Current Applied Science and Technology Vol. 22 No. 6 (November-December 2022)

Research article

Detecting Fraud Job Recruitment Using Features Reflecting from Real-world Knowledge of Fraud

Boonthida Chiraratanasopha[1]*and Thodsaporn Chay-intr[2]

[1]*Faculty of Science Technology and Agriculture, Yala Rajabhat University,Yala, Thailand*
[2]*School of Engineering, Tokyo Institute of Technology, Tokyo, Japan*

Received: 18 August 2021, Revised: 28 December 2021, Accepted: 17 February 2022

DOI: 10.55003/cast.2022.06.22.008

Keywords

fake job advertisement;
internet fraud;
feature design;
fraud detection

Abstract

A common method for text-analysis and text-based classification is to process for term-frequency or patterns of terms. However, these features alone may not be able to differentiate fake and authentic job advertisements. Thus, in this work, we proposed a method to detect fake job recruitments using a novel set of features designed to reflect the behavior of fraudsters who present fake information. The features were missing information, exaggeration, and credibility. The features were designed to represent in the form of a category and an automatically generatable score of readability. Data from EMSCAD dataset were transformed in accordance with the designed features and used to train a detection model for fake job detection. The experimental results showed that the model from the designed features performed better than those based on the term-frequency approach in every applied machine learning technique. The proposed method yielded 97.64% accuracy, 0.97 precision and 0.99 recall score for its best model when used for classifying fake job advertisements.

1. Introduction

With advancement in telecommunication technology, people can easily connect and communicate via the internet, regardless of their location. However, one of the issues with the internet is internet fraud. Internet fraud refers to the crime of using the Internet services to defraud victims, illegally taking advantage of them [1-4]. Such crimes continue to plague the Internet through various methods such as forging fake electronic documents in an attempt to deceive the internet users into divulging sensitive personal information. The term 'sensitive personal information' refers to data required to be protected to safeguard the privacy or security of an individual [2, 5, 6] since such data can be

*Corresponding author: Tel.: (+66) 843133015
E-mail: boontida.j@yru.ac.th

used to validate individual identity in online service. This is, in effect, identity theft. A common method for stealing sensitive personal information is to forge a fake document to deceive internet users to provide the information themselves by imitating legitimate business sources. The forged online documents include e-mail, post, and website.

Online recruitment or E-recruitment is the use of an online website service as a job-listing platform in order to recruit employees. Online recruitment is a popular method since it leads to a much wider applicant pool, consumes less time and money, and involves automated processes such as application screening and communication services. However, the job-listing platform is vulnerable to internet fraud as fraudster can manufacture fake job recruitment advertisements, targeting job seekers' sensitive information including their identification cards, bankbooks, and tax file numbers, some or all of which are normally asked in the job hiring process. It is a difficult task to differentiate fraudulent job advertisements from legitimate ones even by experts since online recruitment platforms have specific required details for job postings, and because fake information can be skillfully crafted.

There have been many research efforts attempting to detect internet fraud such as spam/phishing e-mail detection [7, 8] and faking job advertisement detection [9-13]. In these efforts, researchers applied machine-learning approaches to develop a classification model to differentiate legitimate and forging documents. Since the documents contain textual information, a widely-used textual technique, term-frequency-based analysis was commonly exploited. The term-frequency approach evidently shows high-performance to differentiate documents into categories by calculating word statistics as it signifies the probability of words in their respective categories. However, unlike ordinary text categorization, term-frequency may not perform well in detecting fraudulent job advertisements since the fake information given in the forged advertisements is intentionally crafted to imitate legitimate recruitment information. In fact, fake job recruitment posting usually comes with exaggerations such as higher rate of salary, lower education requirements, and often feature catchphrases to bait more job seekers. Thus, using the ordinary method may not truly distinguish between fraud and legitimate job recruitment postings.

