Intelligent Classification and Automatic Annotation of Violations based on Neural Network Language Model

Yaoquan Yu
Qingyuan Power Supply Bureau
Guangdong, Qingyuan, China
245860898@qq.com

Yuefeng Guo
Qingyuan Power Supply Bureau
Guangdong, Qingyuan, China
673224796@qq.com

Zhiyuan Zhang
South China University of Technology
School of Electric Power Engineering
Guangdong, Guangzhou, China
epzzy@mail.scut.edu.cn
Correspondence author

Mengshi Li
South China University of Technology
School of Electric Power Engineering
Guangdong, Guangzhou, China
mengshili@scut.edu.cn

Tianyao Ji
South China University of Technology
School of Electric Power Engineering
Guangdong, Guangzhou, China
tyji@scut.edu.cn

Wenhu Tang
South China University of Technology
School of Electric Power Engineering
Guangdong, Guangzhou, China
wenhutang@scut.edu.cn

Qinghua Wu
South China University of Technology
School of Electric Power Engineering
Guangdong, Guangzhou, China
wuqh@scut.edu.cn

***Abstract*—To solve the problem of lack on automatic classification and annotation of a large number of violation cases in power supply enterprises, a Neural Network Language Model (NNLM) based on word vector is proposed in this paper. This model adopts word vector to represent text features, extracts essential features of text information by using neural network. Simulation studies are carried out on data collected from South China Grid to evaluate the performance of NNLM, the result of which is compared with Naive Bayes Classifier (NBC) and Logistic Regression (LR). The experimental results show that the classification accuracy of the NNLM is as high as 99%, which is much higher than classification accuracy of the NBC and LR. The NNLM effectively improves the accuracy rate of classification and the efficiency of annotation.**

***Index Terms*—classification, annotation, NNLM, violation**

I. Introduction

Although the self-healing ability of power system is getting stronger and stronger at present, the stability and reliability of power system are getting higher and higher, the production safety accidents can not be completely avoided. All kinds of safety supervision data, such as safety accidents, on-site violation records, inspection and audit problems, are the first-hand data of safety production and the guiding basis of safety production. It is of great practical significance to study the historical safety data of power supply enterprises comprehensively, which can not only reduce the accident probability but also improve the level of safety production.

Historical safety data is a typical multi-source heterogeneous data, the first problem to be solved is to standardize and format the data. In the process of data standardization, the classification and annotation of a large amount of historical safety data are very important. Traditional text classification methods are mainly divided into knowledge engineering classification and machine learning classification.

The knowledge engineering classification is to classify the text manually according to the defined rules [1]. The way of analyzing data manually has many disadvantages: work efficiency is too low, and data format is not standard; the ability of manual processing is limited, and the analysis of safety data is not comprehensive and sufficient; the level of automation and intelligence is low, the safety measures based on human experience are lack of reliability.

Nowadays, the most common used classification method is based on machine learning. Support Vector Machines (SVM), NBC and LR are all commonly used machine methods. NBC [2] is a classical machine learning classification method based on probability calculation. Unfortunately, its performance is poor because it can't handle text data well [3]. SVM puts long text as a research object. When processing short text, the performance of SVM is poor, because the features of short text are few and the data is irregular [4]. LR is a classification method based on linear regression theory [5], [6]. Due to the shortcomings of maximum likelihood method, when the dataset dimension is high, the estimation results may

978-1-7281-6926-2/20/$31.00 ©2020 IEEE

be unstable [7].

The emergence of big data and artificial intelligence technology has brought new development opportunities for the safety supervision of electric power industry [8]. The deep learning technology of artificial intelligence has achieved good results in text classification and has gradually replaced the traditional machine learning methods [9]. Deep learning can automatically extract features from a large number of data and describe the target more precisely [10]. A Convolutional Neural Network (CNN) is proposed for the classification and prediction of microblog public opinion in [11], the experimental results show that it is more accurate than the traditional machine learning algorithm. Unfortunately, the size of CNN filter is fixed, it is difficult to model long series data and the adjustment of filter size parameter is a complex work [12]. However, Recurrent Neural Network (RNN) [13] can extract feature information with variable length.

