Effective Identification of Spam Jobs Postings Using Employer Defined Linguistic Feature

Conference Paper · May 2022

DOI: 10.1109/ICAIC53980.2022.9897059

6 authors, including:

Bishwajeet Pandey
Jain University
305 PUBLICATIONS **2,283** CITATIONS
SEE PROFILE

Naman Bhoj
14 PUBLICATIONS **28** CITATIONS
SEE PROFILE

Hardik Gohel
University of Houston – Victoria
34 PUBLICATIONS **613** CITATIONS
SEE PROFILE

Sivaram Ponnusamy
Raisoni Group of Institutions
48 PUBLICATIONS **8** CITATIONS
SEE PROFILE

Effective Identification of Spam Jobs Postings Using Employer Defined Linguistic Feature

Bishwajeet Pandey
Gyancity Research Labs, India.
bishwajeet.pandey@gyancity.com

Tanisq Kala
Accenture Solution Pvt. Ltd, India
kalatanishq588@gmail.com

Naman Bhoj
Independent Researcher, India
namanbhoj99@gmail.com

Hardik Gohel
University of Houston, Victoria, USA.
hagohel@gmail.com

Abhay Kumar
Jain University, Bangalore, India
k.abhay@jainuniversity.ac.in

P Sivaram
Jain University, Bangalore, India
ponsivs@yahoo.com

2022 1st International Conference on AI in Cybersecurity (ICAIC) | 978-1-6654-0043-5/22/$31.00 ©2022 IEEE | DOI: 10.1109/ICAIC53980.2022.9897059

***Abstract*— With businesses expanding rapidly due to the integration of the internet. Hiring excellent candidates has become really essential for businesses. This has made hiring candidates online very common which has also consequently increased job scams on online job portals which critically puts the privacy of applicants at stake. Pertaining to this in this paper, we investigate the performance of various machine learning algorithms to identify job scams on online job portals based on employer defined linguistic feature. This approach would significantly help job portals to identify scam posting and make the process of identification cost effective, accurate and fast. Our proposed work achieved an accuracy of 98.679% and contributed significantly to the existing knowledge in the domain.**

Keywords-component; online job scams; fake job posting detection; machine learning; deep learning; Bi-LSTM;

I. INTRODUCTION

With continuous advancement in technology, it has become very convenient for applicants to apply for jobs online and also easier for employers to select the candidates online, this consequently creates an opportunity for scammers to execute job scams [1].

Fake jobs scam is a very challenging issue presently. Keeping in mind that due to COVID-19 almost everything is online, this has also led to companies and the government posting job vacancies online for the job seekers. The recruitment follows a given procedure, based on which candidates are shortlisted. This is the exact time frame in the process where scammers call candidates and ask for money and bank details for finalising selection.

In scenarios where these events were noticed it was found that most of the jobs were fake, which led to applicants losing personal data as well as suffering financial loss.

Reports indicate that 188 million people are unemployed in the world, pertaining to which such online job posting portals have become an entity of great importance to help people secure jobs .

This has led to a lot of fake job postings on such online job portals where scammers use a company's logo and name for fraudulent activities which leads to bad name for the company and financial and data loss for the applicant.

Recently, a gang was caught in India which cheated more than 27 thousand job seekers and collected approximately 10 million rupee by the application fees just in a month of November 2020, they used the name of Union Health Ministry for this fraud. It was found that they sent 14 lakhs messages to applicants and in messages linked to fake websites for registration of 13 thousand vacancies, though researchers have also worked on devising effective techniques in identifying malicious websites in the past [2,3]. Due to high unemployment, candidates trusted it to be a real job advertisement and as a consequent suffered from the fraudulent activity of the scammer. very easily in these types of websites to get a job. These job scams are executed mainly using the following techniques [4] :

- Scammers send email to job seekers saying that they found their curriculum vitae (CV) in some job posting website and that your CV is perfect for this post.
- Legitimate job posting site but fake jobs. In this type of job scams the job board is a famous brand but the job may be fraudulent.
- When jobs are fake, the employer is fake and the job posting site is also fake. These types of scams are very difficult to catch, the website looks very professional, they post jobs on the website. Their aim is to collect our personal data as much as they can and ask for our bank details which may lead to financial loss.
- Fake linkedin profiles are created for job recruitment, fake facebook groups who give the opportunity for jobs very easily.