There are several research works trying to solve internet fraud issues using AI techniques. Therefore, detection of online employment scam and fake news are reviewed. Since these two fraud activities involve using textual information to deceive internet users, similar techniques are used for the detection task. Detecting employment scam or forged job recruitment is the task of identifying forged job recruiting posts among recruitment advertisements. Using the publicly available Employment Scam Aegean Dataset (EMSCAD), researchers have investigated employment scams and their detection in recent years. Vidros *et al*. [9] proposed the automatic detection of online recruitment frauds by applying bag-of-word modelling with machine learning-based classification techniques including ZeroR, OneR, Naives Bayes, J48 decision trees, random forest and logistic regression. They also analyzed the data and reported on various characteristics of fake job recruiting content such as the use of captialized words to gain attention, missing information, and short and sketchy textual information. Nasser and Alzaanin [10] applied text pre-processing for term frequency-inversed document frequency (TF-IDF) for feature extraction. This work converted textual information to term frequency features and used them for a classification model generated using machine-learning including Naïve Bay, support vector machine, decision tree, K-nearest neighbor and random forest. Dutta and Bandyopadhyay [11] developed fake job detection by transforming text into encoding categories. The feature vector was then used in ensemble classifiers including random forest, AdaBoost and Gradient Boosting. Experimental results indicated that the random forest classifier yielded the best performance. Mahbub and Pardede [12] proposed to include contextual features for online recruitment fraud detection with a J48 decision tree classifier to improve performance. The contextual features of this work included the validation of a company's internet footprint in order to ascertain creditability of the recruiting company, the identification of suspicious style of content such as the use of bold text in a textual information, and examination of

the range of the content. Then, they converted the validating results to binary features. This method significantly improved the performance of the task.

For the task of detecting fake news, text analysis and the machine-learning approach are also commonly used. Shukla *et al*. [14] applied text-mining using bag-of-word, n-gram, TF-IDF, and part-of-speech as features to develop a classification model for detecting fake news. They exploited LIAR dataset as fake news for training their fake news detection model. Akinyemi *et al*. [15] used TF-IDF to develop a fake news detection model using the PHEME dataset, which contained 5,800 tweets. Elmurngi and Gherbi [16] proposed the use of text-processing as a set of features in the task of fake review detection for movies. The annotated dataset was processed to word-vector and used for training two models, a stop-word excluded model, using machine-learning techniques such as Naïve Bay, support vector machine, decision tree, K-nearest neighbor.

These research efforts relating text-handling can be separated into two approaches, using text frequency to represent text features, and converting text to binary features using knowledge of fraudulent methods. The latter is more helpful since it reflects human knowledge via feature transformation, and it has been proven to improve the performance of a classifier. In this work, we design a novel set of contextual features derived from a knowledge of fraudster behavior and real-world information to specifically help detecting fake job advertisements.

This work specifically designs a set of features for fraud job advertisements. The main concepts of the features rationally align with how a fraudster baits a victim, interesting them in the advertisement. The features are then used to develop a classification model for detecting fraudulence in online job recruitments.

2. Materials and Methods

This work proposes a design of features used in the task of detecting forged job advertisement. Since the forged job advertisement is an attempt of fraud trying to steal sensitive personal information from job seekers, the forged job recruitment post is similarly crafted to contain similar information to a legitimate job advertisement regarding term usage. Hence, we specifically designed features to bring out the implicit nature of fraudulence regarding how fraudsters often use to draw attention of job seekers using exaggerated information and catchphrases. This research uses the public Employment Scam Aegean Dataset (EMSCAD) dataset [17], and instances of recruitment are classified as fraudulent or legitimate by converting the given attributes into the proposed features. The EMSCAD dataset contains 17,880 real-life job recruitment manually annotated into two categories consisting of 17,014 legitimate and 866 fraudulent job recruiting advertisements between 2012 to 2014 in English language from several countries such as USA, Great Britain, and Canada. Table 1 shows the original attributes given in the EMSCAD dataset and their respective converted features.

