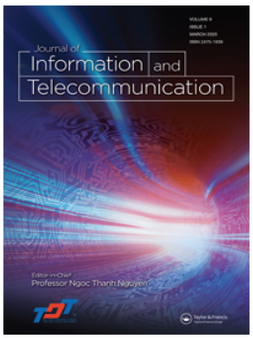

Taylor & Francis
Taylor & Francis Group

Journal of Information and Telecommunication

ISSN: 2475-1839 (Print) 2475-1847 (Online) Journal homepage: www.tandfonline.com/journals/tjit20

Improving fake job description detection using deep learning-based NLP techniques

Dinh-Hong Vu, Kien Nguyen, Khai Thien Tran, Bay Vo & Tuong Le

To cite this article: Dinh-Hong Vu, Kien Nguyen, Khai Thien Tran, Bay Vo & Tuong Le (2025) Improving fake job description detection using deep learning-based NLP techniques, Journal of Information and Telecommunication, 9:1, 113-125, DOI: 10.1080/24751839.2024.2387380

To link to this article: https://doi.org/10.1080/24751839.2024.2387380

Published online: 07 Aug 2024.

JOURNAL OF INFORMATION AND TELECOMMUNICATION
2025, VOL. 9, NO. 1, 113–125
https://doi.org/10.1080/24751839.2024.2387380

Improving fake job description detection using deep learning-based NLP techniques

Dinh-Hong Vu[a], Kien Nguyen[b], Khai Thien Tran[c], Bay Vo[b] and Tuong Le[b]

[a]Natural Language Processing and Knowledge Discovery Laboratory, Faculty of Information Technology, Ton Duc Thang University, Ho Chi Minh City, Vietnam; [b]Faculty of Information Technology, HUTECH University, Ho Chi Minh City, Vietnam; [c]Ho Chi Minh City University of Foreign Languages and Information Technology, Ho Chi Minh City, Vietnam

ABSTRACT

In the modern online society, where social networks are increasingly developed, the authenticity of information is one of the essential needs. There are many problems with detecting fake information, such as fake news detection, fake review detection, etc. Fake job description (FJD) detection is an interesting problem many groups have studied recently. However, current studies still need improvement in predictability. Therefore, this study develops a new method named NLP2FJD that utilizes deep learning-based NLP techniques for improving FJD detection. Firstly, we utilize the pre-trained Word2Vec to extract features from textual information from the dataset. Next, combining textual information and meta-information in the experimental dataset to improve the performance of FJD detection. The above two improvements will help the recommender system significantly improve the predictive ability of the proposed model. Finally, the empirical experiments are conducted to confirm the effectiveness of the proposed method on the experimental dataset compared with cutting-edge methods. The experimental results demonstrate that the NLP2FJD framework transcends other experimental methods for FJD detection on the experimental dataset. Besides, this study also conducts ROC curve analysis to show how to determine the optimal threshold for distinguishing the fake or real job description on the experimental dataset.

ARTICLE HISTORY
Received 4 March 2024
Accepted 28 July 2024

KEYWORDS
Fake job description detection; natural language processing; word embeddings; deep learning

1. Introduction

Nowadays, with the development of computer systems, machine learning is prevalent in all areas of life, from medical diagnosis (Vo et al., 2022a), economics (Le, 2022), engineering (Doan et al., 2021; Nti et al., 2022), etc. In the medical sector, intelligent systems use machine learning to assist medical staff with medical diagnosis. This saves a lot of time in the medical examination and treatment process, increasing the number of patients served. For instance, a computer vision-based framework (Vo et al., 2022a) is developed

CONTACT Tuong Le lc.tuong@hutech.edu.vn Faculty of Information Technology, HUTECH University, Ho Chi Minh City, Vietnam

to detect the kind of shoulder implant via X-ray image classification to support the treatment process. Besides, machine learning helps investors and managers in the economic sector predict several economic issues, such as the possibility of the company going bankrupt (Le, 2022), future oil prices (Vo et al., 2020), the future electricity consumption of customers (Le et al., 2019; Le et al., 2020), and monthly household water consumption (Vo et al., 2022b). Meanwhile, it also contributes a lot to the development of engineering (Nti et al., 2022), such as construction (Doan et al., 2021), robotics (Vo et al., 2019), etc. Through the aforementioned applications, the significance of artificial intelligence (AI) in today's landscape becomes evident.

Natural language processing (NLP), a subfield of artificial intelligence, allows computer programmes to analyze and comprehend human language. There are a lot of applications of deep learning in this field, such as machine translation (Nguyen et al., 2021a; Nguyen et al., 2021b; Rivera-Trigueros, 2022; Tran et al., 2022a), sentiment analysis (Birjali et al., 2021; Tran et al., 2022b; Tran & Phan, 2020), rumour detection (Nguyen et al., 2021c; Nguyen et al., 2022b), text classification (Abayomi-Alli et al., 2022; Minaee et al., 2021; Škrlj et al., 2021), etc. Nowadays, there is much fake information or rumour in a variety of information platforms, including social networking, blogs, as well as online newspapers. Therefore, it's difficult to confirm the trustworthiness of the information. This leads to inappropriate behaviour by the receiver of the information. The detection of fake messages (Jiang et al., 2021; Nguyen et al., 2022a; Raj & Meel, 2021; Sahoo & Gupta, 2021), a kind of text classification, has received a lot of attention recently. Fake news detection is a form of text classification, where algorithms learn to categorize news articles, social media posts, or other textual content into binary categories, typically 'fake' or 'real.' To achieve this, various features and representations of text are employed, including bag-of-words models, word embeddings, syntactic and semantic analysis, and contextual information.