Keeping this in mind in this paper we aim to build a robust system which can identify fake job postings on both fake and authentic websites. Some of the major contributions made in this paper are:

- Investigate performance of machine learning and deep learning approach to detect fake job postings.
- We use employer defined linguistic feature which are also what a real human being would see on such job portals, therefore making our model intuitive. These include the title, department, company profile, job description, requirements and benefits posted by the potential employer.

The rest of the paper is structured as follows : Section 2 delves into existing research conducted on the topic, Section 3 explains the approach undertaken in this paper, Section 4 analyses the results and Section 5 concludes the paper.

II. BACKGROUND RESEARCH

Researchers in the past have conducted various studies to mitigate the problem of fake job postings. In this section we briefly go over the existing research conducted in the literature.

Shawni Datta et al. [5] proposed a methodology to detect the fake jobs scam on the internet using machine learning. Mainly, he used the two types of classifier: single classifier and ensemble classifier, his maximum accuracy was 98.27% in Random Forest classifier.

In their research Ibrahim Nasser et al. [6] proposed different machine learning classifiers (Naive Bayes, Support Vector Machine, Decision Tree KNN and Random Forest) to detect the online fake jobs. The highest accuracy was 98.2% for the Random Forest classifier.

(R.S. Shishupal, Varsha, S. Mane, V. Singh and D. Waseker) [7] proposed a methodology to detect the fake jobs by doing communes through speech and message using Natural language Processing (NLP). Accuracy of 96.2% was achieved using this technique.

In their research FHA. Shibly et al. [1] proposed two different types of Machine learning algorithms to detect the online fake job vacancies. The accuracy was 93.8% for two class Decision Boosted Trees and 95.4% for two class Decision Forest algorithms. S. Vidros, C.Kolias, G. Kambourakis and L. Akoglu [8] proposed that online job recruitment scam is very dangerous for both companies as well as job seekers. In their research Random Forest was the best algorithm for detection of fake jobs, the accuracy achieved was 91.22%.

B. Alghamdi and F. Alharby [9] proposed ensemble classifier approach on the Random Forest algorithm to detect the online fake jobs. He used the Random Forest as an ensemble classifier to detect and classify the cyber attack on the job scams.The accuracy achieved was 97.41% in their research.

Pertaining to existing research in the literature, in our paper we employ deep learning techniques to enhance the predictive accuracy of the models and build a robust prediction system.

III. PROPOSED WORK

This section aims to introduce readers to the approach undertaken in our research.

A. Dataset

We utilized the fake job posting dataset available on Kaggle [10]. The dataset initially consisted of 18 features. We treated the "fraudulent" feature in the dataset as the dependent feature. We were then left with 17 features out of which we only used 6 text based features to build our model. The detailed approach is discussed in the next subsection.

B. Data Analysis

The dataset initially consisted of 17 independent and 1 dependent features. In our research we only used 6 independent text based features to build our model. This was primarily done so that an intuitive model which can detect fake job posting using only text based description of the job can be modelled. The 6 text based features used in the analysis are namely : {title of job, department information, company description, job description, job requirements and benefits provided by the job}.

We combined the text of all 6 different text based feature in the following order :

- Title
- Department information
- Company description
- Job description
- Job requirements
- Benefits

This provided us with a single concatenated text feature, which can be termed as "Employer Defined Linguistic Feature" as the content on these features is posted by the employer. A sample example of what the feature looked like shown below in Figure 1.

Figure 1, is an example of an employer defined linguistic feature containing all the text data in the mentioned sequence above. This was largely important as any human subject particularly to know more about any job would definitely go through these text features before applying for that job.

