Using Contextual Features for Online Recruitment Fraud Detection

Syed Mahbub *S.Mahbub@latrobe.edu.au*
La Trobe University, Department of Computer Science and Information Technology
Melbourne, Australia

Eric Pardede *E.Pardede@latrobe.edu.au*
La Trobe University, Department of Computer Science and Information Technology
Melbourne, Australia

Abstract

The recent growth of online recruitment and candidate management systems has established yet another media for fraudsters on the internet. The ever-growing size of the candidate pool has forced different industries to move to web-based candidate management systems. The advantages of such web-based systems are substantial. On one hand, they are the best means to filter through thousands of applicants for employers and on the other hand, the candidates find themselves in a convenient position while applying for a position. People with fraudulent motivations explore these systems to lure candidates in a hoax and extract sensitive information (e.g. contact information) using fake job advertisements. In this paper, we analyzed a publicly available dataset and used machine learning algorithms to classify job postings as fraudulent or legitimate. The contribution of this research is the inclusion of contextual features in the feature space, which revealed compelling improvements of accuracy, precision and recall.

Keywords: Online recruitment, Fraud detection, Employment scam, Online recruitment fraud, Contextual features.

1. Introduction

Online recruitment fraud (ORF) is one of the most serious problems in recent times on the internet. Although the problem imposes serious threats on personal & social security and privacy, it has not been addressed by the research community to the extent that matches the demand of the severity. ORF is a form of employment scam where a person with fraudulent intensions posts a fake job advertisement on an online platform targeting job seekers. Naïve or desperate job seekers do not think about the legitimacy of the advertisement and end up revealing personal information. This sensitive information is then used by the fraudsters in many ways, compromising the privacy and security of the job seekers. According to an Australian Competition and Consumer Commission (ACCC) report [1], a total amount of **AUD 212,784** in Australia has been lost during the month of October 2017 alone, due to online recruitment frauds. During that period, a frightening number of **237** employment scams were carried out by fraudsters throughout the country.

There are even more severe consequences of ORF than financial loss. Sometimes the fraudsters can even ask for sensitive information to be handed over to the potential employers which can be used to conduct further criminal activities, such as money laundering, reshipping fraud, identity theft and so on. A news report [2] published by Australian Broadcasting Corporation portrays a shocking story of a young man who was convinced enough by a fake job advertisement to send a copy of his passport, driver's license and tax file number to someone he thought, was a potential employer. The information that these documents hold is more than sufficient to conduct an identity theft or a similar crime. Apart from simple contact information, through ORF, people with fraudulent intentions can gather personal information such as home address, educational background, work experience profile and other socioeconomic data and

sell the consolidated dataset to third parties such as call-centres. Some scammers even ask for money from the applicants at a later stage of a fake recruitment process as visa, travel expenses or started-kit expenses.

Due to the structured nature of online recruitment advertisements, it is very hard to distinguish fraudulent job advertisements from legitimate ones. Most of the online recruitment sites have a specific skeleton for job postings that gives the fraudsters an advantage to blend in. Sometimes, the differences are so insignificant that it is very hard to detect the fraudulent advertisement even with human analysis. The lucrative financial offer, flexible work hours can easily trick people into applying for a well-crafter job advertisement. Similar fraudulent online behaviours have been heavily investigated by researchers in the information system domain [3-14]. Email spamming, phishing, cyberbullying, opinion fraud and many more, are problems that are similar to ORF as they can be categorised as improper user behaviours on the web. Nevertheless, the problem of ORF presents some challenges due to the lack of contextual information on the recruitment sites and a very short time span of interaction between the user and the system.