The model proposed in this paper consists of jieba cutter, word2vector model and a RNN model. The proposed model uses jieba cutter to cut each violation into words. In this step, some stopwords are removed and only representative keywords are selected as learning samples. Then the word2vector model is used to transform keywords into eigenvectors. The eigenvectors of all keywords in a violation form a characteristic matrix. The characteristic matrix representing every violation is put into the RNN model for training, after sufficient training, a deep belief network with the ability of classifying violation categories will be obtained. To evaluate the performance of NNLM, a comparative experiment with NBC and LR models was carried out.

II. Methodology Background

A. Jieba word cutter

Jieba cutter is a probability language model cutter, which can find the best segmentation scheme in all segmentation schemes. The steps of Jieba word cutter are divided into the following three steps. Firstly, based on prefix dictionary, an efficient word graph scanning is implemented to generate a Directed Acyclic Graph (DAG) of all possible words in a sentence. Secondly, dynamic programming is used to find the maximum probability path and the maximum segmentation combination based on word frequency. Finally, for words that are not in the dictionary, Hidden Markov Model (HMM) [14] and and Viterbi algorithm [15] is adopted to cut words.

B. Word2vector

Word2vector is a Natural Language Processing (NLP) [16] model, which transforms a word into a eigenvector. For example, there is a sentence, "cats like to eat fish". If we use vectors to represent every word in this sentence, the first thing we think of is to use one hot code, as shown in Fig. 1. Every vector is orthogonal to each other, that is, there is no relationship between every word. In the sentence above, there are five different words, so the dimension of the vector is 5. However, there are many different words in an article. In this way, the dimension of the vector will be very high and it ignores the relationship between words. For instance, "cats" and "fish" are both animals, they are not unrelated.

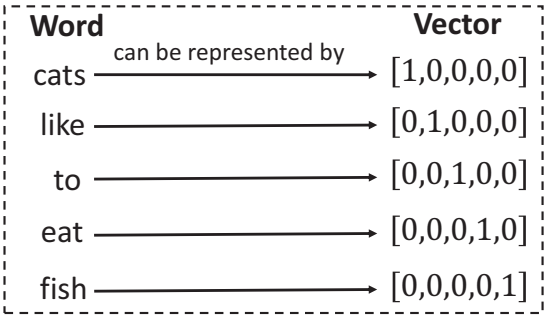

Fig. 1. Word vector diagram.

In order to reduce the vector dimension and express the correlation between words, word2vector uses Continuous Bag-of-Word Model (CBOW) [17] and Skip-Gram [18] to implement word embedding. In this way, the correlation between similar words can be expressed and the dimensions of vectors are greatly reduced.

C. Long Short-Term Memory

Long Short-Term Memory (LSTM) [19] is an excellent variant model of RNN and has better performance than traditional RNN. Compared with the general neural network, RNN can process the data of variable sequence. For example, the meaning of a word will vary according to the previous content, RNN can solve this kind of problem very well. While traditional RNN has the problem of gradient disappearance, LSTM solves this problem well. LSTM can memorize valuable information and give up redundant memory, so as to reduce the difficulty of learning. Compared with traditional RNN, input gate i_t, forget gate f_t, output gate o_t and internal memory unit are added to the neurons of LSTM.

As shown in Fig. 2, the first step is to check the previous hidden vector h_{t-1} and input value x_t, and decide whether the information should be discard. The next step is to decide how much new information should be added into the cell state. According to the decision of the above two steps, the cell state c_t is updated. The i_t, f_t and c_t can be expressed as

$$i_t = \sigma\left(W_{\mathrm{xi}}x_t + W_{\mathrm{hi}}h_{t-1} + W_{\mathrm{ci}}c_{t-1} + b_{\mathrm{i}}\right) \quad (1)$$

$$f_t = \sigma\left(W_{\mathrm{xf}}x_t + W_{\mathrm{hf}}h_{t-1} + W_{\mathrm{cf}}c_{t-1} + b_{\mathrm{f}}\right) \quad (2)$$

$$c_t = f_t c_{t-1} + i_t \tanh\left(W_{\mathrm{xc}}x_t + W_{\mathrm{hc}}h_{t-1} + b_{\mathrm{c}}\right) \quad (3)$$