Original data from the EMSCAD are converted to be more reflexive to fraud detection. We decide to apply three concepts to address fraud: missing information, exaggeration, and credibility of data. First, a legitimate job advertisement from a genuine company should not contain missing information since the required information is common and public detail. The missing information thus can be used as a clue for fraudulent conduct. Second, exaggeration is an act to overstate the information to gain interest of job seekers. Exaggeration is a frequent tactic seen in fraudulence texts and examples are inflated salaries and benefits. Since the details of the job are fake, it is possible to give out exaggerated data to trick job seekers into applying. Third, credibility of the job details and company is one of the possible clues to detect fake information. However, credibility is a difficult attribute to be evaluated. In this work, we decide to apply a readability score calculation to represent how understandable a text is depending on the complexity of vocabulary and syntax used.

Table 1. Set of attributes from EMSCAD dataset and their respective converted features

No.	Feature	Datatype	Conversion
1	title	string	category (6)
2	location	string	category (7)
3	department	string	category (4)
4	salary_range	string	category (9)
5	company_profile	string	readability score
6	description	string	readability score
7	requirements	string	readability score
8	benefits	string	readability score
9	telecommuting	binary	category (2)
10	has_company_logo	binary	category (2)
11	has_questions	binary	category (2)
12	employment_type	string	category (6)
13	required_experience	string	category (8)
14	required_education	string	category (5)
15	industry	string	category (2)
16	function	category	category (3)
17	fraudulent (Label)	binary	binary (2)

The selected readability score in this work is Flesch-Kincaid readability [18-22], which can measure how difficult a passage in English is to understand, and offers results in a reading grade level. The higher the grade, the more complexity the text contains. We expected that the legitimate job advertisements would be different from forged ones in terms of official language style usage. The formula for Flesch-Kincaid readability is as follows.

$$0.39\left(\frac{total\ word\ counts}{total\ sentence\ counts}\right)+11.8\left(\frac{total\ syllable\ counts}{total\ word\ counts}\right)-15.59 \quad (1)$$

For feature conversion, we separated the features into three groups: non-processing, categorizing, and readability score calculations. The non-processing is for the binary (Boolean) features which are used as they are. For the categorizing feature, we adopted the concept of missing information and exaggeration together to design the feature. The conversion of each feature thus is different in terms of the nature of the information. Details of converting are given in Table 2.

For the salary feature, we used the current GDP of the given location to calculate the percentage of the salary. This calculation hence should be able to determine whether the given salary is exaggerated or not. The location feature is categorized based on how detail the information given is. It should help to ensure credibility and completeness of the information. Other features were then categorized regarding the type of relevant information.

To the features in the form of textual explanation including company_profile, description, requirements, and benefits, the aforementioned readability score was applied to calculate the complexity of the given text. The complexity of content was expected to determine the language style of the textual information. Legitimate advertisements were expected to contain good written content with a higher academic language while fraudulent were expected to have a more casual writing style.

Table 2. Criteria to convert data into a specified category

Feature	Categories and their criteria
title	1. No position + Short detail 2. 1 position 3. 1 position + short detail 4. >1 position 5. >1 position + short detail 6. No position + Invitation advertising
location	1. No Location 2. Country 3. Country + (state/region/Other Code/) 4. Country + (state/region/Other Code/) + (City/City Code/Other) 5. Country + (state/region/Other Code/) + >1 (City/City Code/Other) 6. Country + (City/City Code/Other) 7. Country + >1 (City/City Code/Other)
department	1. No Department 2. 1 Department 3. >1 Department 4. Other
salary	1. No salary specified 2. No location specified 3. ≤50% GDP 4. 50% GDP < salary ≤80% GDP 5. 80% GDP < salary <100% GDP 6. = GDP 7. 100% GDP < salary ≤120% GDP 8. 120% GDP < salary <150% GDP 9. ≥150% GDP
employment_type	1. No_Employment_type 2. Full-time 3. Part-time 4. Temporary 5. Contract 6. Other
required_experience	1. No_required_experience 2. Not Applicable 3. Associate 4. Internship 5. Entry level 6. Mid-Senior level 7. Director 8. Executive
required_education	1. No_req_education 2. Unspecified 3. Under Bachelor's Degree 4. Bachelor's Degree 5. Higher than Bachelor's Degree
industry	1. No_industry 2. In_industry
function	1. No_function 2. has_function 3. Other