Recently, the problem of FJD detection represents a specific case within the broader context of fake message detection. Firstly, (Vidros et al., 2017) introduced the Employment Scam Aegean (ESA) dataset in 2017. Their study found that logistic regression emerged as the most accurate approach for detecting FJD. Next, (Vo et al., 2021) developed an efficacious framework named FJD-OT (Fake Job Description Detection Using Oversampling Techniques) for dealing with the class imbalance problem occurring on the ESA dataset. The results in (Vo et al., 2021) confirmed that the FJD-OT model significantly improves in terms of the predictability of FJD detection on the ESA dataset in terms of several famous performance metrics in an imbalanced data scenario. Specifically, this method achieved 0.96 in AUC (area under the ROC curve). Although the results of the above study were relatively high, this method did not take advantage of the latest natural language processing techniques to get the best results. Hence, this study introduces the NLP2FJD method, which leverages deep learning-based NLP techniques to enhance FJD detection. The primary contributions of this research are outlined as follows:

- Utilize the pre-trained Word2Vec model from the Google News dataset to extract features from textual information in the ESA dataset.
- Combine textual information and meta-information from the ESA dataset within the NLP2FJD framework to enhance the performance of fake job description (FJD) detection.

- Conduct experiments to demonstrate the efficacy of the NLP2FJD framework on the ESA dataset, comparing it with state-of-the-art methods.
- Perform threshold analysis to determine the optimal threshold for distinguishing between fake and real job descriptions in the ESA dataset.

The structure of this study unfolds across four distinct sections. Section 2 delves into related works concerning various challenges in NLP, including text classification, fake news detection, and FJD detection. Section 3 introduces the proposed framework, leveraging deep learning-based NLP techniques to enhance FJD detection performance. The experimental findings are detailed in Section 4. Finally, the study culminates in a synthesis of results and outlines potential avenues for future research in the concluding section.

2. Related works

Several problems in NLP have been studied in recent times, such as machine translation (Nguyen et al., 2021a; Nguyen et al., 2021b; Rivera-Trigueros, 2022; Tran et al., 2022a), sentiment analysis (Birjali et al., 2021; Tran et al., 2022b; Tran & Phan, 2020), rumour detection (Nguyen et al., 2021c; Nguyen et al., 2022b), text classification (Abayomi-Alli et al., 2022; Minaee et al., 2021; Škrlj et al., 2021), etc. In this section, we summarize some research on text classification. Then we will review some studies on detecting fake information, including fake news detection, fake review detection, and FJD detection.

Text classification is the most interesting problem. In 2021, (Škrlj et al., 2021) developed tax2vec for short text classification, which is a parallel model for forming taxonomy-based features. In the proposed model, semantic information available in taxonomies is utilized to create semantic features to enhance the performance of the problem of short-text classification. In the same year, (Luo, 2021) used a machine learning model named support vector machines (SVM) for classifying text and documents in English. The experiment was conducted on an experimental dataset with 1033 documents. The experimental results confirmed that SVM outperforms the other classifiers. Next, (Prabhakar & Won, 2021) utilized two deep learning architectures, including a quad-channel hybrid long short-term memory (LSTM) deep learning model and a hybrid bidirectional gated recurrent unit (BiGRU) model with multi-head attention for medical text classification.

Fake news detection (Jiang et al., 2021; Nguyen et al., 2022a; Raj & Meel, 2021; Sahoo & Gupta, 2021) is a specific problem of text classification problems. There are two classes, including fake of real news, for distinguishing. (Jiang et al., 2021) used the stacking approach that combined results of Logistic Regression (LR), SVM, k-Nearest Neighbour (k-NN), Decision Tree (DT), Random Forest (RF), Convolutional Neural Network (CNN), gated recurrent network (GRU), LSTM for improving the performance of the problem of fake news detection. Next, (Sahoo & Gupta, 2021) developed a framework for fake news detection utilizing machine learning and deep learning classifiers. This framework analyzes both user profiles and news content features in the Facebook platform. The proposed method based on LSTM obtained excellent performance at 99.4% accuracy in detecting fake news in real-time. Meanwhile, (Raj & Meel, 2021) introduced a multimodal Coupled ConvNet architecture for efficient fake news detection. This approach combines two data modules and efficiently classifies online news based on its textual and visual contents. Recently, (Nguyen et al., 2022a) developed a deep multi-domain multimodal

model for fake news detection in the Vietnamese language. This approach was verified on a real-life dataset in Vietnamese and achieved impressive results.

In 2017, (Vidros et al., 2017) introduced the ESA dataset, which comprises 18,746 job descriptions collected from 2012 to 2014. This dataset includes 17,880 real job descriptions and 866 fake ones. In their study (Vidros et al., 2017), various machine learning models were employed for fake job description (FJD) detection, with logistic regression achieving the highest accuracy for this task. Later, (Vo et al., 2021) introduced a machine learning model incorporating an oversampling technique called the FJD-OT framework, designed to enhance the detection of fake job descriptions (FJDs) in the imbalanced environment of the ESA dataset. While (Vidros et al., 2017) and (Vo et al., 2021) demonstrated the effectiveness of logistic regression and FJD-OT framework for detecting fake job descriptions (FJDs), there is a need to explore the potential of more advanced machine learning techniques, such as deep learning methods, to further enhance detection accuracy. Therefore, this study develops an enhanced model for Fake Job Description Detection using advanced deep learning-based NLP techniques.

