https://dx.doi.org/10.22606/fsp.2020.41005

Spam Comment Recognition Based on Wide & Deep Learning

Meiling Fu, Daji Ergu*

Key Laboratory of Electronic and Information Engineering,Southwest Minzu University,State Ethnic Affairs Commission, Chengdu, Sichuan 610000, China
Email: ergudaji@163.com

Abstract. The flood of e-commerce platform spam comments affects consumers' purchasing decisions, which greatly damages the interests of consumers. In the process of spam comment recognition, the explicit discrete features of spam comments were usually used as the input of the model. This paper combines the implicit semantic features of spam comments and the explicit discrete features of spam comments to identify the spam comment. First, SMOTE oversampling method is used to balance positive and negative sample sets. Then, wide & deep model, a recommendation system model proposed by Google, is improved and applied to one of the public datasets of spam comment recognition and the commodity datasets collected from one of the biggest e-commerce platform in China. The experimental results show that the improved algorithm can achieve good results in both the gold-standard opinion spam datasets and the commodity datasets.

Keywords: Wide & deep, spam comment, SMOTE, recognition.

1 Introduction

In recent years, with the development of the Internet and the diversified demands of consumers for consumption channels, offline -shopping has gradually been replaced by online -shopping, and a large number of e-commerce platforms have become the choice of consumers[7]. The number of e-commerce platforms has thus been rapidly increased including an increasing scale and a growing variety of commodities. China's singles' day sales on Tmall in 2018 reached 213.5 billion yuan ($30.67 billion), showing that online shopping has become one of the main ways people consume today[①]. The quality and quality of online products are mostly uncontrollable and experiential, consumers often browse the reviews of products before they shop online, and the polarity of product reviews will directly affect consumers' purchasing decisions[7]. According to the surveys, 43.3% of customers will refer to other customers' comments before purchasing[②]. Another survey found that positive comments prompted 87% of people to make a purchase decision, while 80% of customers decided to abandon the purchase because of negative comments. It can be seen that the historical review of commodities is crucial to the decision-making of potential consumers. The polarity and quality of the comments could largely determine whether a consumer will purchase the product, which results in many of the Spam comments. Spam comments are usually divided into fake comments and irrelevant comments. Driven by the interests, more and more businesses hire people to write fake comments to influence consumers' buying decision, which greatly deplete the interests of consumers, or to raise some irrelevant comments in order to increase the sales of the stores, affecting the environment of the online shopping platform[7]. To manage the shopping environment of the e-commerce platform, the Fifth Session of the Standing Committee of the 13th National People's Congress voted to pass the *E-commerce Law*, which will be implemented in January 2019.

The identification of spam comments is mainly divided into three categories: (1) identification of spam comments; (2) identification of spammers; and (3) identification of spam groups. Most of the existing methods mainly extract the signs manually or by algorithms, and then classify them by classifiers. For instance, Shehnepoor et al. [1] extracted the comments and user behavior characteristics in the dataset to construct a heterogeneous information network. Then, a weighting algorithm was used to calculate the importance of each node, the weighted and updated features were classified as input to

[①] http://finance.sina.com.cn/china/2018-11-12/doc-ihnstwwq7580023.shtml

[②] https://max.book118.com/html/2017/0114/83652263.shtm the classifier. Rajamohana et al. [2] proposed a global optimization technique—the adaptive binary pollination algorithm extracts features to reduce the set of selected features. Jia et al. [4] employed the LDA model to extract language features and classified spam comments into false reviews and real comments. Xu et al. [5] constructed suspicious commenter graphs based on user behavior characteristics, and detected spam posting groups based on CPM in a completely unsupervised manner. Li et al. [6] proposed a new feature extraction method based on the characteristics of spammers and spam comments, and the gradient enhancement tree algorithm was used to construct the spam review classifier.

