Industrial Engineering Journal
ISSN: 0970-2555
Volume : 52, Issue 5, May : 2023

ONLINE FAKE JOB ADVERTISEMENT DETECTION USING MACHINE LEARNING

SYED MUJEEB UL HASSAN[1], HOD – IT, ISL ENGINEERING COLLEGE[1,2,3,4,5]

MOHAMMED ARSHAD HUSSAIN[2], Assistant professor.

SYED WALI MOHIUDDIN[3], MOHAMMED MUJTABA MUQTADIR[4], QAZI MOHAMMAD FAISAL UDDIN[5]

ABSTRACT

Machine learning algorithms handle numerous forms of data in real-world intelligent systems. With the advancement in technology and rigorous use of social media platforms, many job seekers and recruiters are actively working online. However, due to data and privacy breaches, one can become the target of perilous activates. The agencies and fraudsters entice the job seekers by using numerous methods, sources coming from virtual job-supplying websites. We aim to reduce the quantity of such fake and fraudulent attempts by providing predictions using Machine Learning. In our proposed approach, multiple classification models are used for better detection. This paper also presents different classifiers' performance and compares results to enhance the results through various techniques for realistic results.

KEYWORDS

Machine learning, Random Forest, Fake jobs detection, Classification

INTRODUCTION

Every organization nowadays is the internet and social media dependent. Systems like enterprise applications, management information systems, Information systems for Human Resources, and office automation applications are pivotal for running work. Creating an effective workforce recruitment process is considered by employing online applications, as it is more convenient for applicants. The majority of the human asset specialists and associations empower the online application framework for the enlistment and choice cycle. It has many benefits. Candidates can apply without the time and transfer their educational program vitae for additional references. Managers additionally can channel the applications rapidly and make waitlists within a brief period. In this way, electronic enrollment makes human resource capacities fast. It gives an ideal chance to online scammers to exploit their distress on these needy occasions when thousands and millions of individuals seek jobs. Over time, there is an expansion in these fake job posts where ads appear to be very ordinary, frequently these organizations will likewise have a site and will have an enlistment interaction like different firms in the area (Ward, Gbadebo, & Baruah, 2015). Online Recruitment Fraud (ORF) is becoming a severe issue in recent times. Due to hype in social media, online job advertisements are growing rapidly, but with advantages, there are many scammers, fraud employers scam them for money or taking personal information. Deceitful jobs ads can be posted using a well-known organization for disregarding their validity (Ward et al., 2015). Detecting fake job posts has taken consideration for acquiring an automatic tool, recognizing fake ads positions, and revealing them to individuals to stay away from the application for such positions.

RELATED WORK

All Fraud jobs advertisements can be viewed as bogus data on the web and as a type of scam. Information on the internet can be false, which is divided into misinformation and disinformation. If information is falsely created by misunderstanding or misconception, disinformation is purposefully made to cheat per user (Kumar & Shah, 2018). Fake job ads are considered disinformation. Supervised and

unsupervised learning solves disinformation-related problems such as fake news and reviews. Bondielli and Marcelloni (2019) suggested two approaches in their paper. The first approach uses factchecking websites for source information validation, it is named knowledge-based detection and the second approach uses the key attributes and extraction of essential features from source information. Fake or bogus news datasets are created manually based on multiple resources that are:

• Creation of Fact-checking websites such as FakeNewsNet Dataset (Murtagh, 1991).

• Using document samples labeled dataset in Burfoot Satire News Dataset by Burfoot and Baldwin

• Credbank Dataset by Mitra and Gilbert (2015) approach by dataset gathering by using expert judgment. For classification, supervised and unsupervised, both algorithms can work. Random forest agave learning-based approach where each classifier comprehends numerous tree-like classifiers applied to various examples, and each tree votes in favor of the most fitting class. Another helpful technique can be boosting, which can work with multiple classifiers for a single classifier to improve classification results. Extended innovation applies an algorithm for classifying the weighted adaptations of training data and chooses the grouping of the more significant voting classifier. AdaBoost illustrates a procedure of boosting, which delivers better effectiveness (Murtagh, 1991). Expanding algorithms implies tackling issues with spam filtration viably. In addition, Gradient boosting is an extra boosting procedure for a Classifier dependent on the decision tree rule (Prentzas et al., 2019). It likewise limits the deficiency of accuracy. Algorithms approaches that can distinguish fake advertisements in online media are the decision forest. Models of a quick, controlled ensemble. The decision tree can be the best model assuming the need to anticipate a target for up to two tests. It is suggested to train and test different models by utilizing the Tune Model Hyperparameters system. Alghamdi and Alharby (2019) provided a model for detecting scam posts in online job ads systems. The authors had used the EMSCAD dataset on various machinelearning algorithms. The methodology is divided into 3 steps preprocessing, selection of features, and identifying scams by the classifier:

• In step, one unwanted noise and tags are removed from the data and bringing into general text.