3. NLP2FJD: an improving model using deep learning-BASED NLP techniques for fake job description detection

3.1. ESA dataset

The Laboratory of Information and Communication Systems Security (ICSS) of the University of the Aegean, Greece, introduced the ESA dataset (Vidros et al., 2017). This dataset has 17,014 real job descriptions and 866 fake ones, introduced on the Internet from 2012 to 2014 in English. Two pieces of information, including textual information and meta-information related to the jobs, are included in the dataset. ***Textual information*** includes 'title', 'company profile', 'description', 'requirements', and 'benefits' fields to describe the job. ***Meta-information*** includes 'telecommuting', 'has company logo', 'has questions', 'employment type', 'required experience', 'required education', 'industry', and 'function' fields to describe more information about the company and the job.

In (Vo et al., 2021), the authors focus solely on utilizing textual information for detecting fake jobs. However, the dataset includes valuable meta-information that can significantly enhance the accuracy of fake job detection. Furthermore, the approach in (Vo et al., 2021) relies solely on Bag-of-Words (BoW), which neglects the intricate relationships between words. Incorporating these semantic relationships could greatly improve the effectiveness of detecting fake job postings. Therefore, in this study, we integrate both textual and meta-information. Moreover, we leverage word embeddings and deep neural networks to capture and utilize the semantic relationships between words in the text. With ***textual information,*** we concatenate all its fields to get one long text, then do some preprocessing like Stopwords removal and tokenizing. After that, we get a word vector to present the word and its relations with other words by Embedding layer and extracting textual information by CNN network. We also get BoW vectors like in (Vo et al., 2021) to combine with other features. With ***meta-information***, we vectorize them using the one-hot technique to create a 204-dimensional meta vector. Exploiting both textual information and meta-information will significantly improve the predictability of FJD detection.

3.2. The overall architecture

In this section, we present the comprehensive flowchart of the NLP2FJD framework, illustrated in Figure 1. The framework consists of four essential modules: text feature extraction, global feature extraction, meta feature extraction, and output modules. Each module is detailed in the following subsections to provide a thorough understanding of our proposed approach.

3.3. Text feature extraction module

This module is used to extract features from ***textual information*** to use for the output module. Cleaned text goes through the word ***embedding layer*** to get word vectors. Word embedding is one of NLP's most famous representations of document vocabulary. This technique aims to transform the individual word into a vector. This vector is then learned in a way that resembles a neural network. The vectors try to capture various characteristics of that word with regard to the overall text. These characteristics can include the semantic relationship of the word, definitions, context, etc. These numerical representations of word embedding methodology enable this model to do many tasks, like identifying similarities or dissimilarities between words. Currently, Word2Vec (Mikolov et al., 2013), a state-of-the-art type of word embedding methodology by Tomas Mikolov et al., has been widely used because of the high results it brings. This approach employs an external neural network for learning word embeddings, which can be achieved through two methods: Skip-gram and Continuous Bag of Words (CBOW).

In the CBOW architecture, the model predicts the current word based on a window of surrounding context words (neighbouring words). Meanwhile, the continuous skip-gram architecture uses the current word to predict the surrounding window of context words.

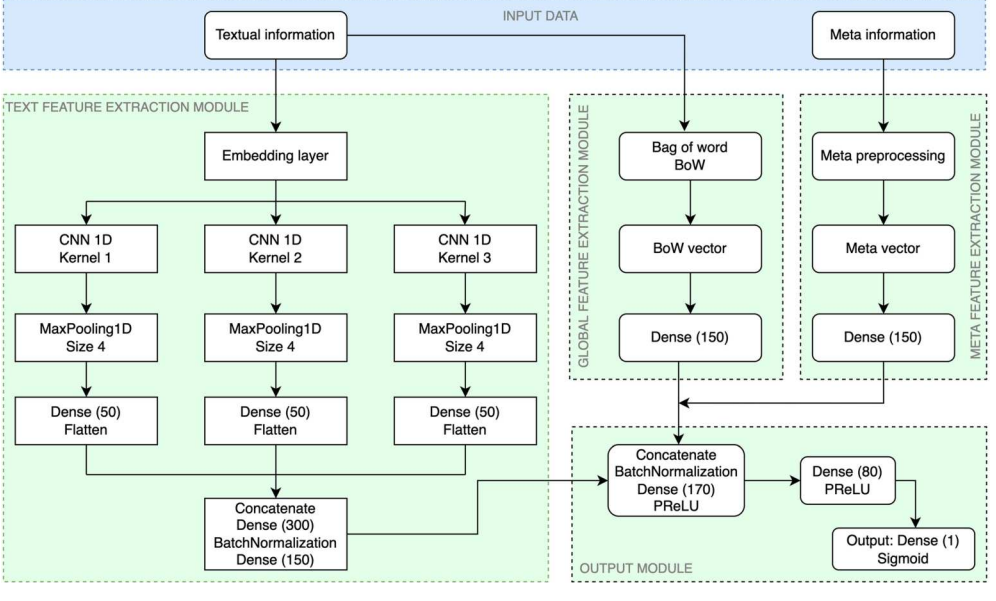

Figure 1. The NLP2FJD framework for fake job description detection.