With the designed features, the data were trained for the classification model. In this work, we chose K-nearest neighborhood (KNN), support vector machine (SVM) and decision tree (DT) as machine-learning techniques. Though we chose various techniques, we did not aim to compare their performance but to demonstrate that the designed features could improve the performance regardless of the applied technique. Due to the difference in the number of categories consisting of 17,014 legitimate and 866 fraudulent job recruiting advertisements, the dataset was unbalanced. Therefore, we needed to adopt the synthetic minority oversampling technique (SMOTE) [23-27] to reduce the effect of unbalanced data in the learning of the classification model. Without SMOTE, the performance of the classification would have favored the larger class because there was more data for machine-learning to observe.

3. Results and Discussion

3.1 Experiment settings

This work used EMSCAD dataset consisting of 17,014 legitimate and 866 fraudulent job recruiting advertisements. To prevent the unbalanced data issue, SMOTE was used to generate data for 20,000 legitimate and 20,000 fraudulent advertisements, i.e. for a total of 40,000 instances. Ten-fold cross-validation was applied to test the classification model. The features from the instances were converted following the criteria mentioned in the previous section. The learned model was analyzed and evaluated using metrics including accuracy, precision, and recall. To prevent the issue of overestimation of the model's performance and generalizability from SMOTE in testing data, the given results were from only the original data before calculating for evaluation results.

The machine learning techniques used were K-nearest neighborhood (KNN), support vector machine (SVM) and decision tree (DT). For the KNN parameter settings, K was set as 5. It should be noted that we did not aim to compare their performance of the techniques but to demonstrate that the designed features could improve the performance regardless of the techniques applied. For comparison, we set up three models as shown in Table 3.

Table 3. Models for comparing performance

	Categorization	**Readability**
No Conversion (NC)	no (TF-IDF)	no (TF-IDF)
Semi Conversion (SC)	yes	no (TF-IDF)
Full Conversion (FC)	yes	Yes

The conversion of the models is as mentioned in the previous section. For full-conversion, every feature was processed according to the proposed conversion method. The semi-conversion was to convert only for categorization but used term-frequency inverse-document frequency (TF-IDF) instead of readability. Last, the no conversion refers to use TF-IDF for all features.

3.2. Experimental results and discussions

The results of the models in terms of accuracy, precision, and recall are given in Figures 1 and 2. The results indicate that the full-conversion yielded the best performance for all measurement metrics while the semi-conversion gave slightly lower results than those of the full-conversion. The no-conversion which represents the traditional method, however, gained between 60 and 75 for all measurement metrics. The results signify that the proposed features strongly assisted the task of fraud detection for job recruiting advertisement.