Much attention has been paid to the Spam comments by constructing models based on the characteristics of comments and comment texts, ignoring merchant features, characterizing models based on manually selected features or algorithms, or employing unsupervised or supervised methods to train classifiers. The imbalance of the datasets is ignored in the supervised training process. Therefore, the smote method is used to balance the spam comments datasets by combining the explicit features of the merchant-reviewer-comment text, the wide & deep model is improved and applied to spam comments recognition.

The remainder of this paper is organized as follows. Section 2 reviews the existing models. In Section 3, the datasets required for the experiment are described. Then, we conducted an experiment and analysed the experimental results in Section 4. The paper is summarized and concluded in Section 5.

2 Introduction of Model

2.1 Construction of Characteristic Indicators

According to the feature selection and practice in the current studies, we select 8 discrete features as the input of the wide part of the improved wide & deep model.

Whether the comment text is liked: the real comment will be praised by many other commentators, reflecting the credibility of the product to a certain extent.

Whether the comment text is copied: In order to save time and earn higher profits, the click farm will write a simple comment or directly copy other people's comments.

Whether the comment text is commented: A true commentary may cause other commenters to discuss it under the comment, or ask the reviewer for some product information to determine whether they will buy the item.

The subject classification of the text: The comment text of the product can be divided into many categories according to different themes. Most of the garbage reviews made by click farm are typed casually in order to save time, that is, comments of irrelevant items. SVD model can be used to extract the topic of comments and conduct topic classification of the comment text.

Whether the reviewer's time for reviewing is the same day: the click farm often repeats the same thing every day, they will be asked to add comments to the product, but the added comments are not as good as the real reviewers will add after the product has been used. In order to get the benefits as soon as possible, they will add directly on the day after the comment, and they will complete the order.

The credential rating of the reviewer: The level of the reviewer reflects the credibility of the reviewer to a certain extent. A click farm will have multiple accounts, and most of their accounts are newly registered with a very low level. We divide commenters' credit rating into 5 levels such as Registered members, Bronze members, Silver members, Gold members, and Diamond members.

Whether the business is cheating: In order to arouse some reviewers' desire to buy, some merchants write on the front page of the goods such as positive feedback and other violations of business rules.

The credit rating of the merchant: Each merchant will have an overall credit rating, which is a response to the overall credit of the company for many years, and also reflects the overall embodiment of the reputation of the merchant who purchased the product in the store.

2.2 Smote Model

The sample imbalance problem will cause the trained classifier to be ineffective because of the unbalanced distribution of data sets, the SMOTE model [17] is used to oversample the data sets and solve the problem of sample imbalance. The application will have a large deviation in the actual dataset.

The basic idea is to manually synthesize new samples from a few samples, and calculate the Euclidean distance from each of the minority samples *I* and the surrounding minority samples, and select the *K* neighbors with the shortest distance. The sample imbalance ratio is the sampling magnification *N*. For each minority sample *I*, the *N* samples are randomly selected in the *K* nearest neighbor, and the generated new sample is any point between the minority sample *I* and the randomly selected sample.

2.3 Wide & Deep Model

The wide & deep [13] recommendation model is transplanted to the classification of spam reviews. The wide & deep model is a recommendation model released by Google in year 2016 and applied to the app recommendation system. The wide part of the model uses Logist linear regression, and inputs some discrete and cross-cutting features selected manually. The deep part uses the deep neural network to discover feature combinations that have not appeared in the historical data. The parameters of two models are optimized at the same time to obtain the optimal model.

2.4 Improved Wide & Deep Model

There is no connection between nodes and nodes in each layer of the traditional neural network. For natural language processing, the meaning of many words must be known according to the context. RNN appears to process sequence data, but when processing long sequence data, the problem of gradient disappearance occurs, LSTM is a time recurrent neural network[18]. Its appearance solves the problem of gradient disappearance of RNN since it is an excellent variant model of RNN. The LSTM inherits the characteristics of most RNN models and can extract the comments in the text, mine the implicit semantics of comment text. In this paper, the wide & deep model [13] is applied to spam recognition. The input part of wide is 8 features such as comment text-merchant-reviewer feature or crossover feature. The deep part changes the DNN part in Figure 1 to LSTM. The structure diagram of the improved wide & deep model is as follows:

