Media Police Microblog Platform

Leyao Chen, Lei Hong(✉), and Jiayin Liu

Jiangsu Police Institute, Nanjing 210000, China
honglei@jspi.edu.cn

Abstract. This article aims to analyze the microblog data released by a new social media platform for police in a certain province in China, and find out the characteristics of those microblogs which are more likely to be reposted from the perspective of modern police affairs. A new topic-based model is proposed in this article. In the first step, the LDA topic clustering algorithm is adopted to extract the topic categories with forwarding heat from the most-reposted microblogs, with a Naive Bayesian algorithm as followed to predict the topic categories. The pre-processed sample data is applied on the proposed method to predict the type of microblog reposting. The experimental results utilizing a large number of online microblog data demonstrate that the proposed method can accurately predict the reposting of microblogs with high precision and recall rate.

Keywords: Microblog 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. In particular, new media platform for public security and government affairs are gradually sprung up in recent years. As the "second battlefield" of public security, the new media for public security government plays a pivotal role in integrating new media technology with police work. In recent years, more and more attention has been attracted by local public security organs on the construction of new media. They have opened official microblog and WeChat public account to adapt to the new Internet era and the refreshing expectations of people. This makes the masses more and more aware of the work of public security. On the other hand, the development of the new media for the police also enables the public opinion database to be fully developed. Accordingly, the police can also make use of the social network behaviors of the masses (browse, like, repost, comment, etc.) and transmission range of the message to identify the social affairs and problems that the masses concern.

X. Sun et al. (Eds.): ICAIS 2021, LNCS 12736, pp. 666–675, 2021.
https://doi.org/10.1007/978-3-030-78609-0_56

2 The Status of Research

In recent years, there are numerous studies in the literature regarding different perspectives of microblog communication, including social influence of the accounts, the viewpoint features in microblog 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 reposting amount under the emergency event and using the SIM-LSTM model to predict the microblog reposting is too complicated and redundant [1, 3, 4, 6], and is not intuitive and interpretable enough. The disadvantages of other methods are also the same with above methods. Through the analysis of forwarding microblogs, we conclude that the highly reposted microblogs have certain thematic features. According to our proposed LDA-based naive Bayes algorithm, it is more intuitive and efficient to predict whether microblog will be reposted. At present, there is no research on the reposting forecasting of the microblog for public security police in China and current research on new media policing still focus on how to deal with public opinion. Hence, our research has the following advantages:

Fill in the current research gap, create a collection of methods that are exclusively targeted at public security microblogs, provide positive thoughts for public security work to guide public opinion, and enhance communication between groups of netizens.

This is a more accurate and intuitive algorithm. Comparing with the performance of the social network that has too much emphasis on reposting microblogs, 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 report of 2018 China Microblog User Scale and Usage, the data shows that the number of microblog users in China was 337 million in 2018, increasing by 34.56 million compared with the end of 2017. The proportion of microblog users among the total number of Internet users reached 42.3%. The increase in the number of user groups also implies that more and more problems and emergencies tend to appear during the communication process via microblog [2, 5, 7, 8]. However, the previous method is too complicated, and it is not easy to apply directly to the official microblog account of the public security. Therefore, we propose a method based on the LDA topic clustering model.

Firstly, we will introduce the model and research methods as followed:

3.1 Latent Dirichlet Allocation

Latent Dirichlet Allocation (LDA) is a probabilistic topic model, which assigns a probability distribution to the topic of each documents in the document set, and accordingly clusters or classifies topics by analyzing documents and extracting their potential topics. At the same time, it is also a typical word bag model, which views the document as composition of a group of words with no order between them.

The traditional way to measure the similarity between two documents is to calculate the number of words that appear in two documents simultaneously [9], such as TF-IDF. However, the semantic association behind the text and the polysemy of the word are not taken into account in this method. It is highly probable that there are few or no words in common between two documents, but they are very similar with closely related topic. LDA can handle this situation well by extracting the potential topics of the document and quantifying the existing 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 between the subject and words, the greater the conditional probability, and vice versa. In the topic model, a topic represents a concept, or an aspect, representative by a series of related words, and is the conditional probability of these words. More specifically, the theme is a bucket of words with high probability of occurrence. These words have a strong correlation with this theme.