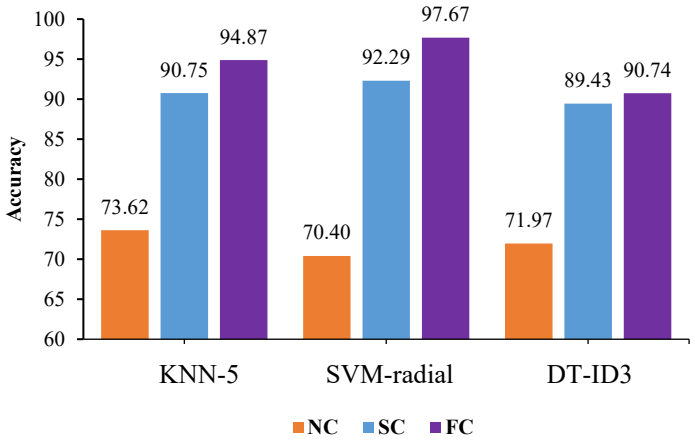

Figure 1. Accuracy results of the models

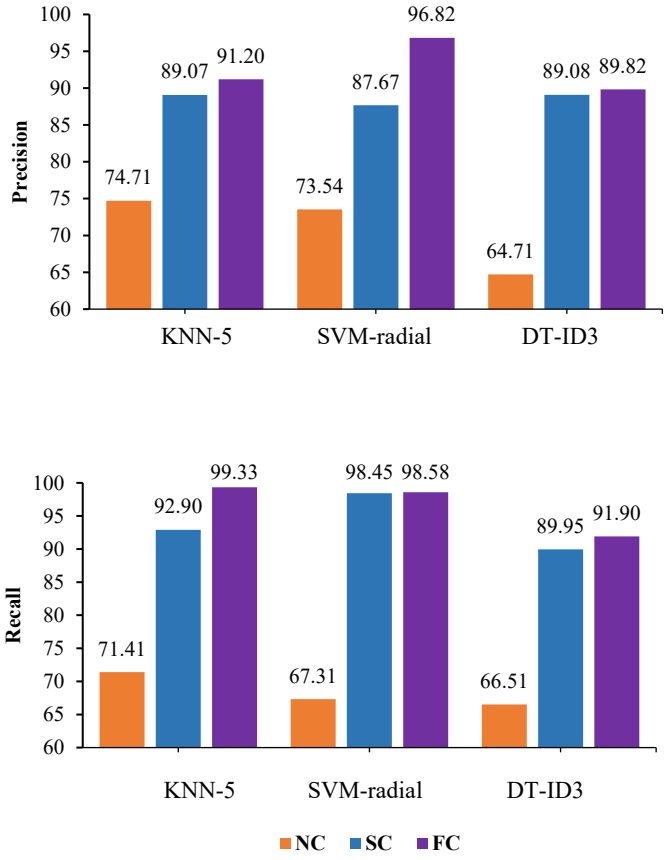

Figure 2. Precision and recall results of the models

From the result of the analysis of incorrect classification, there were two cases: false positive (predicted to be fraudulent but actually being legitimate), and false negative (predicted to be legitimate but actually being fraudulent). The job post instances of false positive cases included those with many missing details (7-14 features). For the false negative cases, the instances were a few missing values (1-2 features), and the forging details were realistic. For these cases, they were difficult to differentiate even for competent people.

To further analyze the results, we explored the original data of fraudulent ads and legitimate job posts for comparison. This comparison insight was then used to confirm why the purposed method could improve the classification results. First, missing data were common in both the fraud and legitimate posts, as shown in Figure 3. The fraud job posts had at least one missing feature and about 43% of them contained 8 to 11 missing features from a total of 16 features. Thus, the fraud job posts had more missing data on average, and the most common missing features were *company profile*, *department*, and *required education*. With such statistics, the transformation of features into categories was able to represent the informative aspects of data used to differentiate fraud and legitimate job posts.