• To reduce extraversion features that are not in use selective features are selected using a support vector machine and random forest classifier.

• It is reported that the detect fake job posts classification accuracy showed 97.4%.

METHODOLOGY

What is the best suitable classification algorithm for detecting Fake job advertisements? What are the appropriate and important features for fraudulent job detection? This research aims at constructing a suitable model to detect fraudulent job advertisements, to protect the expatriates from falling into the trap. This research falls under the category of an empirical study that would be based on observation, testing, evaluations, and comparison of the applied algorithms.

PROPOSED APPROACH

The research understudy can be described as a three-tier approach starting with the dataset preprocessing, feature selection, and classifying by applying different machine learning models and evaluating them. Let us look at the research that has already been done in this field of detecting fraudulent advertisements or detection of spam emails etc., over a period. It is observed that many researchers have applied several classification algorithm including random forest.

DATASET

This research works on a dataset from Kaggle to categorize a job advertisement as fraudulent or not based on some attributes derived from the advertisements available on different sources. The

Industrial Engineering Journal
ISSN: 0970-2555
Volume : 52, Issue 5, May : 2023

data was available in a CSV file having 17880 instances of jobs advertisements. Each advertisement is defined in terms of attributes on which we are working, that data is then preprocessed and classified through several algorithms.

PREPROCESSING DATASET

The initial dataset had 17 attributes based on which this model would be predicting the status of an advertisement. These 17 attributes include job id, title, location, department, salary range, company profile, description, requirements, benefits, and telecommuting, has the company logo, has questions, employment type, required experience, required education, industry, and function. Each attribute contains either object or integer data. The label is binary for the specific problem domain, i.e., 0 for non-fraudulent and 1 for fraudulent. The preprocessing phase starts after analyzing the dataset for missing values and some basic statistical operations on the integer data. Our integer fields include job id, telecommuting, has the company logo, has questions, and the final label of being fraudulent or not. Figure 1 describes the number of missing values in each field; this description justifies the deletion of job IDs and salary range containing the maximum missing values. The integer fields were ten checked for the correlation, and Figure 2 depicts the correlation heat map.

IMPLEMENTATION OF CLASSIFIER

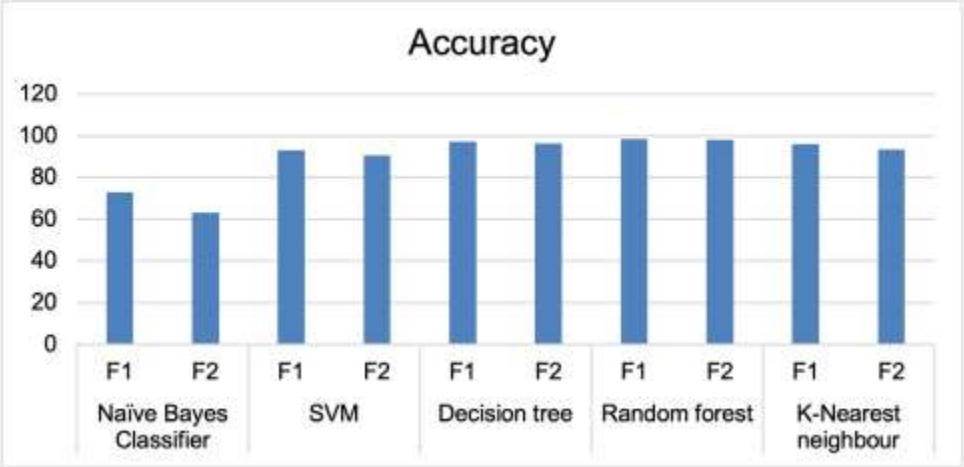
The proposed approach compares the performance of classifiers on two different feature sets. The first feature set includes the processed data discussed above and the second feature set has the integer attributes, benefits, and location. In this research, several classifiers are engaged, Random Forest Classifier, classifying job posts as fake. Note that 'fraudulent' is the target class for the research under discussion. For both the feature sets, the classifiers are passed on to the training phase with 80 percent of the entire dataset, the remaining 20 percent would be used for the prediction phase. Training the classifiers for the proposed approach starts with choosing the right and tuned parameters as default parameters do not guarantee the best and promising results. After the prediction of the testing data, the model would be then evaluated on metrics such as Accuracy, F-measure, and Cohen- Kappa score. They are keeping the work on both the feature sets in parallel. The best classifier would be chosen to have outstanding performance among all the peer classifiers for each feature set.