Figure 1.Concatenated Employer Defined Linguistic Features

"Account Executive - Sydney Sales Adthena is the UK’s leading competitive intelligence service for Google search advertisers. A dthena is loved by major brands and digital agencies alike and provides a great opportunity to work in the high growth adtech s pace. Our patent-pending technologies provide unparalleled accuracy for clients to understand their competitors’ keywords, budg ets, spend, CPCs, Adcopy and more. We're profitable, fast growing and love what we do. Are you interested in a satisfying and f inancially rewarding role in a high growth technology company? You’ll work in a casual yet high energy environment alongside pa ssionate people delivering the leading competitive intelligence solution to major global brands. With the continued rapid growt h of digital marketing and PPC a huge opportunity exists to further expand the Adthena enterprise client base. We are seeking a n experienced Account Executive to develop and close new business in enterprise accounts. A large opportunity exists across all major search engine verticals including: Auto, Retail, Gaming, Travel, Finance, Insurance and Education.Primarily you’ll use a consultative approach to determine customer needs and deliver presentations and technical demonstrations. Where required, you w ill work with a client’s advertising agency to uncover value for prospective customers. You’ll be supported by marketing and bu siness development reps to build a pipeline of accounts. You’ll need to be smart and passionate and have 2 years experience sel ling software/Saas ideally including familiarity with PPC and marketing technologies. Excellent presentation and communication skills as well an understanding of marketing technologies in enterprise organisations. You should be an entrepreneurial self-st arter who is looking for a high growth technology environment and have strong skills in #URL_8d92932a488fb7e172d73a0f6813d06d46 4f1f03705d2825f86b2c7947d60a86#, Powerpoint and Google Apps. In return we'll pay you well, give you some ownership in the compa ny (stock options) and importantly provide you with excellent opportunities for advancement and professional development. Oh, a nd we’ll give you a new pair of Adidas trainers when you join. "

Then we analyzed our dataset for null values. We found our dataset to be containing 14095 null values which were then dropped off. This was followed by major text pre-processing steps for cleaning our text and reducing noise in our data.
The pre-processing steps employed in the research were: lowercasing text, removing punctuations as they would only not provide any significant information to identify a job posting as fraudulent, removing stop words as they do not provide any insights to job posting authenticity. This was followed by lemmatizing our text. These pre-processing steps helped us to clean our text data. This also made visualization of frequently occurring words in both fraudulent and authentic job cases easier. Figure 2 and Figure 3 shows important words occurring in the vocabulary of authentic and fraudulent job postings.

Figure 2. Important words in authentic job postings

Figure 3. Important words in fraudulent job postings

This was then followed by distributing our dataset in an 80-20 split, where 80% of the data was used for training purposes and 20% of the data was used for testing purposes.
The machine learning and deep learning models used in our research are discussed in the next subsection.

C. Models

1) **Random Forest (RF)** : Random Forest [11] algorithm is a vigorous supervised learning algorithm efficient for performing both regression and classification. Algorithm builds on a forest of decision trees using the Information Gain, Gini Index approach. The forest chooses the class having the most votes. So overall in random forests, the more the number of trees, the higher will be the accuracy.

2) **Support Vector Machine (SVM)** : Support vector machines [12] is a machine learning technique used for both classification and regression problems. It helps us to place new data points in the exact category and basically separates the plane between each class. The best boundary that

separates the plane is called the hyperplane. The extreme cases chosen by the SVM in making the hyperplane are called support vectors. Algorithm is executed with a kernel which makes it more powerful by transforming the input data into desired form.

3) **Bi-Direcrtional Long Short-Term Memory (B-L)** : Bi-Directional LSTM [13] are also a variation of Recurrent Neural Network and an extension of traditional Long Short-Term Memory models. These particularly improve the performance of sequence classification problems as they train two Long Short-Term Memory models instead of one on the input sequence. In our research we treat text of job posting as sequence data and feed it to our Bi-LSTM model. The model architecture consists of an input embedding dimension of 32, vocabulary size of 10000 and max length of 500. This was followed by a SpatialDropout of 0.2, Bi-Directional layer with 64 units of neurons, finally a dense layer with 24 units followed by a dense output layer with 1 neuron.

In order to extract features from the text for feeding them to machine learning algorithms we used word embeddings [14]. Tensorflow with keras backend was used to implement it with a vocabulary size of 100000, embedding dimension of 32 and maximum length of 500 .
The results of our approach are discussed in detail in the next section.
.

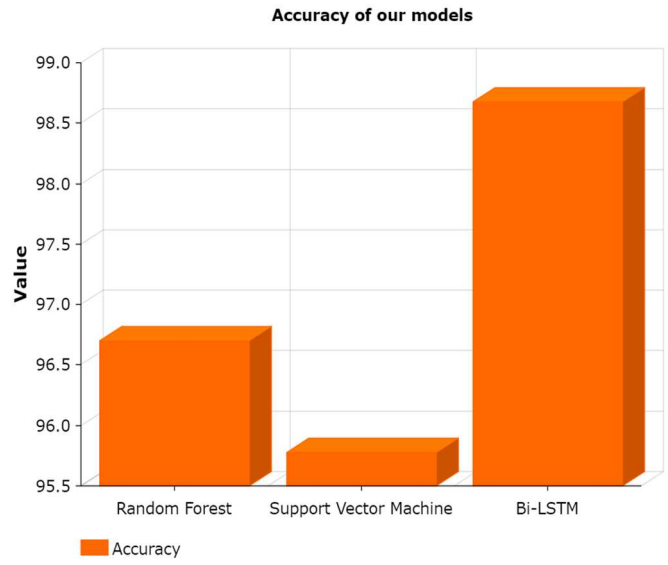
IV. Result and Analysis

In this section we discuss and analyze the results in two phases :

1) Comparative analysis of our models.
2) Comparative analysis of our best model with existing work in the literature.