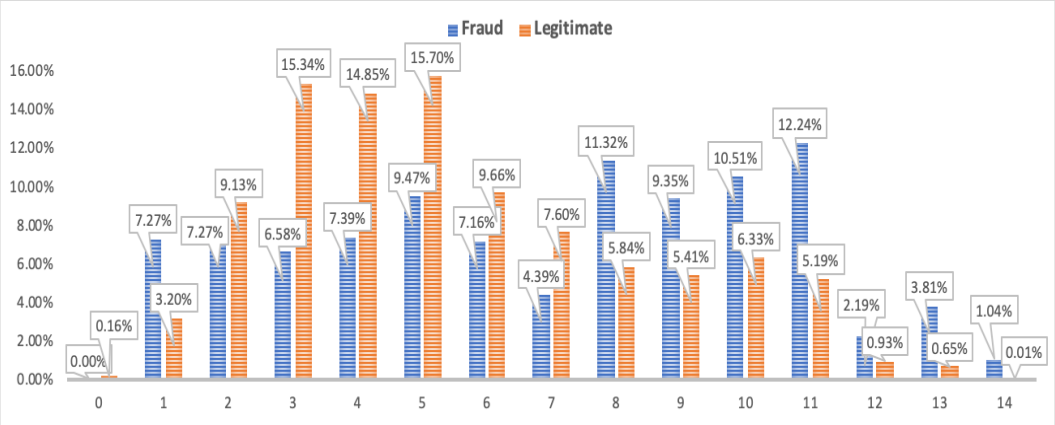

Figure 3. Percentage of missing features from raw data

In terms of text-based data such as *company profile* and *job description*, we could classify the raw data based on text range using word count and language style, as given in Figure 4. The statistics apart from missing features showed that most of the data for fraud posts were in the short and medium range, but those for legitimate posts were in the medium and long range. In addition, the complexity of word usage from legitimate instances was much higher. Examples of text data were "Are you looking to be paid to take vacations?", which was obvious fraud data while "[anonymous] is made up of a team of passionate consultants with the drive for putting people who wish to grow themselves through education whilst working into long term and relevant job roles." was from a legitimate company profile. Although some of fraud instances were well forged and included complex words of a longer range, the differences in the text nature were of assistance in determining what was fraud and what was legitimate.

To investigate further, we looked into the information gain (IG) score [28-31] generated in a process of decision tree. The IG score can indicate the significance of the features in the model. Therefore, we rank the top-5 features using average IG score of the 10-fold-cross validation from the full-conversion models. The top-5 features with IG score were as follows.

1. Profile (readability): 1.0
2. Location (category): 1.0

3. Requirements (readability): 0.8083
4. Description_(readability): 0.7652
5. Required education (category): 0.6530

Fraud			**Legitimate**		
	simple word	complex word		simple word	complex word
no given information	30.34%		no given information	14.66%	
short range (less than 10 words)	24.46%	3.12%	short range (less than 10 words)	9.57%	8.88%
medium range (10 to 30 words)	26.10%	9.84%	medium range (10 to 30 words)	21.15%	17.92%
long range (over 30 words)	2.80%	3.35%	long range (over 30 words)	14.91%	12.91%

Figure 4. Statistics of text-data from raw data

This ranking shows which features have an impact on classification. Among the five top-ranked features, there are 3 features that are converted using the readability score; hence, it can be assumed that readability score is relevant to how a fraudster forged the content. This shows that complexity in terms of vocabulary and syntax of the content can represent a quality of content that can be one of the key attributes to detect fake information, and is similar to a method used to detect fake news [32] and click bait detection [33].

Regarding the application of SMOTE, additional data in this work were synthesized for training. This however may have had an effect as the generalized content that creates a bias for classification model. This work handles the information according to how data are given, not what the data are. For example, according to Table 2, the *location* feature, which is the top feature, is considered by how many details of location are given such as the data are only country, and the data contains both country and city instead of taking the individual location into consideration.

Despite yielding high performance comparing to the baseline, the designed features have a disadvantage as some features need real-world information to assign the feature value. For example, the feature salary requires the GDP of the country where the company belongs to for assigning the proper category. This however increases more manual work for feature conversion. For readability score calculation, there are several formulae, but this work adopted the Flesch-Kincaid readability since it is designed for grading a reading passage for native English speakers. Furthermore, acquiring the readability score does not require extra effort since there are several tools for calculating it, yet the results are accountable for a task of internet fraud detection.

4. Conclusions

To tackle employment scams which are one of internet fraud issues, we proposed a set of features specifically designed to detect fake job advertisements. Based on annotated EMSCAD job advertisement data analysis, we studied fraudsters' patterns of thinking and the characteristics of forged information in recruiting posts. Accordingly, we designed features with respect to three types of concepts, i.e. blank information, exaggeration, and credibility of data. Exaggeration is a trait of fake information that aims to catch people's attention by overstating key details while credibility can be determined by suitable complexity of textual description according to job position. We combine the concepts of blank information and exaggeration into features in a form category type.