The output value is determined by the cell state. The cell state is processed by tanh function (to get a value between -1 and 1) and multiplied by the output of σ gate.

$$o_t = \sigma\left(W_{\mathrm{xo}}x_t + W_{\mathrm{ho}}h_{t-1} + W_{\mathrm{co}}c_t + b_{\mathrm{o}}\right) \quad (4)$$

$$h_t = o_t \tanh\left(c_t\right) \quad (5)$$

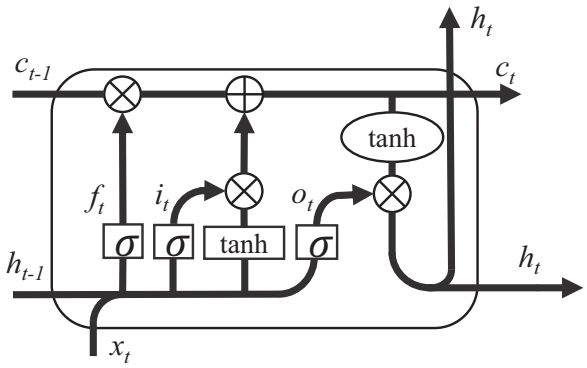

Fig. 2. The structure of an LSTM unit.

TABLE I
MEANING OF VARIABLE IN FORMULA

Variable	Meaning
σ	the sigmoid function
b_{i}	the bias for input gate
b_{f}	the bias for forget gate
b_{o}	the bias for output gate
b_{c}	the bias for memory cell
b_{y}	the bias for output
W_{xc}	the weight between input and memory cell
W_{hc}	the weight between hidden vector and memory cell
W_{hy}	the weight between hidden layer and output
W_{xi}	the weight between input and input gate
W_{xf}	the weight between input and forget gate
W_{xo}	the weight between input and output gate
W_{hi}	the weight between hidden vector and input gate
W_{hf}	the weight between hidden vector and forget gate
W_{ho}	the weight between hidden vector and output gate
W_{ci}	the weight between memory cell and input gate
W_{cf}	the weight between memory cell and forget gate
W_{co}	the weight between memory cell and output gate

Finally the output y_t is calculated by

$$y_t = W_{\mathrm{hy}} h_t + b_{\mathrm{y}} \tag{6}$$

All variables appearing in the above formulas are illustrated in Table I.

III. NEURAL NETWORK LANGUAGE MODEL

A. Stopwords Filter

Because the violation records obtained from power supply enterprises contain some functional words and symbols, such as determiners, prepositions, punctuation marks and mathematical symbols, the first problem to be solved is to remove these stopwords to save storage space and improve search efficiency.

In this paper, jieba cutter is used as filter to preprocess violation records, which filters out stopwords and extracts representative keywords in the violation records. The process of filtering words is shown in Fig. 3, for example, " one staff didn't wear the helmet " is cut into six words, " one " and " the " are judged as stopwords and filtered out by filter, " staff ", " didn't ", " wear " and " helmet" are judged as keywords and preserved to represent this violation.

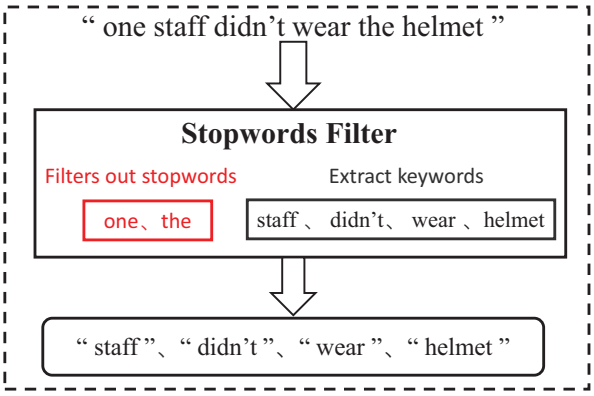

Fig. 3. Stopwords Filter.