This approach assigns greater weight to nearby context words compared to those further away. This study utilized Word2Vec (Mikolov et al., 2013) with pre-trained vectors trained on the Google News dataset, which consists of approximately 100 billion words. This pre-trained model contains 300-dimensional vectors representing 3 million words and phrases. The phrases were generated using a straightforward data-driven approach detailed in (Mikolov et al., 2013). Therefore, for each job's text, we obtain a matrix of size *N*x300, where *N* represents the maximum length of text among all jobs.

Once the embedding matrix is obtained, it undergoes processing through three branches of 1D Convolutional Neural Networks (CNNs) equipped with kernels of sizes 1, 2, and 3, respectively. These specialized CNN branches meticulously examine the text, capturing intricate word relations across three vital levels: 1-gram, 2-gram, and 3-gram. These nuanced word associations play a pivotal role in uncovering indicators of fraudulent job postings. The outputs of three CNNs are flattened, concatenated, and passed through two Dense layers to generate feature vectors for the text. These vectors constitute the module's output.

3.4. Global feature extraction module

This module uses to capture global information of ***textual information*** for fake job detection. We call it global because the input of this module is Bag-of-Words (BoW) vectors, which contain information about the word and the number of repetitions of the word. These vectors do not have information about the relationship between words, but they give an overview of a text. BoW vector is created through the following three steps. Firstly, the model is to convert the sentences in our corpus into tokens or individual words. The second step creates a dictionary that contains all the words in our corpus as keys and the frequency of the occurrence of the words as values. Step three is to create the Bag of Words Model by creating a matrix. In this matrix, the columns correspond to the most frequent words in our dictionary, while the rows correspond to the document or sentences.

The cleaned text undergoes the BoW module to generate a text vector. Each job's text yields a unique BoW vector. In our experiment, each vector spans 173,157 dimensions. These vectors subsequently pass through a Dense layer designed to distil and preserve key features. The resultant output from this Dense layer serves as the final output of the module.

3.5. Meta feature extraction module

The module described is focused on extracting ***meta-information*** features from job postings, which can be crucial for tasks such as fake job detection.

Meta-information is passed through the pre-processing layer to produce a 204-dimensional vector for each job by one-hot technique as follows: (1) With fields 'telecommuting', 'has company logo' and 'has questions': their values are 0 and 1, for each field we convert to a two-dimensional one-hot vector. (2) With fields 'employment type', 'required experience', 'required education', 'industry', and 'function': they are category forms where 'employment type' has six categories, 'required experience' has eight categories, 'required education' has 14 categories, 'industry' has 132 categories, and 'function' has 32

categories. Each field turns into a one-hot form with dimensions equal to the number of categories. After converting these fields into one-hot encoded vectors, we concatenate them to form a single 204-dimensional meta-vector for each job.

Furthermore, we apply a Dense layer to distil and preserve the essential features within the meta vectors. The output of this Dense layer serves as the module's final output.

3.6. Output module

This pivotal module integrates insights gleaned from the preceding three modules to render a definitive verdict on the authenticity of a given job posting. The outputs from the text feature extraction, global feature extraction, and meta feature extraction modules are concatenated and subsequently processed through two Dense layers, each followed by a Parametric Rectified Linear Unit (PReLU) activation function. Finally, the output layer employs a Sigmoid function to determine whether the job posting under scrutiny is genuine or fraudulent.

$$\phi(z) = \frac{1}{1 + e^{-z}} \tag{1}$$

where $\phi(z)$ is the output of the proposed method. This value is from zero to one (probability estimate). Meanwhile, z is the input value.

4. Experiments

4.1. Experiment setting

The study's experiments are implemented in Python programming language, utilizing version 3.7. The operating system is CentOS Linux 7. The computational infrastructure comprises a robust system equipped with an Intel(R) Xeon(R) CPU E5-2698 v4 @ 2.20 GHz (40 cores) and 125GB of RAM. To facilitate various essential functionalities, the imbalanced-learn (Lemaitre et al., 2017) and scikit-learn (Pedregosa et al., 2011) packages are installed.

Five common metrics, including Accuracy (ACC), Geometric mean (G-mean), Recall, Sensitivity, and Specificity, are used to evaluate classifiers. These metrics are calculated as follows:

$$ACC = \frac{TP + TN}{TP + TN + FP + FN} \tag{2}$$

$$Sensitivity = \frac{TP}{TP + FN} \tag{3}$$

$$Specificity = \frac{TN}{TN + FP} \tag{4}$$

$$G - mean = \sqrt{Sensitivity \times Specificity} \tag{5}$$

$$Recall = \frac{TP}{TP + FN} \tag{6}$$

Besides, the receiver operating characteristic (ROC) (Fawcett, 2006) curve is also presented due to the class imbalance problem. The ROC curve can be utilized to visualize the trade-off between the true positive rate and the false positive rate. On this curve, each point indicates the performance of a model for a specific threshold. In the above scenario, the area under the ROC curve (AUC) of different models is compared with each other. In this metric, the higher the values are the better the models' performances.

The experimental methods include two state-of-the-art models for the detection of FJDs on the experimental dataset ESA such as logistic regression (LR) (Vidros et al., 2017) and Fake Job Description Detection Using Oversampling Techniques (FJD-OT) (Vo et al., 2021). Next, this study also uses two famous machine-learning models, including Random Forest and AdaBoost, combined with BoW for comparison (denoted by BoW-RF and BoW-AB, respectively). Then, several under-sampling techniques such as Repeated Edited Nearest Neighbours (RepENN), Tomek Links (TLink), Instance Hardness Threshold (IHThres), Near Miss (NMiss), and Neighbourhood Cleaning Rule (NCRule) combined with logistic regression are also used to deal with imbalanced data which are denoted by U_RepENN-LR, U_TLink-LR, U_IHThres-LR, U_NMiss-LR, and U_NCRule-LR respectively.