For complexity of content, we adopted the concept of readability score which grades a text for difficulty using the criteria of vocabulary and syntax. In this work, we selected the Flesch-Kincaid readability to grade the text content. Data from the EMSCAD dataset were transformed following the designed features and used to train a detection model for fake job detection using KNN, SVM and decision tree machine-learning methods. The experimental results indicate that models generated from the designed features yielded 97.64% accuracy, 0.97 precision and 0.99 recall scores from their best model when being used to classify fake job advertisements. Moreover, the models generated from the designed features outperformed the baseline which applies the traditional term-frequency based approach without machine-learning techniques.

References

[1] Fan, Q., 2015. The types, characteristics and countermeasures of internet fraud crime. *Proceedings of the International Scientific Conference "Archibald Reiss Days"*, Belgrade, Serbia, March 3-4, 2015, pp. 315-319.

[2] Eze-Michael, E., 2021. Internet fraud and its effect on NIGERIA's image in international relations. *Covenant Journal of Business and Social Sciences*, 11(3), 1-25.

[3] Ye, N., Cheng, L. and Zhao, Y., 2019. Identity construction of suspects in telecom and internet fraud discourse: from a sociosemiotic perspective. *Social Semiotics*, 29(3), 319-335.

[4] Norris, G., Brookes, A. and Dowell, D., 2019. The psychology of internet fraud victimization: A systematic review. *Journal of Police and Criminal Psychology*, 34(3), 231-245.

[5] Huang, Z., 2017. Causes and prevention of telecommunication network fraud. *Proceedings of the 2nd International Conference on Humanities Science and Society Development (ICHSSD 2017)*, Xiamen, China, November 18-19, 2017, pp. 164-173.

[6] Galbraith, M.L., 2012. Identity crisis: Seeking a unified approach to plaintiff standing for data security breaches of sensitive personal information. *American University Law Review*, 62, 1365-1397.

[7] Alqatawna, J., Faris, H., Jaradat, K., Al-Zewairi, M. and Adwan, O., 2015. Improving knowledge based spam detection methods: The effect of malicious related features in imbalance data distribution. *International Journal of Communications, Network and System Sciences*, 8(5), 118-129.

[8] Zareapoor, M. and Seeja, K.R., 2015. Feature extraction or feature selection for text classification: A case study on phishing email detection. *International Journal of Information Engineering and Electronic Business*, 7(2), 60-65.

[9] Vidros, S., Kolias, C., Kambourakis, G. and Akoglu, L., 2017. Automatic detection of online recruitment frauds: Characteristics, methods, and a public dataset. *Future Internet*, 9(6), https://doi.org/10.3390/fi9010006.

[10] Nasser, I. and Alzaanin, A.H., 2020. Machine learning and job posting classification: A comparative study. *International Journal of Engineering and Information Systems*, 4(9), 6-14.

[11] Dutta, S. and Bandyopadhyay, S.K., 2020. Fake job recruitment detection using machine learning approach. *International Journal of Engineering Trends and Technology*, 68(4), 48-53.

[12] Mahbub, S. and Pardede, E., 2018. Using contextual features for online recruitment fraud detection. *Proceedings of the 27th International Conference on Information Systems Development,* Lund, Sweden, August 22-24, 2018, p. 60.

[13] Alghamdi, B. and Alharby, F., 2019. An intelligent model for online recruitment fraud detection. *Journal of Information Security*, 10(03), 155-176.

[14] Shukla, Y., Yadav, N. and Hari, A., 2019. A unique approach for detection of fake news using machine learning. *International Journal for Research in Applied Science and Engineering Technology*, 7(VI), https://doi.org/10.22214/ijraset.2019.6087.

[15] Akinyemi, B., Adewusi, O. and Oyebade, A., 2020. An improved classification model for fake news detection in social media. *International Journal of Information Technology and Computer Science*, 12(1), 34-43.

[16] Elmurngi, E. and Gherbi, A., 2017. An empirical study on detecting fake reviews using machine learning techniques. *Proceedings of the 7th International Conference on Innovative Computing Technology (INTECH 2017)*, Luton, UK, August 16-18, 2017, pp. 107-114.