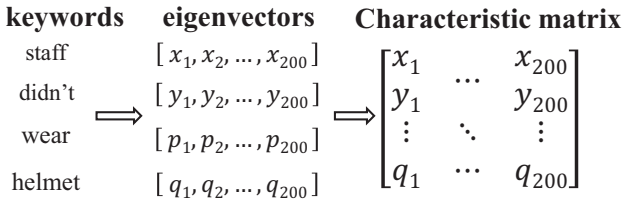

Fig. 4. Characteristic matrix generation.

B. Characteristic matrix generation

Through the Jieba filter, the keywords representing every violation are obtained, but the data to be sent to the neural network for learning must be in numerical form rather than in text form. Word2vector is adopted to convert keywords into eigenvectors in this paper, the dimension of eigenvectors is set to 200. Each keyword is represented by a eigenvector, and all eigenvectors in each violation form a characteristic matrix to represent the violation. As illustrated in Fig. 4, the Characteristic matrix in Fig. 4 can represent " one staff didn't wear the helmet ".

C. The Proposed Model

In this paper, a model for classifying and annotating violations is proposed, the schematic diagram of which is given in Fig. 5. The implementation of the model is divided into the following steps:

1) Use Jieba cutter to preprocess the original violation records, remove the stopwords and extract the keywords that can represent the violation.

2) Transform the keywords obtained in the first step into eigenvectors by using word2vector.

3) Convert the eigenvectors of all keywords in each violation record into a characteristic matrix, and make the corresponding label of each violation record.

4) Put the characteristic matrixes and the corresponding labels of the violation records into the LSTM neural network for training.

5) Classify the testdata and analyze the loss rate and classification accuracy.

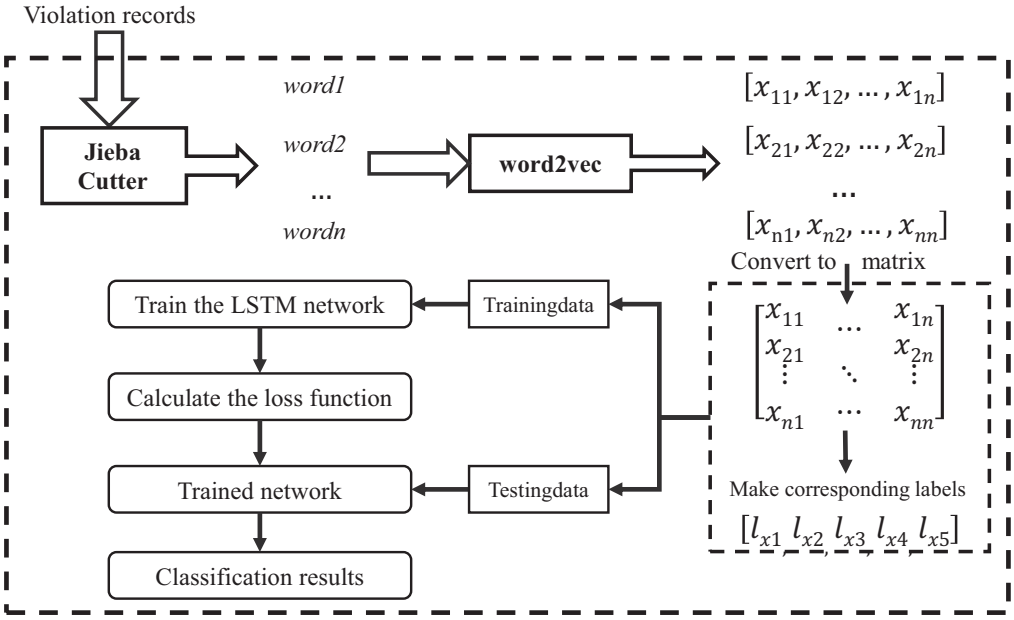

Fig. 5. The proposed model.