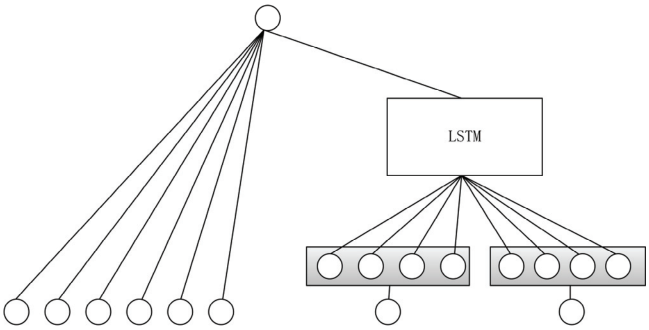

Figure 1. Improved wide & deep model

Firstly, the data is collected and pre-processed, then the characteristic index of the collected data is constructed, and the problem of sample imbalance of the data set is solved by SMOTE algorithm. We take the feature index extracted from the balanced data set as the input of the wide part, and then insert the text into the LSTM of the deep part to extract the text. The implicit semantics of this model is optimized by optimizing the parameters of the two models at the same time.

3 Datasets

3.1 Gold-Standard Opinion Spam Datasets

The gold-standard opinion spam dataset on Hotel Comments is a data set of 400 false comments constructed by Ott et al [16] through crowdsourcing platform and real comments on 20 Chicago hotels screened by multiple data sets on TripAdvisor website. This way can effectively imitate human behaviour. The data set is representative because of its physical characteristics. The data composition of the data set is shown in Table 1.

Table 1. Gold-standard opinion spam dataset.

	Number
Spam Comments	400
True comments	400

3.2 JD Datasets

In this paper, a mobile phone comment data set are collected by a crawler in Jingdong (in short JD), one of China's largest B2C online retailers, to test the improved model. The JD data set contains more than 8000 comments. After manual labelling, we found that there was obvious data imbalance in the data set. Thus the smote algorithm is employed to pre-process these imbalance data, and the positive and negative sample equalization are achieved. After smote oversampling, the total number of samples is more than 12,000, in which there are more than 6000 positive and negative samples respectively. The data set includes not only comment texts, but also eight features of merchants, sellers and reviews. Data sets are annotated manually, and multi-vote decision-making mechanism is adopted to annotate the data sets. The structure of JD Data Set is shown in Table 2.

Table 2. The structure of JD datasets

	Number	Feature
Spam Comments	6720	Whether the comment text is liked Whether the comment text is copied Whether the comment text is commented The subject classification of the text
True comments	6720	Whether the reviewer's time for reviewing is the same day The credential rating of the reviewer Whether the merchant is cheating The credit rating of the merchant

3.3 Evaluating Indicator

To evaluate the improved model, the commonly used evaluation indicators such as accuracy, recall rate and F1 value are used in this paper, as shown in Table 3. For spam reviews and real reviews, confusion matrix is used to describe TP, TN, FP, FN:

Table 3. Confusion matrix

	Spam Comments	True comments
Predicted as spam comment	TP	FP
Predicted as true comment	FN	TN

Accuracy, the proportion of the correct number of samples for the classifier to the total sample.

$$Accuracy = \frac{TP + TN}{TP + TN + FP + FN} \tag{1}$$

Recall, the proportion of the number of samples predicting the correct false comment as the sum of the total number of false samples and the number of samples that predicted the true comment as a false sample.

$$Recall = \frac{TP}{TP + FP} \tag{2}$$

F1 value, in reality we measure the quality of a classifier, we hope that the higher the accuracy and recall rate, the better, but in fact these two indicators cannot be high at the same time, so the F1 value is commonly used as a comprehensive indicator, comprehensive measurement accurate rate and recall rate.

$$F_1 = \frac{2TP}{2TP + FP + FN} \tag{3}$$

4 Result and Discussion

To verify the validity of the proposed model, the improved wide & deep model is used to calculate the above mentioned evaluation indicators based on the gold-standard opinion spam dataset and the JD dataset. We found that a high accuracy on both data sets could be obtained by the proposed model. The experimental results are shown in Figure 2.