[17] Bansal, S., 2020. *Real/Fake Job Posting Prediction*. [online] Available at: https://www.kaggle.com/shivamb/real-or-fake-fakejobposting-prediction.

[18] Kanungo, T. and Orr, D., 2009. Predicting the readability of short web summaries. *Proceedings of the Second ACM International Conference on Web Search and Data Mining*, Barcelona, Spain, February 9-12, 2009, pp. 202-211.

[19] Tsai, Y., 2010. Text analysis of patent abstracts. *The Journal of Specialized Translation*, 13, 61-80.

[20] Fabian, B., Ermakova, T. and Lentz, T., 2017. Large-scale readability analysis of privacy policies. *Proceedings of the International Conference on Web Intelligence*, Leipzig, Germany, August 23-26, 2017, pp. 18-25.

[21] Potthast, M., Kiesel, J., Reinartz, K., Bevendorff, J. and Stein, B., 2017. A stylometric inquiry into hyperpartisan and fake news. *arXiv preprint*, arXiv:1702.05638.

[22] Martinc, M., Pollak, S. and Robnik-Šikonja, M., 2021. Supervised and unsupervised neural approaches to text readability. *Computational Linguistics*, 47(1), 141-179.

[23] Sáez, J.A., Luengo, J., Stefanowski, J. and Herrera, F., 2015. SMOTE–IPF: Addressing the noisy and borderline examples problem in imbalanced classification by a re-sampling method with filtering. *Information Sciences*, 291, 184-203.

[24] Gu, Q., Wang, X.M., Wu, Z., Ning, B. and Xin, C.S., 2016. An improved SMOTE algorithm based on genetic algorithm for imbalanced data classification. *Journal of Digital Information Management*, 14(2), 92-103.

[25] Basgall, M.J., Hasperué, W., Naiouf, M., Fernández, A. and Herrera, F., 2018. Smote-bd: An exact and scalable oversampling method for imbalanced classification in big data. *Proceedings of the VI Jornadas de Cloud Computing and Big Data (JCC&BD 2018),* Buenos Aires, Argentina, June 25-29, 2018, pp. 23-18.

[26] Fernández, A., Garcia, S., Herrera, F. and Chawla, N.V., 2018. SMOTE for learning from imbalanced data: progress and challenges, marking the 15-year anniversary. *Journal of Artificial Intelligence Research*, 61, 863-905.

[27] Yan, Y., Liu, R., Ding, Z., Du, X., Chen, J. and Zhang, Y., 2019. A parameter-free cleaning method for SMOTE in imbalanced classification. *IEEE Access*, 7, 23537-23548.

[28] Rajpura, H.R. and Diwanji, H., 2013. Enhancement of fake website detection techniques using feature selection and filtering algorithms. *International Journal of Advanced Research in Computer Science*, 4(3), 132-137.

[29] Patel, R. and Thakkar, P., 2014. Opinion spam detection using feature selection. *Proceedings of the 2014 International Conference on Computational Intelligence and Communication Networks*, Bhopal, India, November 14-16, 2014, pp. 560-564.

[30] Joshi, A., Pattanshetti, P. and Tanuja, R., 2019. Phishing attack detection using feature selection techniques. *Proceedings of the International Conference on Communication and Information Processing (ICCIP)*, Chongqing, China, November 15-17, 2019, pp. 1-7.

[31] Shabudin, S., Sani, N.S., Ariffin, K.A.Z. and Aliff, M., 2020. Feature selection for phishing website classification. *International Journal of Advanced Computer Science and Applications*, 11(4), 587-595.

[32] Stahl, K., 2018. Fake news detection in social media. *California State University Stanislaus*, 6, 4-15.
[33] Zhou, X., Jain, A., Phoha, V.V. and Zafarani, R., 2020. Fake news early detection: A theory-driven model. *Digital Threats: Research and Practice*, 1(2), 1-25.