IV. Simulation studies

A. Data Description

The data of the experiment is collected from South China Grid. As shown in Fig. 6, it includes five categories: management, behavior, two-ticket, tools and environment, 1660 samples in total. The data virtually includes all types of violations in the current power grid. If there are new types of violations in the power grid, the classification network will train and learn the new types of violations, so as to ensure the reliability and accuracy of the classification network. Management violation refers to operation and behavior in violation of grid management regulations. Behavior violation refers to the personal behavior in violation of grid safety regulations, such as being improperly dressed or operating without authorization. Two tickets refer to work ticket and operation ticket, two-ticket violation refers to the behavior of working without tickets or working in violation of tickets. Tools violation refers to the behavior of using or placing tools irregularly. Environment violation refers to that the working environment of workers does not meet the requirements of safety regulations. Each category gives an example in Table II. From each category, 80% of the samples are randomly selected as training data, which are used to build the classifier model, and the remaining 20% of the data are used to verify the accuracy of the classifier.

B. Performance Evaluation

Four evaluation criteria, precision (Pre), recall (Rec), F1-score (F1) [20] and Missing Alarm (MA) are selected to evaluate the experimental results. In this paper, four classification results are defined.

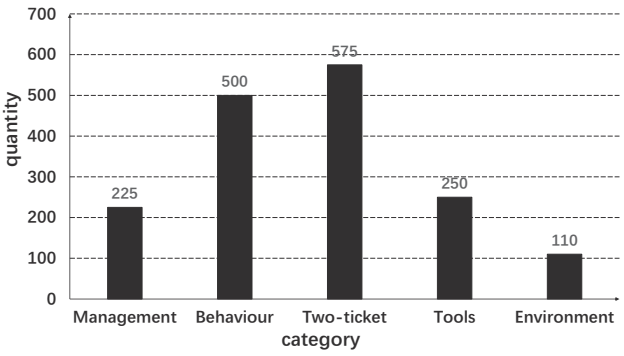

Fig. 6. Category and quantity of data.

TABLE II
Examples of each category of violation

Category	Example of each category
Management	Management No safety agreement with contractor
Behavior	One staff didnt wear the helmet
Two-ticket	Use of non-standard operation ticket
Tools	The knife gate is seriously rusted
Environment	There is no fence at the work site

T_P: the classification is correct.

F_P: violations that do not belong to this category are classified as this category.

F_N: Violations of this category are classified as other categories.

T_N: Violations that do not belong to this category are classified as other categories.

Four evaluation criteria are defined as

TABLE III
ACCURACY ANALYSIS OF CLASSIFICATION RESULTS

Violation category	NBC				LR				NNLM			
	Pre	**Rec**	**F1**	**MA**	**Pre**	**Rec**	**F1**	**MA**	**Pre**	**Rec**	**F1**	**MA**
Management	0.900	0.900	0.900	0.100	0.778	0.700	0.737	0.300	**1.000**	**1.000**	**1.000**	**1.000**
Behavior	0.727	0.800	0.762	0.200	0.778	0.700	0.737	0.300	**0.990**	**0.990**	**0.990**	**0.010**
Two-ticket	0.75	0.900	0.818	0.100	0.727	0.800	0.762	0.200	**1.000**	**0.990**	**0.995**	**0.010**
Tools	0.727	0.800	0.762	0.200	0.818	0.900	0.857	0.100	**0.990**	**1.000**	**0.995**	**0.000**
Environment	**1.000**	0.600	0.750	0.400	0.900	0.900	0.900	0.100	**1.000**	**1.000**	**1.000**	**0.000**

Note: Bold values indicate the result corresponding to the highest accuracy in each case.

$$\mathrm{Pre} = \frac{\mathrm{T_P}}{\mathrm{T_P} + \mathrm{F_P}} \tag{7}$$

$$\mathrm{Rre} = \frac{\mathrm{T_P}}{\mathrm{T_P} + \mathrm{F_N}} \tag{8}$$

$$\mathrm{F1} = \frac{2\mathrm{Pre} \cdot \mathrm{Rre}}{\mathrm{Pre} + \mathrm{Rre}} \tag{9}$$

$$\mathrm{MA} = \frac{\mathrm{F_N}}{\mathrm{T_P} + \mathrm{F_N}} \tag{10}$$

where Pre reflects the overall performance of the classifier, the higher the pre is, the higher the accuracy of classifier is. Rec measures the ability of classifier to recognize positive samples. F1 is a weighted harmonic average of Rec and Pre, the higher F1 is, the better the performance of classifier is. MA reflects how many positive cases have been missed, the smaller Ma is, the better the performance of classifier is.