The model is built as follows:

Select a document d_i according to the prior probability $p^{(di)}$.
Extract the topic distribution θ_i of the generated sample document d_i from the α of the Dirichlet distribution,
Extract the subject $z_{(i, j)}$ of the j_{th} word from the polynomial distribution θ_i of the sample.
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 $\omega(i, j)$ (Fig. 1).

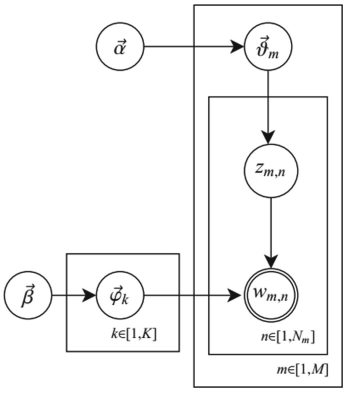

Fig. 1. LDA model generation flow chart

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 seems 'naïve' 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. Just like evaluating the quality of a good when you are shopping online. If you do a random guess, there will be 50% chance for this product to be an acceptable one. However, if you have the prior knowledge that the shop has a good reputation, then this information will increase the probability that this good will be an acceptable one. The advantage of Naive Bayes classification is that it will not be affected by "data impurities" or irrelevant variables. 'Naïve' means it assumes that features are independent with each other. A given item will be classified into the category that has the largest predictive probability.

The Bayesian formula is given below:

$$P(B\mathrm{i}|A) = \frac{P(A|Bi)P(Bi)}{\sum_{j=1}^{n} P(A|Bj)P(Bj)} \tag{1}$$

The Naive Bayes calculation formula (the expression under multiple features) is:

$$P(C = \mathrm{c}|A1 = a1...An = an) = \alpha P(C = c)\prod_{i=1}^{n} P(Ai|C = c) \tag{2}$$

In order to predict whether microblog will be reposted or not, we need to calculate the posterior probability of reposting. According to Bayesian inference, the adjustment factor is based on the probability that the event has occurred, the event may occur, and the ratio of the probability of detection. Adjusting the prior probability by this ratio can help us obtain the posterior probability, so as to make accurate prediction.

Through the screening of the acquired sample data, the frequently reposted sample microblogs are identified as the input data of the LDA model. The model aims to output the topic features of the frequently reposted text, and as the input of the naive Bayesian algorithm, the final prediction result is the output. The flow of our proposed research method is shown in the following Fig. 2:

First, we clean all the captured microblog sample data by removing those with repostnum equals to zero. Then, we perform the second-round data cleaning according to our proposed rule that reposting number is not less than 1.5 times of the average number. 22w+ strips of the full microblog sample are filtered into 10w+ strips data with high reposting number left to be further processed.

Next, input those frequently reposted data obtained in the first step into the LDA topic model, train the theme distribution of the text data through machine learning algorithm, set the number of displayed words for each topic, and leave the obtained subject as the simplest in the next link the subject classification feature of Bayesian.

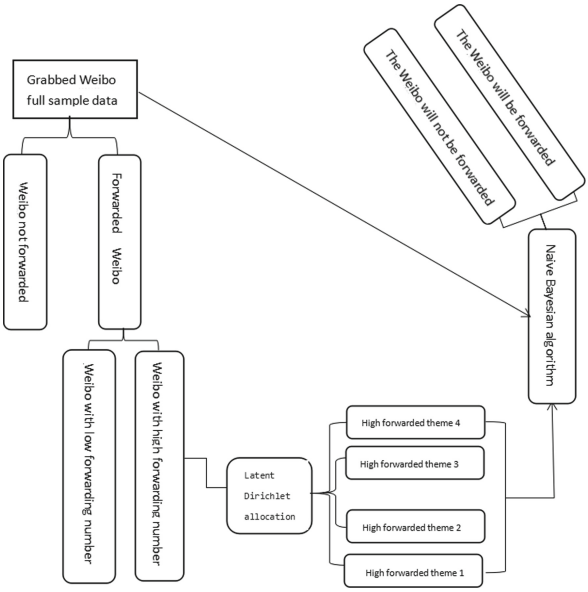

Fig. 2. Research method flow chart

The topics obtained in the second step are listed as the features for naive Bayesian algorithm, on which the training data are applied. In this step, we use the test data on the trained model to predict which microblogs will be reposted and which will not, waiting for the last step of verification.