For evaluating the performance of our models we use Accuracy and F1-Score as evaluation measures which are briefly defined below.
Accuracy: In a multiclass classification problem accuracy is defined as the sum of True Positive and Negative divided by the number of instances.
F1-Score: It is simply the harmonic mean of precision and recall and is indicative of the test's accuracy.

First Phase Analysis:
In this phase of our analysis we focus on comparing and analyzing the performance of approaches undertaken in our paper.
Figure 4. Comparison of Accuracy of our models.



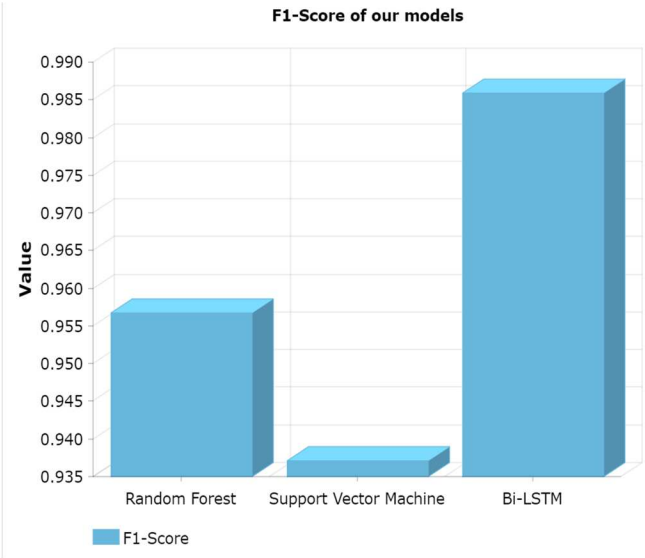


Figure 5. Comparison of F1-Score of our models

It is evident from Figure 4 and Figure 5 that our deep learning based Bi-LSTM model outperforms the machine learning model in identifying fake job postings using employer defined linguistic feature.
This is due to the fact that the Bi-LSTM model is capable of learning sequence based relationships in our text data and was able to more efficiently model the text based data compared to the machine learning models.
The quantitative comparison results of the models are aggregated in Table 1.

Table 1. Highest Accuracy model and its comparison with other models

Metric	Bi-LSTM	Improvement RF	Improvement SVM
Accuracy	**98.679%**	1.982	2.906
F1-Score	**0.98588**	0.02921	0.04883

Second Phase Analysis:

In this subsection we aim to compare the results of our model with existing research in the literature.

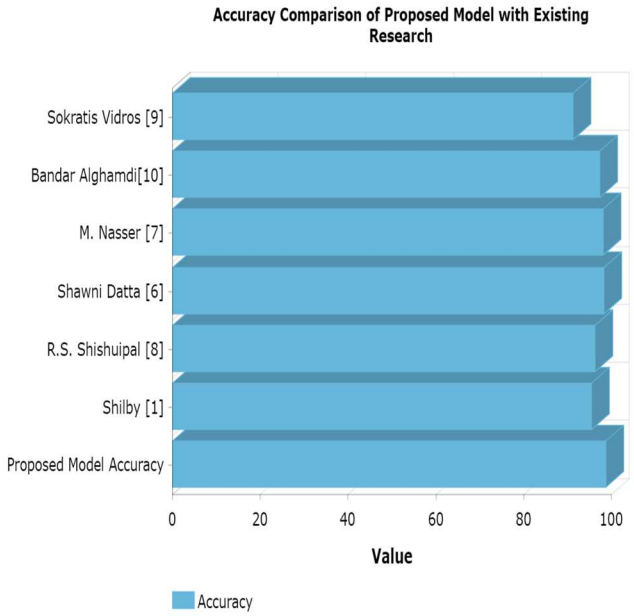


Figure 6. Comparison of Proposed Model with Existing work

We can clearly see from Figure 6 and Table 2 that our proposed model outperforms the majority of existing research work considered in this paper..