C. Comparison of three classification models

Naive Bayes model originates from classical mathematical theory, it performs well on small-scale data and can handle multiple classification tasks. Its algorithm is also relatively simple, commonly used in small-scale text classification. NBC assumes that the attributes of each variable are independent of each other, which is often not true in practical application. When the attribute correlation is small, NBC has good performance. However, when the number of attributes is relatively large or the correlation between attributes is large, the classification effect is not good. NBC model needs to know the prior probability, which depends on the hypothesis. Therefore, in some cases, its classification effect will not be ideal due to the improper hypothesis.

LR is a traditional machine learning model, which is widely used because of its simplicity and efficiency. It can also solve the problem of overfitting by regularization of parameters. However, it is easy to under fit in the training process, resulting in low accuracy. Moreover, it is essentially a linear classifier, so it can't deal with the nonlinear correlation between features well. And it lacks the ability to deal with a large number of multi class features and variables. As a result, when the sample size is large, LR classification performance is poor.

LSTM is an excellent variant model of RNN and has better performance than traditional RNN. Traditional RNN can only memorize part of the sequence, so its performance in long sequence is far inferior to that in short sequence, which results in the decrease of accuracy once the sequence is too long. LSTM solves this problem by adding three control units, namely input gate, output gate and forgetting gate. The information will be remembered if it is in accordance with the rules; if not, it will be forgotten. In this way, the long-sequence dependence problem in networks can be solved. LSTM can also solve the gradient disappearance problem resulting from the process of gradient backpropagation, which is superior to the traditional RNN model. Moreover, the length of the text is usually variable, LSTM can extract feature information with variable length, which can improve the accuracy of classification.

D. Comparative Experiment

In order to make the comparative experiment convincing, NBC and LR are used to test the same data. The classification results of three models are detailed in Figs. 7, 8, 9 and 10. As shown in figures, the proposed model has the highest Pre, REC and F1 and the smallest MA, which indicates that the proposed model has the highest classification accuracy among three models in all categories.

To make a more comprehensive comparative study, numerical results of the three models for all categories have been summarised in Table III. As can be seen from Table III, the average Pre, Rec and F1 of NNLM (99.6%, 99.6% and 99.6%) are much higher than NBC (82.1%, 80% and 79.8%) and LR (80%, 80% and 79.9%), and the MA of NNLM (0.4%) is much smaller than NBC (20%) and LR (20%), which means that the proposed model performs much better than NBC and LR.

V. CONCLUSION

This paper proposed an intelligent classifier based on LSTM and NLP technology. Jieba cutter and word2vector model are used to transform the violation records in text form into the characteristic matrices in numerical form that can be recognized and processed by computer. LSTM is adopted to implement intelligent classification and automatic annotation of violation records. As an excellent variety of RNN, LSTM can extract feature information with variable length. Besides, compared with the traditional RNN, LSTM solve the problem of gradients disappearing or exploding by adding three gates to neurons. Comparative experiments with NBC and LR have been carried out on the same data from several power

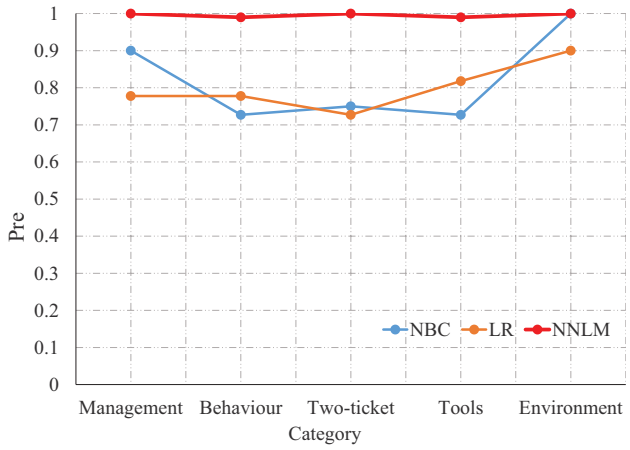

Fig. 7. The pres of three models.