In our last step, we write a verification script to check whether the predicted repost_num of the microblog is true by calculating the recall rate and the precision rate. Finally, 2 criteria are adopted to measure the prediction accuracy of our proposed algorithm.

4 The Experimental Process

4.1 Data Acquisition and Data Mining

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 microblog account has a uniquely specified uid, so we can automatically complete the URL that will be crawled each time according to this rule. Use the URL query string to realize the automatic jumping from one page to another 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 adopts the principle of depth to climb down the historical records on each official microblog account in chronological order, and then inputs the microblog uid of thirteen fixed accounts into the queue program with loops to crawl the data. A total of 266,266 pieces of microblog information, 180,203 pieces of microblog comment data

and related information such as the number of fans were obtained. For the storage of data, we use the non-relational mongodb database. Its light but abundant functions provide great help for the next step of text analysis. The crawled data is saved in the database in json format.

The first step of cleaning: use the json module in python to process the original data, converting the json format into a dictionary to operate further. 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 (reposting number) in the acquired data. Here we first calculate the average number of reposting numbers in all the obtained microblog data. After calculating the full sample average, we empirically eliminate those samples with reposting number greater than the average but not less than the threshold, which is 1.5 times of the average number. On this basis, the first step of data processing is completed and a frequently-reposted 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 (Fig. 3).

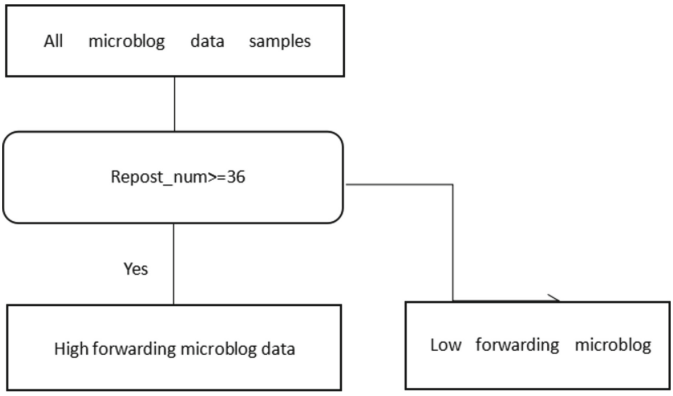

Fig. 3. Data cleaning diagram

4.2 Data Mining

In the LDA topic model processing procedure, the cleaned microblog text data is first processed by 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 can be calculated to establish the LDA model and start our training procedure. After several times of trainings, the stop words are selected and included in the stop

word library, and the model is imported to continue training. After the training process, we can identify the output theme and the keywords of each topic as needed. Since the microblog text is very short, it makes the topic classification more difficult and time-consuming to debug and find the right theme, the number of keywords, and the stop thesaurus. To achieve the balance between classification accuracy and interpretability closed to common-sense, we need to keep updating the parameters. 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):

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):

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 frequently reposted theme through the LDA topic model, the features of the predicted category will be viewed as the input 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 microblog is the first item of this topic, and then based on the algorithm, we classify the microblog into the subject category it belongs to. More specifically, we carry out the training phase of machine learning algorithms based on the training set and then conduct the related tests on the test set to calculate whether the categories of these microblogs belong to the frequently reposted category. If they belong to that, they will be reposted, otherwise they will not be reposted. After all algorithm processing is performed, the output data is retained and left for verification.

The output data of Naive Bayes algorithm is presented as follows (Table 3):

Table 3. The output data of Naive Bayes algorithm

	Total	The number of microblog judged not to be forwarded	The number of microblog 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 reposted in the test set are extracted. Comparing with the truly reposted microblogs obtained by the first round of data cleaning, we can finally evaluate the accuracy of our proposed method using the criteria of the predicted and verified recall and precision.

5 Analysis of Verification Results

For the results obtained, we use precision and recall rate to measure the correctly predicted rate and accuracy of the above experimental model. The precision rate refers to the proportion of microblogs that are correctly predicted to be reposted among the items that are predicted positive. For example, the microblogs being reposted have an accuracy rate of a/(a + c). The recall rate is the correctly predicted microblog among all relevant elements that are truly positive. 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 precision rate, and the microblog concerning safety warnings issued by the public security official microblog (pufa push, fraud prevention, security common sense push, etc.) tend to be easier to be

reposted. It is reposted and shared by users, which is also a true reflection of the real life of the masses. The model can accurately predict those microblogs that are widely spread and more likely to cause sensation, which has a positive effect on guiding the public security departments to better build the microblog account of the police.