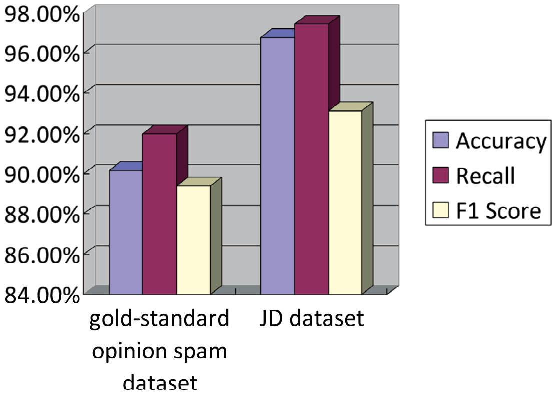

Figure 2. Experimental results on different datasets.

In the vertical comparison, the improved model is compared with some other popular models that are mainly used in the study of the spam comments such as SVM, Naive Bayes, CNN etc. The experimental results show that the improved wide & deep model is slightly better than other models in terms of accuracy, recall and F1 on the same dataset, as shown in Figure 3 and Figure 4.

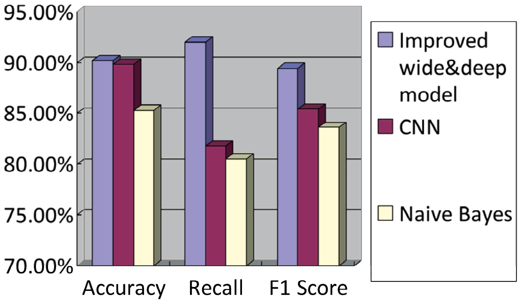

Figure 3. Comparison of different algorithms on the gold-standard opinion spam datasets.

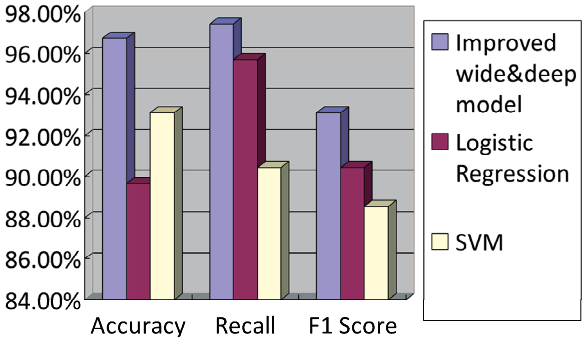

Figure 4. Comparison of different algorithms on the JD datasets.

It can be seen from Figure 3 and Figure 4, the accuracy, recall rate and F1 value of the proposed model are 90.18%, 92% and 89.41% on the gold-standard opinion spam dataset, and 96.78%, 97.45% and 93.12% on the JD dataset, respectively. Compared with some popular machine learning algorithms, such as SVM and Naive Bayes, the experimental results show that the improved model is more effective and accurate than other models in both tested Datasets.

5 Conclusions and Future Work

In the previous spam comment research, most of the models are extracted by hand-extracted features. Based on the hand-extracted features of the predecessors, the 8 features extracted from the review text-business-reviewer are combined in the wide & deep model. In the input of the wide part, comment text is represented by *word vector*, as the input of the deep part. The implicit semantic relationship is mined, and two parts of wide & deep model are trained together to adjust the parameters of the whole model, and it is proved that the proposed model is effective and accurate both in horizontal or vertical comparison. However, the spam comment research relies heavily on the manually labeled data set to some extent. Therefore, we will look for a better quality data set to further verify the validity of the proposed model in future. At the same time, the method is equally applicable to the construction of corpora.

Acknowledgements. This work was supported by grants from the National Natural Science Foundation of China #71774134 and #71373216,in part by the Innovation Scientific Research Program for Graduates in Southwest Minzu University(No.CX2019SZ12).