Existing methodologies for detection of ORFs [15,16] utilizes textual and structural information from the job postings but fails to take into consideration, the importance of contextual information about the organization that offers the job. The theoretical and practical importance of contextual features about different actors, have been analysed by researchers in different domains, such as cyberbullying detection [10,11], opinion fraud detection [3] and crowdturfing detection [13,14]. In the case of recruitment fraud, the actor whose contextual information plays an important role in successful detection, is the offering organization. The contextual information includes organization's history, reputation, internet footprint and so on, which are described in the methodology section of this paper. Policy makers such as government organizations also suggest job seekers to validate an organization's legitimacy by gathering these pieces of contextual information [17,18] before applying for a position advertised by the organization. Popular job advertisement portals like *SEEK*[1] also suggest their users to search the internet for a company's footprint before applying for a job.

Keeping these insights in mind, our research focuses on a novel feature space design that not only covers the textual or structural features but also analyses contextual features, in order to improve the detection quality of ORF. The research uses the public EMSCAD dataset [19] and classifies instances of recruitment circulars as fraudulent or legitimate based on the proposed feature space. The learned model is analysed and evaluated using traditional data mining metrics, such as accuracy, precision and recall.

The remaining paper is organized in the following sections: *Section 2* discusses some relevant research works, *Section 3* elaborates the research methodology and proposed feature space, *Section 4* describes the experimental setup and evaluates the proposed model and *Section 5* concludes the paper with remarks on the managerial implications of our work and future improvement scopes.

2. Related work

Research in the field of cybercrime and deception detection are more often than not, domain specific. The generalized area consists of, but not limited to domains, such as phishing, email spam, cyberbullying, Wikipedia vandalism, trolling, opinion fraud, astroturfing, malware attack, cross-site scripting, online predation, financial fraud, identity theft, employment scam and so on. It is outside the scope of this paper to cover previous research works in all of these domains. However, in terms of related approaches, a number of research works [3-16] focused on feature space design, natural language processing (NLP) and machine learning techniques, which we cover in this section.

[1] http://www.seek.com.au

The research work by Stringhini, et al. [3] designed a feature space based on empirical analysis on social networks traits and trained a Random Forest classifier to detect email spam, whereas Boykin and Roychowdhury [4] focused on extracting features from the message body and message header for training a Bayesian classifier. Yeh, et al. [5] on the other hand, focused on meta-heuristics to propose a feature space based on user behavior. The approach of extraction of binary features from online text was adopted by Dinakar, et al. [6] to classify YouTube comments to detect cyberbullying, whereas other researchers used NLP techniques such as Term Frequency–Inverse Document Frequency (TF-IDF), lexical analysis and syntactic-semantic analysis to detect improper behaviors such as phishing, Wikipedia vandalism and cyberbullying [7,8,9]. User contexts of online social network (OSN) such as gender information and user activity history were considered by several research works [10,11,12], in the domain of cyberbullying and trolling. Group and individual contextual characteristics were also utilized by several research works [13,14] to identify crowdturfing groups on OSN. However, these approaches [3-14] of feature space design are not adequate for ORF detection, as the categories of features, included in the feature space, largely depend on specific problem domain. Additionally, the structured nature of online recruitment advertisements begs careful consideration towards feature selection for successful fraud detection.

To the best of our knowledge, in the domain of ORF or employment scam, one research group conducted only two related research [15,16]. The research work done by Vidros, et al. [15] analyzed the problem of employment scam for the first time. The authors explained, in details, the workflow of hiring a candidate and the role of Application Tracking Systems (ATS) within that flow. The authors also mentioned the severity of exploitation of such ATSs. Identity theft, financial loss and loss of privacy were some of the main highlights of their motivation. They drew the differences and discussed the similarities of recruitment fraud with some of the highly studied problem domains such as email spam, cyberbullying, phishing, trolling and Wikipedia vandalism. This research work listed some of the challenges of recruitment fraud domain which include lack of adherence to any communication protocol, short and one-time interaction of users with job advertisements and impersonation of fraudsters as an existing business and so on. Their analysis of real-world workable data generated a set of empirical rules.