4.2. Experimental results

Firstly, we evaluated the performance of our NLP2FJD model using two types of word embeddings: BERT (Devlin et al., 2019) and Word2Vec, as detailed in Table 1.

The results indicate that while BERT achieves slightly higher accuracy and specificity, Word2Vec outperforms BERT in critical metrics such as recall, G-Mean, sensitivity, and AUC. Notably, Word2Vec demonstrates superior performance in handling imbalanced data, which is crucial for our NLP2FJD model's effectiveness. One significant factor contributing to BERT's lower performance is its token limitation of 512, which leads to truncation of longer texts and potential loss of context. In contrast, Word2Vec, being less constrained by token limits, provides more comprehensive embeddings that capture nuanced semantic relationships in texts. In conclusion, our experiments underscore the suitability of Word2Vec embeddings for enhancing the performance of the NLP2FJD model, particularly in scenarios involving complex or lengthy textual data. These findings guide our choice of embedding technique, emphasizing practical effectiveness in real-world natural language processing tasks.

Secondly, the study compares the ROC curves of several experimental algorithms, including LR (Vidros et al., 2017), FJD-OT (Vo et al., 2021), BoW-RF, BoW-AB, and the proposed model (NLP2FJD). In this experiment, k-fold cross-validation is utilized with $k = 5$. The results are shown in Figure 2, which are the average values of 5 folds. It is easy to see the green ROC curve of the proposed algorithm (NLP2FJD) outside the ROC curves of other algorithms. Specifically, the AUC of NLP2FJD achieved 0.984, an impressive result. Meanwhile, the AUC of LR, FJD-OT, BoW-RF, and BoW-AB are 0.976, 0.965, 0.936, and 0.96, respectively.

Table 1. Experimental results of NLP2FJD model using BERT and Word2Vec embedding.

Model	ACC	Recall	G-Mean	Sensitivity	Specificity	AUC
BERT-NLP2FJD	0.98 ± 0.002	0.80 ± 0.033	0.89 ± 0.018	0.80 ± 0.033	0.99 ± 0.002	0.983 ± 0.004
Word2Vec-NLP2FJD	0.96 ± 0.011	0.93 ± 0.018	0.94 ± 0.011	0.93 ± 0.018	0.96 ± 0.012	0.984 ± 0.004

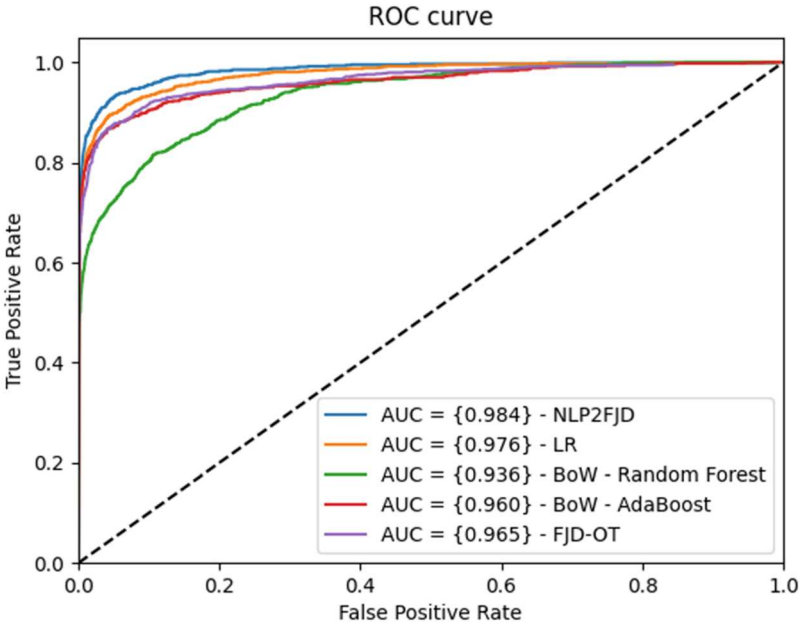

Figure 2. ROC curves of the experimental methods for the ESA dataset.

Thirdly, full experimental results in terms of Accuracy, G-mean, Recall, Sensitivity, Specificity, and AUC of all experimental methods for the ESA dataset are reported in Table 2. For the Accuracy metric, there are four methods, including LR, FJD-OT, BoW-AB, and NLP2FJD, obtaining accuracy rates more prominent than 0.9, while the others have accuracy rates lower than 0.9. The proposed method got the best accuracy score at 0.96, while LR and BoW-AB are in second place with accuracy at 0.94.

For the Recall metric, the proposed method (NLP2FJD) is the best performer, achieving a recall of 0.93 ± 0.018. This means it correctly identifies 93% of the actual positive instances in the dataset, with minimal variability across different runs. Logistic Regression (LR) is in second place with a recall of 0.92 ± 0.017, indicating it correctly identifies 92% of the positive instances with stable performance. These results demonstrate that the proposed method is highly effective in minimizing false negatives, which is crucial in applications where missing a positive instance can have significant consequences. Other

Table 2. Experimental results of the experimental method for the ESA dataset.

