Cloud recruitment false information detection method based on Entity bias and BERT-BiLSTM

[1st] Peiwen Gao[a*], [2st] Liang Zhang[b]

Hangzhou Normal University, Institute of Information Science and Technology, Hangzhou, China

[a*]stevenhznu@163.com, [b]zhl@hznu.edu.cn

Abstract. The Internet has swept the world, and the way of job hunting has undergone earth-shaking changes. Cloud recruitment has become the mainstream of the times. However, with the development of cloud recruitment, non-hair elements publish false recruitment information online, which induces job seekers to be deceived and their money is empty. Monitoring false recruitment information is helpful for people to identify false recruitment in advance and cut off the conditions for fraud at the source. One of the difficulties in monitoring false recruitment information is feature extraction, and extracting effective features from recruitment information is the focus of subsequent detection. The other difficulty is that the time effect of high-frequency information entities makes the model lack generalization ability. Therefore, we propose a BERT-BiLSTM model without entity deviation, which can effectively improve the detection ability and generalization ability of the model while fully extracting information context features. The experimental results show that our model has reached 99.02% accuracy and F1 score of 0.92.

Keywords: cloud recruitment, false recruitment detection, entity bias, BERT, Bi-LSTM.

1 Introduction

In recent years,online recruitment has won the favor of more and more job seekers with faster recruitment efficiency and better job-seeking experience. According to the Research Report on Internet Recruitment Industry in China in 2022 released by research[1], the market size of Internet recruitment industry in China reached 19.86 billion yuan in 2021, and online recruitment accounted for 85.1%. However, due to factors such as asymmetric information and inadequate supervision in online recruitment, criminals publish false information, steal job seekers' private information and defraud job seekers' money [2], which seriously intrudes the normal job-seeking and employment environment. According to the Report on Consumer Rights Protection of Internet Users in China in 2017 [3], in 2016, the amount of false part-time reports accounted for one-fifth of the total amount of Internet users' rights protection, most of which occurred

C. Chen et al. (eds.), *Proceedings of the 3rd International Conference on Digital Economy and Computer Application (DECA 2023)*, Atlantis Highlights in Computer Sciences 17,
https://doi.org/10.2991/978-94-6463-304-7_56

on well-known recruitment platforms. Therefore, how to effectively identify the false information in cloud recruitment has become the focus of academic attention.

This paper proposes a BERT-BiLSTM model based on Entity Deflection Framework (ENDEF)[6], The main contributions of this paper are as follows:

1.In the aspect of feature extraction, this paper applies BERT model in the field of false recruitment information detection to extract text feature information more effectively.

2.In the aspect of removing bias, this paper introduces the Entity biasframework into the field of false recruitment detection for the first time, and improves the framework to adapt to the research field of this paper.

3.In the aspect of data set, ADASYN[9] is used to balance EMSCAD data set, and the model is built on the balanced and unbalanced data set.

2 Literature review

Based on the perspective of machine learning, in 2017, Vidros et al. [4] used random forest classifiers to predict false information on recruitment platforms, with an accuracy rate of 91%;

Based on the perspective of deep learning, Anita et al[5] established a classification model based on word2vec and LSTM network to detect this problem.

There are three problems in the above research: (1)word2vec, as a word vector tool, has its limitations in extracting text features.(2) Although deep learning has a stronger generalization ability than machine learning, it does not solve the entity deviation of data sets [6].(3) The researchers of all the above studies used the open EMSACD[9], but they didn't consider the extreme data imbalance in this data set. To sum up, taking full advantage of text features and considering the influence of entity bias in the data set, we put forward a BERT-BiLSTM/ENDEF model to eliminate entity bias, and use ADASYN technology to improve the data set in order to detect the problem of false information in cloud recruitment.

3 Way

3.1 BERT-BiLSTM model

The model structure is shown in Figure 1.

Input layer: this layer receives text features from recruitment information.

Embedding layer: the content input by the input layer is transformed into a low latitude vector by BERT[7].

Feature extraction layer: the low latitude vector is extracted by Bi-LSTM neural network.

Output layer: The last layer of the model is a layer with a single neuron and Sigmoid activation function, which outputs a value between 0 and 1, indicating the probability that the information is false.