In order to prove the hypothesis and applicability of the empirical ruleset, Vidros, et al. [16] conducted a more comprehensive and extensive research. In this second stage of research, the authors generated a real-life dataset of 17,880 instances of job ads where 17,014 were legitimate and 866 were fraudulent. The dataset was made public by the authors and is known as the EMSCAD [19] dataset. The authors conducted bag-of-word (bow) modelling and empirical analysis on a subset of the EMSCAD dataset. Their empirical analysis on geographical constraints, textual analysis of spam words, analysis of HTML elements and binary analysis provided affective baseline information for ORF detection. The empirical ruleset generated in the first paper [15] was expanded in the second [16] and served as a base for the ruleset based binary features. They achieved a highest accuracy of 90.56%, recall value of 0.906 and precision value of 0.906 using J48 decision tree classifier of *WEKA* for their binary analysis of features.

Although Vidros, et al. [16] analyzed different aspects of the problem domain by taking different approaches to model a solution for detection, the contextual features indicating organization's online profile were not considered by the authors. Also, due to the short span of interaction between the user and the online recruitment systems, many aspects of the recruitment advertisement itself, are usually ignored by the regular job seekers. To overcome these challenges, our research considers the contextual attributes that gives an overall idea about the advertising organizations, which are imperative to design a better detection tool. The next section elaborates our research methodology for designing a novel feature space that takes into account the contextual information about the organization's background.

3. Research methodology

Our research is mainly driven by the motivation of designing a feature space that can train a learning model to detect an instance of job circular or advertisement more accurately. As discussed in the related work section, to the best of our knowledge, the closest relevant research we could find in the domain was the research conducted by Vidros, et al. [16]. This research makes use of the EMSCAD [19] dataset that was made public by the authors of the paper [16]. The dataset contained 17,880 real-life job ads posted by Workable [20], a renowned online recruitment portal. Each instance of the public dataset is labelled. Each record in the dataset had a set of attributes and a binary class label indicating whether or not an instance of a job circular is fraudulent. For each record in the dataset, a class label 't' indicates a fraudulent job posting whereas, class label 'f' indicates that the posting is a legitimate job advertisement. Apart from the class label, there are 16 attributes pertaining to each record in the public dataset. The values for these attributes for each record are either pure text, or text with HTML tags. Table 1 lists down all the fields and their short descriptions.

Table 1. Description of attributes in the EMSCAD dataset.

Name	**Description**
Title	The title of the job circular
Location	The geographic location of the job
Department	Internal department of the organization
Salary range	Indicative salary range
Company profile	A brief profile of the company
Description	A detail description of the job
Requirements	Requirements of the job
Benefits	Offered benefits of the job
Telecommuting	True/False based on whether or not the job requires telecommunication
Company logo	True/False based on whether or not the ad contains company logo
Questions	True/False based on the presence of screening questions
Employment type	Type of employment (full-time, part-time, contract, etc.)
Required experience	Level of experience (Executive, entry level, intern, etc.)
Required education	Level of education (Master's, bachelor's degree, etc.)
Industry	Specific industry (IT, healthcare, etc.)
Function	Specific area of functionality (Engineering, research, sales, etc.)
Fraudulent	Class label ('f' means legitimate job ad and 't' means fraudulent job ad)

In our research, the principle idea behind the design of contextual feature space was to mimic the human behavior while trying to validate the authenticity of a job circular. An initial analysis of the dataset fields reveals that, it is quite challenging for a job seeker to decide whether or not a circular is fraudulent just by looking into the field values, which are in terms, a direct representation of the real-life job ads. Further investigation of the available government and private online articles revealed some insight. For example, an article published by Australian Government [17] strongly advices job seekers to search the internet for the company website and other information they can find about the company that indicates towards the reputation and dependability of the company as an organization. In Australian scenario, if the job posting is made by an organization inside Australia, the Government also suggests checking the Australian Business Register [18] for the company details. The idea is to have a way to validate the existence of the company's internet footprint. Also, instinctively, when we need to find out something about an entity, whether or not the entity is an organization, the first place we look for is the internet. Given the information availability provided by the search engines, the existence of a company can be validated by a human in several minutes who has basic knowledge of using the internet. If the corresponding employer of a job ad turns out to be an invalid company, it is almost certain that the job ad is also invalid.