No	Model	ACC	Recall	G-Mean	Sensitivity	Specificity	AUC
1	LR (Vidros et al., 2017)	0.94 ± 0.020	0.92 ± 0.017	0.93 ± 0.014	0.92 ± 0.017	0.95 ± 0.021	0.976 ± 0.005
2	FJD-OT (Vo et al., 2021)	0.92 ± 0.019	0.89 ± 0.034	0.91 ± 0.010	0.89 ± 0.034	0.92 ± 0.021	0.965 ± 0.010
3	BoW-RF	0.86 ± 0.042	0.85 ± 0.047	0.86 ± 0.010	0.85 ± 0.047	0.86 ± 0.046	0.936 ± 0.007
4	BoW-AB	0.94 ± 0.014	0.88 ± 0.019	0.92 ± 0.014	0.88 ± 0.019	0.95 ± 0.014	0.960 ± 0.011
5	NLP2FJD	0.96 ± 0.011	0.93 ± 0.018	0.94 ± 0.011	0.93 ± 0.018	0.96 ± 0.012	0.984 ± 0.004
6	U_RepENN-LR	0.86 ± 0.029	0.84 ± 0.035	0.85 ± 0.006	0.84 ± 0.035	0.86 ± 0.030	0.93 ± 0.01
7	U_TLink-LR	0.86 ± 0.030	0.84 ± 0.035	0.85 ± 0.006	0.84 ± 0.034	0.86 ± 0.033	0.93 ± 0.01
8	U_ IHThres-LR	0.86 ± 0.032	0.84 ± 0.040	0.85 ± 0.004	0.84 ± 0.040	0.86 ± 0.035	0.92 ± 0.01
9	U_NMiss-LR	0.73 ± 0.045	0.62 ± 0.045	0.67 ± 0.011	0.62 ± 0.045	0.73 ± 0.049	0.73 ± 0.02
10	U_NCRule-LR	0.88 ± 0.027	0.83 ± 0.038	0.85 ± 0.007	0.83 ± 0.038	0.87 ± 0.030	0.93 ± 0.01

models, such as FJD-OT, BoW-RF, and BoW-AB, show lower recall values of 0.89, 0.85, and 0.88, respectively, indicating they are less effective at identifying all positive cases. The remaining models generally have even lower recall values, indicating a high rate of false negatives.

For the G-Mean metric, NLP2FJD achieves the best value compared to other methods, with a G-Mean score of 0.94 ± 0.011. G-Mean, or geometric mean, is a metric that balances sensitivity and specificity, providing an overall measure of the model's performance by combining the accuracy for both the positive and negative classes. A high G-Mean indicates that the model performs well in correctly identifying both positive and negative instances, minimizing both false positives and false negatives. The low standard deviation (± 0.011) associated with the proposed method's G-Mean score indicates that this high performance is consistent across different runs. In contrast, other methods achieve lower G-Mean scores, demonstrating their comparatively lower effectiveness in maintaining a balance between sensitivity and specificity.

For the Sensitivity metric, NLP2FJD also achieves the highest value compared to other methods, with a Sensitivity score of 0.93 ± 0.018. Sensitivity, also known as recall, measures the proportion of actual positive cases that are correctly identified by the model. A high Sensitivity score indicates that the model is effective at detecting positive instances, minimizing the number of false negatives. The proposed method's Sensitivity score of 0.93 means it accurately identifies 93% of the positive instances, making it highly reliable in scenarios where missing positive cases is costly or dangerous. Other methods, such as Logistic Regression (LR) and FJD-OT, have lower Sensitivity scores of 0.92 ± 0.017 and 0.89 ± 0.034, respectively, indicating they are less effective at identifying positive cases compared to the proposed method. This demonstrates the robustness of the NLP2FJD method in providing high recall and reliability in detecting true positives.

The proposed method stands out with a specificity of 0.96 ± 0.012, indicating its ability to accurately identify true negative instances in the dataset. This high specificity is crucial in applications like FJD detection, where minimizing false positives is essential for reliable predictions. Additionally, the method achieves an AUC of 0.984 ± 0.004, reflecting its strong discriminatory power in distinguishing between positive and negative instances. These metrics highlight the effectiveness of leveraging deep learning-based NLP techniques, demonstrating their capability to enhance predictive performance and reliability in tackling FJD detection challenges.

Finally, we conducted a ROC curve analysis to ascertain the optimal threshold on the ESA dataset. This analysis involved randomly partitioning the data into an experimental dataset using a 70:30 ratio. Specifically, 70% of the ESA dataset was allocated for training purposes, while the remaining 30% was set aside for validation. After training and prediction, the ROC curve of this experiment is obtained and displayed on the left side of Figure 3. Youden's J statistic (Youden, 1950) is defined as follows.

$$J = Sensitivity + Specificity - 1 \tag{7}$$

Youden's J statistic serves as a straightforward technique for identifying the optimal threshold, determined by selecting the threshold that yields the highest J statistic value, as defined in Equation (7). Employing this method, our study identified the optimal threshold at 0.04, visually depicted on the left side of Figure 3.