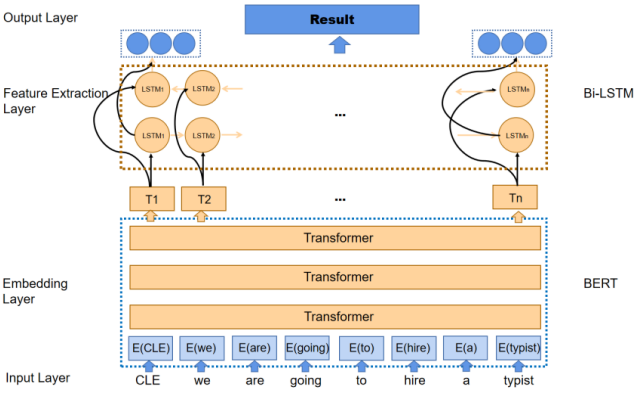

Fig. 1. BERT-BILSTM Model

3.2 ENDEF framework

ENDEF framework is an Entity bias framework proposed by Zhu et al. [6], and it is applied in the field of fake news detection. Entities will have an impact on the content and forecast results of recruitment. We need to establish the causal relationship among entities, recruitment information and recruitment information authenticity labels, model the impact of entities and recruitment information on the authenticity of recruitment information respectively in the training stage, and directly remove the part based on entity forecasting the authenticity of recruitment information in the testing stage to remove the entity deviation

ENDEF framework architecture is shown in figure 2.

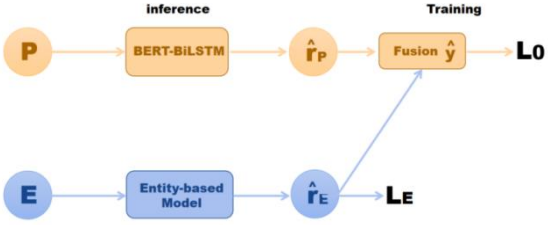

Fig. 2. ENDEF Frame

The calculation formula of model processing is as follows

Use cross entropy loss function:

$$L_O = \sum_{(P,y)\in D} -y\,log(\ y\ ) - (1-y)\,log(1-\ y\ ) \tag{1}$$

Set an auxiliary loss function for the entity separately:

$$L_E = \sum_{(P,y)\in D} -y\,log(\sigma(\hat{r}_E) - (1-y)\,log(1-\sigma(\hat{r}_E) \tag{2}$$

βIs a super parameter, set to 0.2, and the whole process training loss function is:

$$L = L_O + \beta L_E \tag{3}$$

In the verification stage, we can think that only using it is equivalent to removing the influence of entity deviation to the maximum extent.$\hat{r}_P$

4 Experiments

4.1 Data set and data processing

4.1.1 Data set

The benchmark data set of this paper is the "Real/Fake Job Posting Prediction" data set published on kaggle [8], which contains 17014 real job information and 866 false job information, including 16 independent features.

4.1.2 Data balance.

In this paper, the sample information of false positions in the data set is less, which will lead to the poor recognition effect of classifiers on these categories, and the experiment will be over-fitted. ADASYN[9] can increase the number of samples in these categories by generating synthetic samples, thus improving the comprehensive performance of the model. The balanced data set contains 17,014 real job information and 16,895 false job information.

4.1.3 Pretreatment

Firstly, we analyzed the data features and counted the number of all features in the data set, and found that a large number of features were missing among all features.It can be found that the features of department, salary_range, required_experience and required_education are the most missing in the data set, so we removed the first five features with the most serious missing and kept 12 features,and then merge all features into one feature to facilitate subsequent entity extraction.

Finally, we use Text-Smart technology to detect the entities in the data set. Text-Smart[6] can detects the entities in the text. We add the entity list detected by text-smart as a feature to the cleaned data set. We divide the reconstructed data set into training set and test set according to the ratio of 7:3. So far, the data set has been processed as shown in Table 1.

Table 1. The data set after extracting the entity

Fraudulent	Text	EntityList
0	marketing intern marketing we're...	[itern,food...]
1	customer service cloud video...	[engneer,punch...]
...	...	...