In order to acquire information about a company, the first thing we needed was the name of the company. The EMSCAD [19] public dataset does not contain such information. Our

approach to resolve this problem involved two significant steps. First, to extract a name for a company from the text listed as company profile in the dataset. Second, to extract some contextual information about the company using simple Google search.

The problem with the implementation of the idea is that the company name extraction from text requires significant natural language processing efforts, as Named Entity Recognition (NER) algorithms perform well with human names, not organization names. Moreover, due to updated privacy policies of Google, Google Search API [21] now, has restricted flexibility and lets the developer create search engines that can search only within a specific website, not the entire web. Keeping these challenges in mind, we decided to keep the extraction process of contextual features as manual and also keep a manual validation process for the output of the name extraction algorithm, as a safeguard. The automation of contextual feature extraction will be the primary objective of future extension of this work.

Due to the fact, that the manual extraction of features from the web requires significant time and efforts, the convenient option was to work with a subset of the EMSCAD [19] dataset. One major criteria for selecting the records from the entire dataset of 17,880 instances was the presence of company profile as the name extraction algorithm was designed to work with company profile text. The pre-processing program extracted a small subset of records and for each record in the subset, the company profile field in the dataset was *not* empty. The *Java Collections* class was used to randomise the records. The ratio between the positive and negative instances was also kept within the 4:1 range. As a result, we ended up with a dataset of **368** instances among which, **94 (~25%)** instances were fraudulent records and **274** instances were legitimate job ads.

In order to extract the name of the company, we used the Stanford CoreNLP [22] natural language processing toolkit. The algorithm first parsed the HTML company profile using Jsoup [23] and then used a combination of the Named Entity Recognition (NER) and Parts of speech (POS) tagging libraries from the CoreNLP toolkit to come up with a name of the company. Manual validation revealed that around 85% of the company names were extracted correctly among the instances that had a valid company name within the profile text.

Before the population process of binary contextual features for our dataset, the company website URL (if any), domain age, LinkedIn page URL for the company were extracted manually for each of the records. The domain age was extracted using open source domain age tool [24] which indicated how long the domain has been occupied by the company. The domain age provided an indicative baseline to decide whether or not the website was recently created. The rational for keeping the website age information was the tendency of the fraudsters to create fake websites just before posting a fake job advertisement. Nowadays, the ease of creating websites in a few clicks has increased the number of such fake websites where the company does not actually exist, but the website does. Several news reports and government agencies list these sorts of discrepancy [25,26]. Finally, to keep a track record of the fraudsters, the company profile texts of fraudulent job ads were kept in a look-up table as fraudsters often tend to use the same company profile for posting different fraudulent job advertisements. Fig. 1 shows a summary of our contextual attribute extraction process.

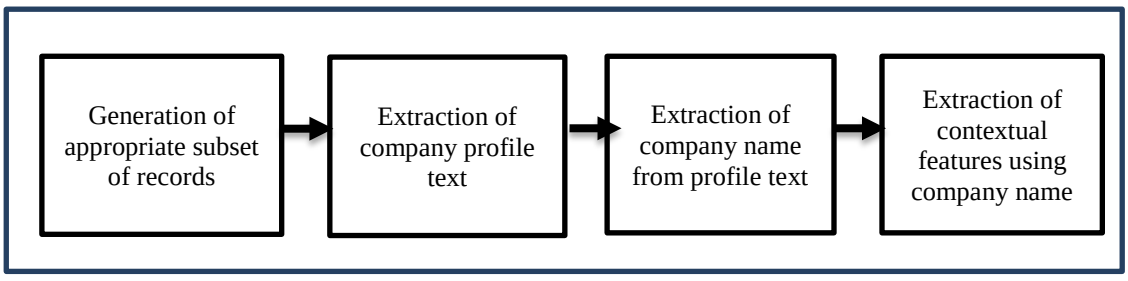

Fig. 1. Contextual feature extraction process.