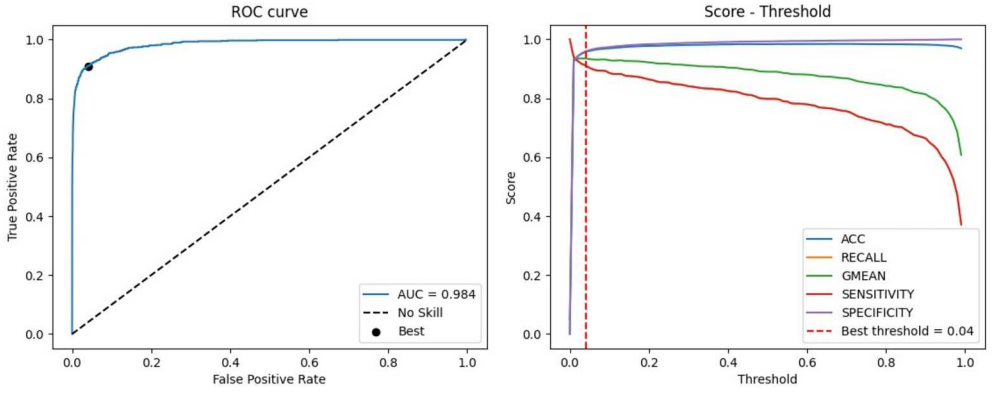

Figure 3. ROC curve analysis for determination of optimal threshold on the ESA dataset.

Later, the study displays a graph showing the values of ACC, Recall, G-Mean, Sensitivity, and Specificity corresponding to the variation of the threshold from 0 to 1, depicted on the right side of Figure 3. At the optimal threshold of 0.04, as determined by Youden's J statistic method, the model demonstrates exceptional performance, evidenced by the following metric values: Accuracy (0.96), Recall (0.91), G-Mean (0.94), Sensitivity (0.91), and Specificity (0.96). These figures underscore the robustness and efficacy of the proposed algorithm in effectively discriminating between genuine and fake job descriptions within the ESA dataset. Therefore, this optimal threshold (0.04) is recommended for use in the proposed algorithm for solving this task on the ESA dataset.

5. Conclusion

This study harnesses deep learning-based Natural Language Processing (NLP) techniques to introduce a potent methodology dubbed NLP2FJD, aimed at enhancing FJD detection within the ESA dataset. Initially, the pre-trained Word2Vec model sourced from the Google News dataset is leveraged to extract intricate features from textual data within the ESA dataset, a process conducted within the Text Feature Extraction Module. Subsequently, the Global Feature Extraction Module is deployed to encapsulate overarching textual information, thereby enriching the model's understanding of the dataset. Furthermore, this study incorporates meta-information from the ESA dataset through the Meta Feature Extraction Module. Subsequently, these extracted features are amalgamated within the final model, denoted as the Output Module. This module plays a pivotal role in determining the authenticity of each job description, distinguishing between genuine and fake postings based on the amalgamated features.

The conducted experiments serve to showcase the effectiveness of NLP2FJD on the ESA dataset in comparison to state-of-the-art methodologies. The findings unequivocally demonstrate that our proposed approach outperforms existing methods in FJD detection on the experimental dataset. Additionally, we conducted a ROC curve analysis to elucidate the process of determining the optimal threshold for the experimental dataset. The results of this analysis indicated that a cut-off threshold of 0.04 is recommended for utilization in our proposed model. This threshold serves as a critical benchmark for distinguishing

between genuine and fake job descriptions, thereby enhancing the accuracy and reliability of our model's predictions.

For future work, our focus will be on developing an online learning model tailored specifically for FJD detection, a necessity given the exponential growth of data in this domain. Additionally, we aim to broaden the scope of our application by collecting Vietnamese FJD data, thus enabling the extension of our methodology to encompass diverse languages. Moreover, our research agenda includes the exploration and proposal of machine learning models customized to address the intricacies of FJD detection within the Vietnamese language and other linguistic contexts.

Disclosure statement

No potential conflict of interest was reported by the author(s).

References

Abayomi-Alli, O., Misra, S., & Abayomi-Alli, A. (2022). A deep learning method for automatic SMS spam classification: Performance of learning algorithms on indigenous dataset. *Concurrency and Computation: Practice and Experience*, *e6989*.

Birjali, M., Kasri, M., & Beni-Hssane, A. (2021). A comprehensive survey on sentiment analysis: Approaches, challenges and trends. *Knowledge-Based Systems*, *226*, 107134.

Devlin, J., Chang, M. W., Lee, K., & Toutanova, K. (2019). Bert: Pre-training of deep bidirectional transformers for language understanding. *NAACL-HLT*, *1*, 4171–4186.

Doan, Q. H., Le, T., & Thai, D. K. (2021). Optimization strategies of neural networks for impact damage classification of RC panels in a small dataset. *Applied Soft Computing*, *102*, 107100.

Fawcett, T. (2006). An introduction to ROC analysis. *Pattern Recognition Letters*, *27*(8), 861–874.

Jiang, T. A. O., Li, J. P., Haq, A. U., Saboor, A., & Ali, A. (2021). A novel stacking approach for accurate detection of fake news. *IEEE Access*, *9*, 22626–22639.

Le, T. (2022). A comprehensive survey of imbalanced learning methods for bankruptcy prediction. *IET Communications*, *16*(5), 433–441.

Le, T., Vo, M. T., Kieu, T., Hwang, E., Rho, S., & Baik, S. W. (2020). Multiple electric energy consumption forecasting using a cluster-based strategy for transfer learning in smart building. *Sensors*, *20*(9), 2668.

Le, T., Vo, M. T., Vo, B., Hwang, E., Rho, S., & Baik, S. W. (2019). Improving electric energy consumption prediction using CNN and Bi-LSTM. *Applied Sciences*, *9*(20), 4237.