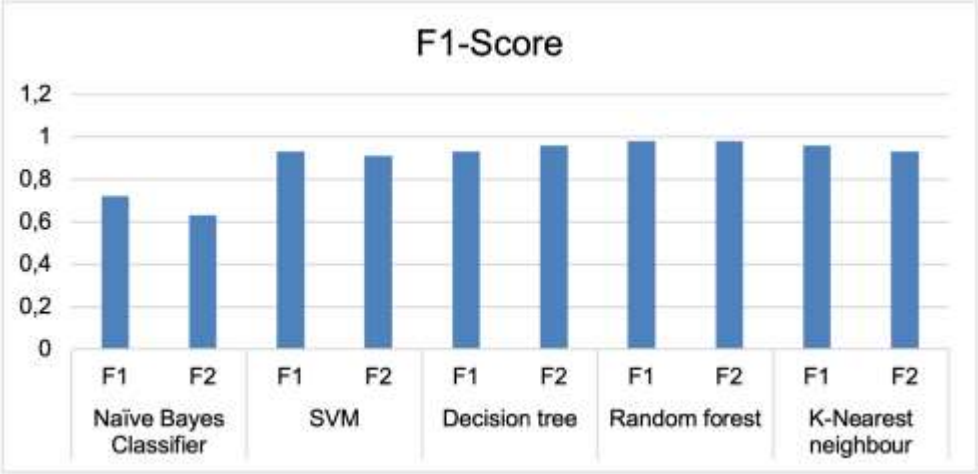
RESULTS

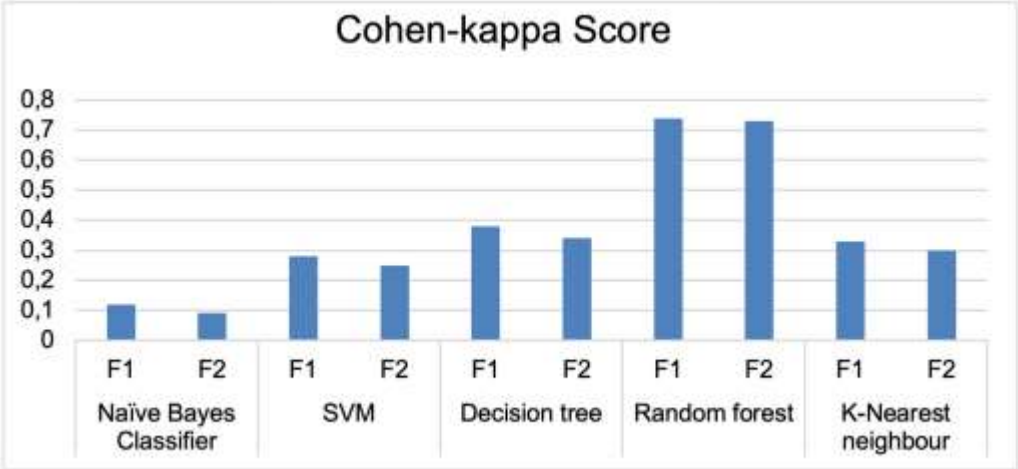
Comparative table of classifiers performance.