Based on gathered knowledge, discussions and feasibility analysis, we decided to include the following binary features in our contextual feature space as part of our novel feature space design:

- *hasCompanyName*: 1 if the name of the company is present in the profile text, 0 otherwise.
- *hasCompanyWebsite*: 1 if the company has a valid website, 0 otherwise.
- *hasMaturedCompanyWebsite*: 1 if the company website is older than 1 year, 0 otherwise.
- *hasLinkedInPage*: 1 if the company has a valid LinkedIn page, 0 otherwise.
- *previouslySeenAsFraudulent*: 1 if the job ad contains a company profile that was used in a previously found fraudulent job ad within the dataset, 0 otherwise.

Apart from the above contextual features, other binary features were also considered based on previous research of Vidros, et al. [16]. The features in the entire feature space were divided into three major categories to differentiate between classes of features more precisely. These categories are, *textual features*, *structural features*, *contextual features*, where the contextual features are the highlights of the novelty of this work. The details of these features are discussed in the experimentation section.

The machine learning toolkit used to conduct experiments was *WEKA: Waikato Environment for Knowledge Analysis* [27]. Once the dataset was mapped into the binary feature space, several classification algorithms, such as J48 decision tree, JRip rule-based classifier, Naïve Bayes classifier were used from the *WEKA* toolkit. It is imperative to notice that the idea behind using several classification algorithms was not to identify a superior one that fits the problem domain better, but to validate the consistency and improvement of performance measures across different experimental setup. Accuracy, precision and recall were the three metrics that was used to measure the performance of the model for different classification algorithms. The details about the performance of each of the classification algorithm are also listed in the experimentation and evaluation section.

4. Experimentation and evaluation

The experiments were conducted in two phases for each classification algorithm used. Each phase was given a particular name to better differentiate between the obtained performance measures. The difference between the two phases was the feature space of the training dataset. **Experiment A** contained textual and structural features in the dataset of 368 instances, whereas **experiment B** contained contextual features along with textual and structural features for the same dataset. Apart from the difference in feature space, the experimental environment, i.e., algorithm used, algorithm parameters, were kept constant across the two experiments, for each classification algorithm.

Table 2. Proposed binary feature space.

Feature Class	Name	Description	Included in experiments
Textual	containsSpamWord?	1 if the title or description contains a spam word such as “easy job”, “work from home”, etc. 0 otherwise	A, B
	hasConsecutivePunctuation?	1 if title or description contains consecutive punctuations (two or more ‘!’), 0 otherwise	A, B
	hasMoneyInTitle?	1 if title contains “money”, “cash”, etc. 0 otherwise	A, B
	hasMoneyInDescription?	Same as above	A, B

	hasExternalPrompt?	1 if description or requirement contains external prompts such as “follow the link”, “send resume at”, etc. 0 otherwise	A, B
	hasNonOrgEmailLinks?	1 if description or requirement contains personal email links, 0 otherwise	A, B
	isTelecommuting?	1 if telecommuting is true, 0 otherwise	A, B
	hasConsecutiveCappitalLetter?	1 if title or description contains consecutive capital letters (10 or more), 0 otherwise	A, B
	educationLevelLow?	1 if required education level is high school or equivalent, 0 otherwise	A, B
	hasBoldTextInDesciption?	1 if bold text is present in description, 0 otherwise	A, B
	hasBoldTextInBenefits?	Same as above	A, B
	basedInUS?	1 if location text contains US, 0 otherwise	A, B
Structural	hasCompanyProfile?	1 for each record of our version of the dataset	A, B
	hasCompanyLogo?	1 if company logo is present, 0 otherwise. Information is directly acquired from EMSCAD dataset	A, B
	hasJobIndustry?	1 if job industry is specified, 0 otherwise	A, B
	hasScreeningQuestion?	Same as above	A, B
	hasJobDescription?	Same as above	A, B
	hasSkillRequirement?	Same as above	A, B
	hasBenefits?	Same as above	A, B
	hasShortDescription?	1 if description is less than 60 words (4 standard sentences), 0 otherwise	A, B
	hasShortCompanyProfile?	Same as above	A, B
	hasShortRequirements?	Same as above	A, B
	hasShortBenefits?	Same as above	A, B
	hasHTMLListInRequirements?	1 if HTML list element is present in requirement, 0 otherwise	A, B
	hasHTMLListInBenefits?	Same as above	A, B
Contextual	hasCompanyName?	1 if profile text contains company name, 0 otherwise	B
	hasCompanyWebsite?	1 if company has a valid website, 0 otherwise	B
	hasMaturedWebsite?	1 if domain age is greater than 1 year, 0 otherwise	B
	hasLinkedInPage?	1 if company has a LinkedIn page, 0 otherwise	B
	previouslySeenAsFraudulent?	1 if a fraudulent ad was seen with the same profile text, 0 otherwise	B