Lemaitre, G., Nogueira, F., & Aridas, C. K. (2017). Imbalanced-learn: A python toolbox to tackle the curse of imbalanced datasets in machine learning. *Journal of Machine Learning Research*, *18*, 17:1–17:5.

Luo, X. (2021). Efficient English text classification using selected machine learning techniques. *Alexandria Engineering Journal*, *60*(3), 3401–3409.

Mikolov, T., Sutskever, I., Chen, K., Corrado, G., & Dean, J. (2013). Distributed representations of words and phrases and their compositionality. *NIPS'13*, 3111–3119.

Minaee, S., Kalchbrenner, N., Cambria, E., Nikzad, N., Chenaghlu, M., & Gao, J. (2021). Deep learning-based text classification: A comprehensive review. *ACM Computing Surveys (CSUR)*, *54*(3), 1–40.

Nguyen, T. T., Nguyen, T. T., Nguyen, T. T., Vo, B., Jo, J., & Nguyen, Q. V. H. (2021c). Judo: Just-in-time rumour detection in streaming social platforms. *Information Sciences*, *570*, 70–93.

Nguyen, T., Nguyen, H., & Tran, P. (2021a). Sublemma-Based neural machine translation. *Complexity*, *2021*, 5935958:1–5935958:9.

Nguyen, T., Nguyen, L., Tran, P., & Nguyen, H. (2021b). Improving transformer-based neural machine translation with prior alignments. *Complexity*, *2021*, 5515407:1–5515407:10.

Nguyen, T. T., Phan, T. C., Nguyen, M. H., Weidlich, M., Yin, H., Jo, J., & Nguyen, Q. V. H. (2022b). Model-agnostic and diverse explanations for streaming rumour graphs. *Knowledge-Based Systems*, *253*, 109438.

Nguyen, T. C. V., Vuong, T. T., Le, D. T., & Ha, Q. T. (2022a). *V3mfnd: A deep multi-domain multimodal fake news detection model for Vietnamese*. Proc. of Asian conference on intelligent information and database systems, Ho Chi Minh City, Vietnam, pp. 608-620.

Nti, I. K., Adekoya, A. F., Weyori, B. A., & Nyarko-Boateng, O. (2022). Applications of artificial intelligence in engineering and manufacturing: A systematic review. *Journal of Intelligent Manufacturing*, *33*(6), 1581–1601.

Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., (2011). Scikit-learn: Machine learning in python. *Journal of Machine Learning Research*, *12*, 2825–2830.

Prabhakar, S. K., & Won, D. O. (2021). Medical text classification using hybrid deep learning models with multihead attention. Computational Intelligence and Neuroscience, Article ID 9425655.

Raj, C., & Meel, P. (2021). Convnet frameworks for multi-modal fake news detection. *Applied Intelligence*, *51*(11), 8132–8148.

Rivera-Trigueros, I. (2022). Machine translation systems and quality assessment: A systematic review. *Language Resources and Evaluation*, *56*(2), 593–619.

Sahoo, S. R., & Gupta, B. B. (2021). Multiple features based approach for automatic fake news detection on social networks using deep learning. *Applied Soft Computing*, *100*, 106983.

Škrlj, B., Martinc, M., Kralj, J., Lavrač, N., & Pollak, S. (2021). Tax2vec: Constructing interpretable features from taxonomies for short text classification. *Computer Speech & Language*, *65*, 101104.

Tran, K. T., Dinh, H. M., & Phan, T. T. (2022a). Building an enhanced sentiment classification framework based on natural language processing. *J. Intell. Fuzzy Syst*, *43*(2), 1771–1777.

Tran, P., Nguyen, T., Vu, D. H., Tran, H. A., & Vo, B. (2022b). A method of Chinese-Vietnamese bilingual corpus construction for machine translation. *IEEE Access*, *10*, 78928–78938.

Tran, K. T., & Phan, T. T. (2020). Capturing contextual factors in sentiment classification: An ensemble approach. *IEEE Access*, *8*, 116856–116865.

Vidros, S., Kolias, C., Kambourakis, G., & Akoglu, L. (2017). Automatic detection of online recruitment frauds: Characteristics, methods, and a public dataset. *Future Internet*, *9*(1), 6.

Vo, H. A., Le, H. S., Vo, M. T., & Le, T. (2019). A novel framework for trash classification using deep transfer learning. *IEEE Access*, *7*(1), 178631–178639.

Vo, H. A., Nguyen, T., & Le, T. (2020). Brent Oil price prediction using Bi-LSTM network. *Intelligent Automation & Soft Computing*, *26*(6), 1307–1317.

Vo, M. T., Vo, A. H., & Le, T. (2022a). A robust framework for shoulder implant X-ray image classification. *Data Technologies and Applications*, *56*(3), 447–460.

Vo, M. T., Vo, A. H., Nguyen, T., Sharma, R., & Le, T. (2021). Dealing with the class imbalance problem in the detection of fake job descriptions. Computers. *Materials & Continua*, *68*(1), 521–535.

Vo, M. T., Vu, D., Nguyen, H., Bui, H., & Le, T. (2022b). *Predicting monthly household water consumption*. Proc. Of international conference on computing and communication technologies, Ho Chi Minh City, Vietnam, pp. 720-724.

Youden, W. J. (1950). Index for rating diagnostic tests. *Cancer*, *3*, 32–35.