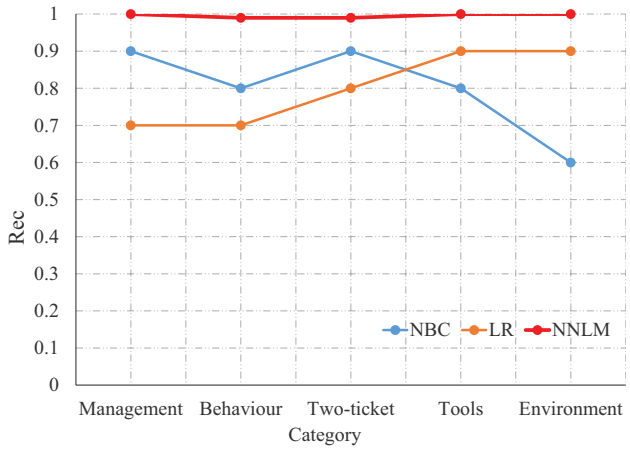

Fig. 8. The Recs of three models.

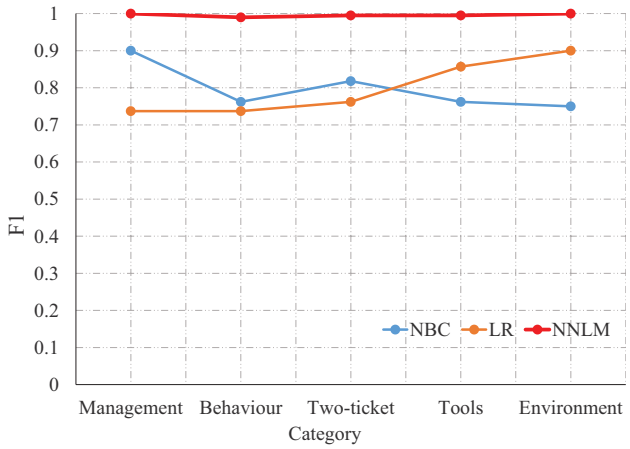

Fig. 9. The F1s of three models.

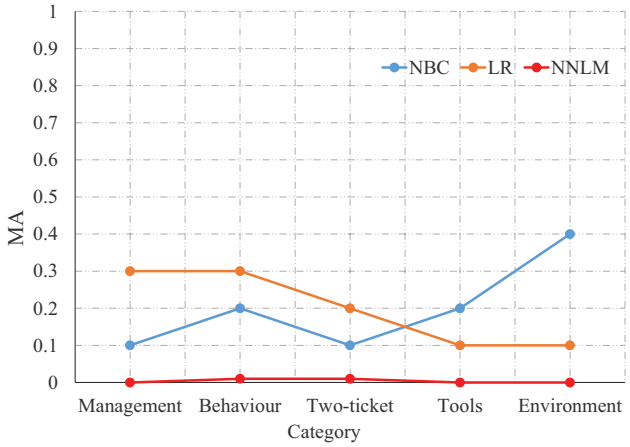

Fig. 10. The MAs of three models.

supply enterprises located in South China. The results have demonstrated that NNLM has much higher accuracy and much lower missing alarm rate than NBC and LR model. What's more, classification network will keep learning and training with the update of violation database. If there are new types of violations in the power grid, the classification network will train and learn the new types of violations, so as to ensure the reliability and accuracy of the classification network.

ACKNOWLEDGMENT

The work is jointly funded by Research Project of Qingyuan Power Supply Bureau, China Southern Grid (GDKJXM20183525) and the Fundamental Research Funds for Central Universities, South China University of Technology (2019MS014).

REFERENCES

[1] F. Miao and P. Zhang and L. Jin and H. Wu. Chinese News Text Classification Based on Machine Learning Algorithm. In *2018 10th International Conference on Intelligent Human-Machine Systems and Cybernetics (IHMSC)*, pages:48-51, Aug 2018

[2] S. Goswami and P. Bhardwaj and S. Kapoor. Naive bayes classification of DRDO tender documents. In *2014 International Conference on Computing for Sustainable Global Development (INDIACom)*, pages:593-597, March 2014

[3] Ikonomakis, Emmanouil and Kotsiantis, Sotiris and Tampakas, V. Text Classification Using Machine Learning Techniques. *WSEAS transactions on computers*, 4(8):966-974, Aug 2005