References

1. Shehnepoor S, Salehi M, Farahbakhsh R, et al. NetSpam: A network-based spam detection framework for reviews in online social media[J]. IEEE Transactions on Information Forensics and Security, 2017, 12(7): 1585-1595.
2. Rajamohana S P, Umamaheswari K, Abirami B. Adaptive binary flower pollination algorithm for feature selection in review spam detection[C]//2017 International Conference on Innovations in Green Energy and Healthcare Technologies (IGEHT). IEEE, 2017: 1-4.
3. Etaiwi W, Awajan A. The effects of features selection methods on spam review detection performance[C]//2017 International Conference on New Trends in Computing Sciences (ICTCS). IEEE, 2017: 116-120.
4. S. Jia, X. Zhang, X. Wang and Y. Liu, "Fake reviews detection based on LDA," 2018 4th International Conference on Information Management (ICIM), Oxford, 2018, pp. 280-283.
5. G. Xu, M. Hu, C. Ma and M. Daneshmand, "GSCPM: CPM-Based Group Spamming Detection in Online Product Reviews," ICC 2019 - 2019 IEEE International Conference on Communications (ICC), Shanghai, China, 2019, pp. 1-6.
6. M. Li, B. Wu and Y. Wang, "Comment Spam Detection via Effective Features Combination," ICC 2019 - 2019 IEEE International Conference on Communications (ICC), Shanghai, China, 2019, pp. 1-6

7. N. A. Patel and R. Patel, "A Survey on Fake Review Detection using Machine Learning Techniques," 2018 4th International Conference on Computing Communication and Automation (ICCCA), Greater Noida, India, 2018, pp. 1-6.
8. J. K. Rout, A. K. Dash and N. K. Ray, "A Framework for Fake Review Detection: Issues and Challenges," 2018 International Conference on Information Technology (ICIT), Bhubaneswar, India, 2018, pp. 7-10.
9. J. Li, "Identification Model of Commodity False Reviews Based on Integrated Features," 2018 International Conference on Virtual Reality and Intelligent Systems (ICVRIS), Changsha, 2018, pp. 395-398.
10. W. Liu, J. He, S. Han, F. Cai, Z. Yang and N. Zhu, "A Method for the Detection of Fake Reviews Based on Temporal Features of Reviews and Comments," in IEEE Engineering Management Review
11. J. Li, "Identification Model of Commodity False Reviews Based on Integrated Features," 2018 International Conference on Virtual Reality and Intelligent Systems (ICVRIS), Changsha, 2018, pp. 395-398.
12. C. He and Y. Shi, "Research on Chinese Spam Comments Detection Based on Chinese Characteristics," 2018 IEEE 4th International Conference on Computer and Communications (ICCC), Chengdu, China, 2018, pp. 2608-2612.
13. Cheng H T , Koc L , Harmsen J , et al. Wide & Deep Learning for Recommender Systems[J]. 2016.
14. Y. Ren and D. Ji, "Learning to Detect Deceptive Opinion Spam: A Survey," in IEEE Access, vol. 7, pp. 42934-42945, 2019.
15. A. Bitarafan and C. Dadkhah, "SPGD_HIN: Spammer Group Detection based on Heterogeneous Information Network," 2019 5th International Conference on Web Research (ICWR), Tehran, Iran, 2019, pp. 228-233
16. Li J, Ott M, Cardie C, et al. Towards a general rule for identifying deceptive opinion spam[C]//Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers). 2014: 1566-1576.
17. W. Feng, W. Huang and W. Bao, "Imbalanced Hyperspectral Image Classification With an Adaptive Ensemble Method Based on SMOTE and Rotation Forest With Differentiated Sampling Rates," in IEEE Geoscience and Remote Sensing Letters.
18. R. K. Jeevan, S. Venu Madhava Rao, P. Shiva Kumar and M. Srivikas, "EEG-based emotion recognition using LSTM-RNN machine learning algorithm," 2019 1st International Conference on Innovations in Information and Communication Technology (ICIICT), CHENNAI, India, 2019, pp. 1-4.