4.2 Experimental Settings

The experimental environment is 64-bit Windows10, the development platform is jupyter notebook, the node of BiLSTM neural network is set to 128, the single-layer neural network, the loss function uses binary cross entropy function, the adam gradient descent optimizer is used, the learning rate is set to 0.001, and epochs is set to 25.

4.3 Experimental indicators

We evaluate our model by accuracy, precision, recall and F1-score.

4.4 Experimental results

In order to verify the validity of the model, we model the data set before and after the balance, and set up the control experimental group.

Table 2. Experimental results of unbalanced data sets

Model	Accuracy	Precision	Recall	F1score
W2vec-LSTM	97.04%	0.85	0.40	0.57
W2vec-BiLSTM	98.71%	0.88	0.62	0.70
BERT-Bi-LSTM	98.87%	0.89	0.68	0.75
BERT-BiLSTM/ENDEF	99.02%	0.91	0.70	0.78

Table 3. Experimental results of balanced data sets

Model	Accuracy	Precision	Recall	F1 score
W2vec-LSTM	97.42%	0.85	0.84	0.83
W2vec-BiLSTM	98.75%	0.88	0.84	0.85
BERT-Bi-LSTM	98.89%	0.89	0.88	0.87
BERT-BiLSTM/ENDEF	99.05%	0.92	0.92	0.93

According to Table 2 and Table 3, we can draw the following conclusions:

With unbalanced data sets:

(1)Using the same word vector W2vec, the detection accuracy of BiLSTM neural network is improved by 1.67% compared with LSTM neural network, which shows that BiLSTM is better than LSTM neural network in extracting text features.

(2)Using the same neural network BiLSTM, the BERT pre-training model can extract the semantic information of the text better than W2Vec, and its detection accuracy is improved by 0.18%.

(3)Using the same pre-training model and the same neural network, the accuracy of using the Entity bias framework ENDEF is improved by 0.15% compared with the model without this framework, showing better prediction ability.

(4)There is over-fitting in the experiment, which shows that the F1 scores of the four experiments are not good enough.

With a balanced data set:

(1)The F1 scores of all the experiments were improved.

(2)Compared with the other three models, the model proposed in this experiment performs best.

5 Conclusion

Aiming at the problem of false information in cloud recruitment, this paper proposes a BERT-BiLSTM model of entity bias. Experiments show that the accuracy, recall and F1 value of our model reach 99.02%, 0.91, 0.70 and 0.78 respectively, which is superior to other advanced models for detecting this problem.

This model has the following shortcomings: (1) It only detects the text features, and has not considered the multi-modal data set; (2) At present, there is only one public data set for this problem, so it is impossible to effectively detect its generalization; In view of the above points, the next step is to deal with the imbalance of data sets by collecting multi-dimensional data sets ourselves. And improve the related model to overcome the above shortcomings.

References

1. Iresearch. Market Development Research Report of China's Online Recruitment Industry in 2022 [R]. Iresearch, 2022.
2. Joyce S (2021) 5 major types of scam jobs and job scams online— Job-Hunt.org . Job Hunt. https://www.job-hunt.org/job-search-scams/
3. Wang Chunge. Analysis of false Information in Online Recruitment [J]. Modern Marketing: Management Edition, 2020(11):2.
4. Vidros S, Kolias C , Kambourakis G .Online recruitment services: another playground for fraudsters[J].Computer Fraud&Security,2016, 2016(3):8-13.DOI:10.1016/S1361-3723(16)30025-2.
5. Anita, C. Schürch et al. “Fake Job Detection and Analysis Using Machine Learning and Deep Learning Algorithms.” (2021).
6. Zhu, Yongchun et al. “Generalizing to the Future: Mitigating Entity Bias in Fake News Detection.” Proceedings of
7. Devlin J, Chang M W , Lee K ,et al.BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding[J]. 2018.DOI:10.48550/arXiv.1810.04805.
8. Kambourakis G (2017) Employment scam Aegean dataset. http://emscad.samos.aegean.g. Accessed 29 Aug 2022
9. He H, Bai Y , Garcia E, Li S (2008) ADASYN: adaptive synthetic sampling approach for imbalanced learning, pp 1322–1328. https://doi.org/10.1109/IJCNN.2008.4633969