[4] C. Yin and J. Xiang and H. Zhang and J. Wang and Z. Yin and J. Kim. A New SVM Method for Short Text Classification Based on Semi-Supervised Learning. In *2015 4th International Conference on Advanced Information Technology and Sensor Application (AITS)*, pages:100-103, Aug 2015

[5] Kotsiantis, S. B. and Zaharakis, I. D. and Pintelas, P. E. Machine learning: a review of classification and combining techniques. *Artificial Intelligence Review*, 26(3):159–190, Nov 2006

[6] M. Y. Helmi Setyawan and R. M. Awangga and S. R. Efendi. Comparison Of Multinomial Naive Bayes Algorithm And Logistic Regression For Intent Classification In Chatbot. In *2018 International Conference on Applied Engineering (ICAE)*, pages:1-5, Oct 2018

[7] H. Park and Y. Shiraishi and S. Imoto and S. Miyano. A Novel Adaptive Penalized Logistic Regression for Uncovering Biomarker Associated with Anti-Cancer Drug Sensitivity. In *IEEE/ACM Transactions on Computational Biology and Bioinformatics*, 14(4):771-782, July 2017

[8] L. Qing and Z. Boyu and L. Qinqian. Impact of big data on Electric-power industry. In *2017 IEEE 2nd International Conference on Big Data Analysis (ICBDA)*, pages:460-463, March 2017

[9] Tang, Duyu and Qin, Bing and Liu, Ting. Deep learning for sentiment analysis: Successful approaches and future challenges. *Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery*, 5: 292-303, Nov 2015

[10] C. Li and G. Zhan and Z. Li. News Text Classification Based on Improved Bi-LSTM-CNN. In *2018 9th International Conference on Information Technology in Medicine and Education (ITME)*, pages:890-893, Oct 2018

[11] X. Wang and J. Li and Y. Liu. Application of Convolutional Neural Network (Cnn)in Microblog Text Classification. In *2018 15th International Computer Conference on Wavelet Active Media Technology and Information Processing (ICCWAMTIP)*, pages:127-130, Dec 2018

[12] J. Cai and J. Li and W. Li and J. Wang. Deeplearning Model Used in Text Classification. In *2018 15th International Computer Conference on Wavelet Active Media Technology and Information Processing (ICCWAMTIP)*, pages:123-126, Dec 2019

[13] Liu, Pengfei and Qiu, Xipeng and Huang, Xuanjing. Recurrent Neural Network for Text Classification with Multi-Task Learning. May 2016

[14] M. I. Mohd Yusoff and I. Mohamed and M. R. A. Bakar. Hidden Markov models: An insight. In *Proceedings of the 6th International Conference on Information Technology and Multimedia*, pages 259-264, Nov 2014

[15] A. J. Viterbi. A personal history of the Viterbi algorithm. *IEEE Signal Processing Magazine*, 23(4):120-142, July 2006

[16] Y. A. Solangi and Z. A. Solangi and S. Aarain and A. Abro and G. A. Mallah and A. Shah. Review on Natural Language Processing (NLP) and Its Toolkits for Opinion Mining and Sentiment Analysis. In *2018 IEEE 5th International Conference on Engineering Technologies and Applied Sciences (ICETAS)*, pages:1-4, Nov 2018

[17] Q. Wang and J. Xu and H. Chen and B. He. Two improved continuous bag-of-word models. In *2017 International Joint Conference on Neural Networks (IJCNN)*, pages:2851-2856, May 2017

[18] C. Zhang and X. Liu and D. Biś. An Analysis on the Learning Rules of the Skip-Gram Model. In *2019 International Joint Conference on Neural Networks (IJCNN)*, pages:1-8, July 2019

[19] V. Jithesh and M. J. Sagayaraj and K. G. Srinivasa. LSTM recurrent neural networks for high resolution range profile based radar target classification. In *2017 3rd International Conference on Computational Intelligence Communication Technology (CICT)*, pages:1-6, Feb 2017

[20] Marina Sokolova and Guy Lapalme. A systematic analysis of performance measures for classification tasks. *Information Processing and Management*, 45(4):427-437, May 2009