6 Summary

This paper first collects a large amount of online microblog data from microblog, simulating the real user environment on the Internet. Then through the data analysis with simple but efficient algorithm, it is found that the microblog text topic category is an important factor affecting its reposting rate. Based on this, we propose an experimental method to predict the reposting behavior of microblog. The method is based on the LDA model and adopts the Naïve Bayes algorithm for prediction. Experiments demonstrate that there are two themes related to popularity of the public security police microblog: social hotspot case notification and life safety. The final recall and precision of the model indicates the accurate prediction ability. Through the predictions of the model, the safety warning class (preventing fraud, etc.) is the most popular type of microblog that are frequently reposted by users. It can also be seen from the keywords of the displayed topic category that the users repost relevant contents most frequently 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 outlaws, it is a tool for committing crime. Today, cybercrime intensifies, for example, telecom fraud is aggravated rapidly recently. The "Xu Yuyu" incident caused a sensation throughout the country, and the public opinion was unprecedented. The public's awareness of prevention cannot be built 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. We 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, be careful when releasing every microblog, and be considerate of everything the masses care about.

In order to achieve the full flexibility of the new media police platform, account operators should actively look for social hot issues. From the perspective of public security, they can edit the microblog daily with the help of some conclusion points in this paper. We should grasp the principle of timeliness of information, grasp the situation of malignant events in some regions where fans make replies and ask for help under the released microblog immediately, and effectively analyze the clues. We need to give this new media platform more interactivity and endow people more opportunities to express themselves, instead of simply instilling and transmitting rigid news to people.

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. Many thanks to my

classmates, they live and study with me, and help me a lot. At last, I will thank my parents, they are my best teacher, and I love them so much, nothing I can do without them.

Funding Statement. This study has been supported by Jiangsu Social Science Foundation Project (Grant No: 20TQC005). Philosophy Social Science Research Project Fund of Jiangsu University (Grant No: 2020SJA0500). The 13th Five-Year Plan Project of Educational Science in Jiangsu Province "Research on the reform and innovation of network public opinion teaching in public security colleges and universities from the perspective of overall national security" (Grant 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" (Grant 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 (Grant No: 2019A30).

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

References

1. Zhang, J., Yin, J., Hu, C.T.: Prediction of microblog forwarding based on active fan forwarding influence model. J. Jiangsu Police Officer Acad. **34**(1), 116–121 (2019)
2. Liu, W., He, M., Liu, Y., Shen, H.W., Cheng, X.Q.: Research on Weibo reposting prediction based on user behavior characteristics. J. Comput. **39**(10), 1992–2006 (2016)
3. Tian, L., Ren, G.H., Wang, W.: Prediction of Weibo user reposting behavior oriented to reading promotion. J. Inf. **36**(11), 1175–1182 (2017)
4. Deng, Q., Ma, Y.F., Liu, Y., Zhang, H.: Prediction of microblog forwarding based on BP neural network. J. Tsinghua Univ. (Sci. Technol.) **55**(12), 1342–1347 (2015)
5. Mu, S.K., Zhang, L.Q., Teng, C.F.: Prediction of microblog forwarding behavior based on recurrent neural network. Comput. Syst. **28**(8), 155–161 (2019)
6. Guan, P., Wang, Y.F., Fu, Z.: Analysis of the extraction of scientific literature subjects based on LDA topic model under different corpus. Libr. Inf. Serv. **60**(2), 112–121 (2016)
7. Zhao, H.D., Liu, G., Shi, C., Wu, B.: Prediction of microblog forwarding based on forward propagation process. Chin. J. Electron. **44**(12), 2989–2996 (2016)
8. Li, Y., Chen, Y.H., Liu, T.: A review of microblog information propagation prediction research. J. Softw. **27**(2), 247–263 (2016)
9. Hagen, L.: Content analysis of e-petitions with topic modeling: how to train and evaluate LDA models? Inf. Process. Manage. **54**(6), 1292–1307 (2018)