Table 2. Quantitative Accuracy Comparison

Existing Research	Metric	Proposed Work
Shawni Dutta et al. [6]	Accuracy = 98.27%	**Accuracy = 98.679%**
M. Nasser[7]	Accuracy = 98.2%	**Accuracy = 98.679%**
R.S. Shishupal et al.[8]	Accuracy = 96.2%	**Accuracy = 98.679%**
Shibly [1]	Accuracy = 95.4%	**Accuracy = 98.679%**
Sokratis Vidros et al. [9]	Accuracy = 91.22%	**Accuracy = 98.679%**
Bandar Alghamdi et al. [10]	Accuracy = 97.41%	**Accuracy = 98.679%**

V. CONCLUSION

In this paper we analysed and experimented with different machine learning models to identify fake job posting on job portals. We used employer defined linguistic feature to perform our analysis which consisted of various parameters defining a job posted on the portal. We experimented using Random Forest, Support Vector Machines and Bi-Directional LSTMs. The Bi-Directional LSTMs model achieved the highest accuracy of 98.679%, whereas Support Vector Machines achieved the lowest accuracy of 95.773%. The use of employer defined linguistic feature will be advantageous to job portals to automatically identify such job posts and mark them as spam or remove them. The future work should focus on Multilingual Spam Job Posting Detection, so as to have a generalized system for all languages.

REFERENCES

[1] FHA. Shibly, U. Sharma,HMM. Naleer. “Performance Comparison of Two Class Boosted Decision Tree Snd Two Class Decision Forest Algorithms in Predicting Fake Job Postings,” Annals of the Romanian Society for Cell Biology, Apr. 2021, pp. 2462.

[2] “Fake job portals cheat 27,000 job-seekers, collect Rs 1.09 crore in one month. Hindustan Times,” 6 November 2020.

[3] N. Bhoj, et al. “Comparative Analysis of Feature Selection Techniques for Malicious Website Detection in SMOTE Balanced Data.” RS Open Journal on Innovative Communication Technologies, vol. 2, issue 3, pp. 1-10, 2021, doi:10.46470/03d8ffbd.993cf635.

[4] S. P. Joyce, “ *5 Major Types of Scam Jobs and Job Scams Online*. Job,” 4 October 2019.

[5] S. Datta and S.K. Bandyopadhyay "Fake Job Recruitment Detection Using Machine Learning Approach."

[6] I. Nasser and A.H. Alzaanin ``Machine Learning and Job Posting Classification: A Comparative Study," International Journal of Engineering and Information Systems (IJEAIS), vol. 4 issue 9 pp.6-14, September 2020,

[7] R. S. Shishupal, Vrasha, S. Mane, V. Singh and D. Wasekar, "Virtual Assistant for Prediction of Fake Job Profile Using Machine Learning," International Journal of Advanced Research in Science, Communication and Technology (IJARSCT), vol. 3, issues 2, March 2021

[8] S. Vidros, C. Kolias, G. Kamboiurakis and A. Akoglu,"Automatic detection of online recruitment frauds: Characteristics, methods, and a public dataset," Future Internet, vol. 9, issue 1, 3 march 2017.

[9] B. Alghamdi, and F. Alharby, (2019) "An Intelligent Model for Online Recruitment Fraud Detection," Journal of Information Security, vol. 10, pp.155-176.

[10] S. Bansal, " [Real or Fake] Fake JobPosting Prediction. Kaggle," 9 February 2020.

[11] L. Breiman, "Random forests," Machine learning 45.1 pp. 5-32, 2001.

[12] V. Vapnik, "*The nature of statistical learning theory*. Springer science & business media," 2013.

[13] T. Lapjaturapit, K. Viriyayudhakom and T. Theeramunkong, "Multi-Candidate Word Segmentation using Bi-directional LSTM Neural Networks," 2018 International Conference on Embedded Systems and Intelligent Technology & International Conference on Information and Communication Technology for Embedded Systems (ICESIT-ICICTES), 2018, pp. 1-6, doi: 10.1109/ICESIT-ICICTES.2018.8442053.

[14] Y. Bengio, R. Ducharme, P. Vincent and C. Jauvin, "A neural probabilistic language model," The journal of machine learning research 3, pp.1137-1155, 2003.