Performance Measure Metric	Naïve Bayes Classifier		SVM		Decision tree		Random forest		K-Nearest neighbour	
	F1	F2	F1	F2	F1	F2	F1	F2	F1	F2
Accuracy	73.03	63.21	93.04	90.8	97.2	96.3	98.27	98	95.9	93.5
F1-Score	0.72	0.63	0.93	0.91	0.93	0.96	0.98	0.98	0.96	0.93
Cohen-kappa Score	0.12	0.09	0.28	0.25	0.38	0.34	0.74	0.73	0.33	0.3
MSE	0.52	0.59	0.06	0.075	0.04	0.04	0.02	0.02	0.041	0.049

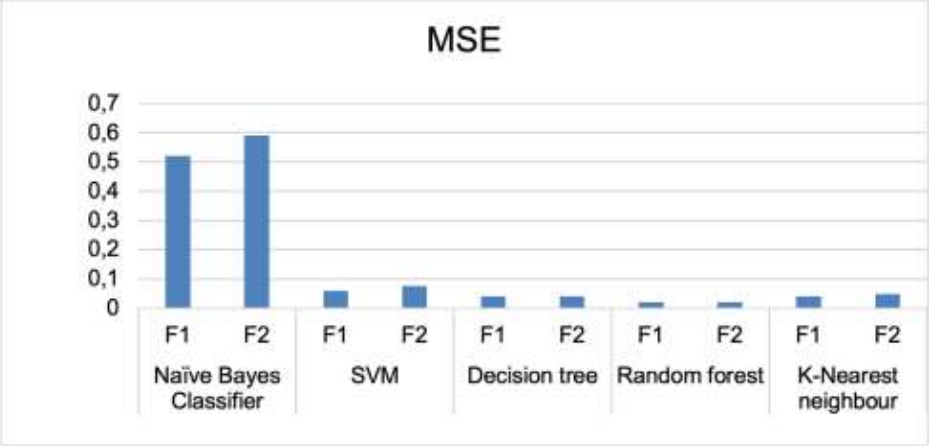
Industrial Engineering Journal
ISSN: 0970-2555
Volume : 52, Issue 5, May : 2023







Industrial Engineering Journal
ISSN: 0970-2555
Volume : 52, Issue 5, May : 2023



CONCLUSIONS

Platforms such as online job portals or social media for job advertisements are an exciting way of attracting potential candidates on which many enterprise companies are dependent on the hiring process. Fake jobs scam detection at an early stage can save a job seeker and make them only apply for legitimate companies. For this purpose, various machine learning techniques were utilized in this paper. Specifically, supervised learning algorithms classifiers were used for scam detection. This paper experimented with different algorithms such as naïve Bayes, SVM, decision tree, random forest, and K-Nearest Neighbor. It is reported that the K-NN classifier gives a promising result for the value k=5 considering all the evaluating metrics. On the other hand, Random Forest is built based on 500 estimators on which the boosting is terminated.

REFERENCES

1. Alghamdi, B., & Alharby, F. (2019). An intelligent model for online recruitment fraud detection. Journal of Information Security, 10(03), 155. https://www.scirp.org/journal/paperinformation.aspx?paperid=93637
2. Bondielli, A., & Marcelloni, F. (2019). A survey on fake news and rumour detection techniques. Information Sciences, 497, 38-55.
3. https://app.dimensions.ai/details/publication/pub.1114201506
4. Islam, M. R., Liu, S., Wang, X., & Xu, G. (2020). Deep learning for misinformation detection on online social networks: a survey and new perspectives. Social Network Analysis and Mining, 10(1), 1-20. https://doi.org/10.1007/s13278-020-00696-x
5. Kumar, S., & Shah, N. (2018). False information on web and social media: A survey. arXiv preprint arXiv:1804.08559.
6. Mitra, T., & Gilbert, E. (2015). Credbank: A large-scale social media corpus with associated credibility annotations. In Ninth international AAAI conference on web and social media. https://ojs.aaai.org/index.php/ICWSM/article/view/14625
7. Mujtaba, G., & Ryu, E. S. (2020). Client-driven personalized trailer framework using thumbnail containers. IEEE Access, 8, 60417-60427. https://ieeexplore.ieee.org/document/9046852
8. Mujtaba, G., & Ryu, E. S. (2021). Human Character-oriented Animated GIF Generation Framework. In 2021 Mohammad Ali Jinnah University International Conference on Computing (MAJICC) (pp. 1-6). IEEE.
9. Mujtaba, G., Lee, S., Kim, J., & Ryu, E. S. (2021). Client-driven animated GIF generation

framework using an acoustic feature. Multimedia Tools and Applications, 1-18. https://link.springer.com/article/10.1007/s11042-020-10236-6