The pre-processing object-oriented program mapped each record of the dataset into a binary feature vector. These feature vectors were then fed to each of the classification

algorithms with a 10-fold cross validation. The cross-validation process randomises the training and test set for each of the folds. Table 2 lists the entire feature space totalling 30 binary features including textual, structural and contextual features along with the information about the corresponding experiments they were used in. The table does not list the binary class label.

For J48 decision tree classification, a 10-fold cross validation process with the training dataset revealed an accuracy of **79.62%** for experiment **A** (without contextual features). The precision and recall values for class label 't' (fraudulent) were **0.651** and **0.436** respectively. On the other hand, for J48 decision tree classification, a 10-fold cross validation with the same training dataset yielded an accuracy of **94.29%** for experiment **B** (with contextual features in the feature space). The precision and recall values for class label 't' (fraudulent) were **0.910** and **0.862** respectively. Experiments conducted with JRip rule-based classifier and Naïve Bayes classifier also demonstrated significant improvement of performance measures for experiment **B** over experiment **A**. Table 3 lists the confusion matrix for experiments A and B for all three classifiers.

Table 3. Confusion matrix for experiments A and B for all three classifiers.

Actual class	Predicted class in experiments			
	J48 experiment A		J48 experiment B	
	f (legitimate)	t (fraudulent)	f (legitimate)	t (fraudulent)
f (legitimate)	252	22	266	8
t (fraudulent)	53	41	13	81
	JRip experiment A		JRip experiment B	
f (legitimate)	246	28	268	6
t (fraudulent)	57	37	8	86
	Naïve Bayes experiment A		Naïve Bayes experiment B	
f (legitimate)	236	38	221	53
t (fraudulent)	59	35	8	86

As it can be seen from the confusion matrix, for J48 (experiment A), a total of **53** fraudulent job advertisements were incorrectly classified as legitimate (false negative) whereas, **22** instances of legitimate job postings were incorrectly classified as fraudulent (false positive). The confusion matrix explains the low precision and recall values for class label 't' for experiment A. For experiment B with J48, a total of **13** fraudulent job advertisements were incorrectly classified as legitimate (false negative) whereas, only **8** instances of legitimate job postings were incorrectly classified as fraudulent (false positive).

For experiments with J48 decision tree the number of false negatives was reduced from **53** to **13** from experiment A to experiment B. On the other hand, the number of false positives was reduced to **8** in experiment B from **22** in experiment A. Moreover, the accuracy increased up to **94%** in experiment B from only **79%** in experiment A. The value of precision and recall for class label 't' also increased from **0.651** to **0.910** and **0.436** to **0.862**, respectively. These measures clearly demonstrate that the inclusion of contextual features such as company's website information, existence of LinkedIn page and so on, makes a significantly positive impact on the outcome of the model and detects more fraudulent instances compared to the feature space without them. Fig. 2 illustrates the differences in performance measures obtained in experiments A and B across different classifiers.

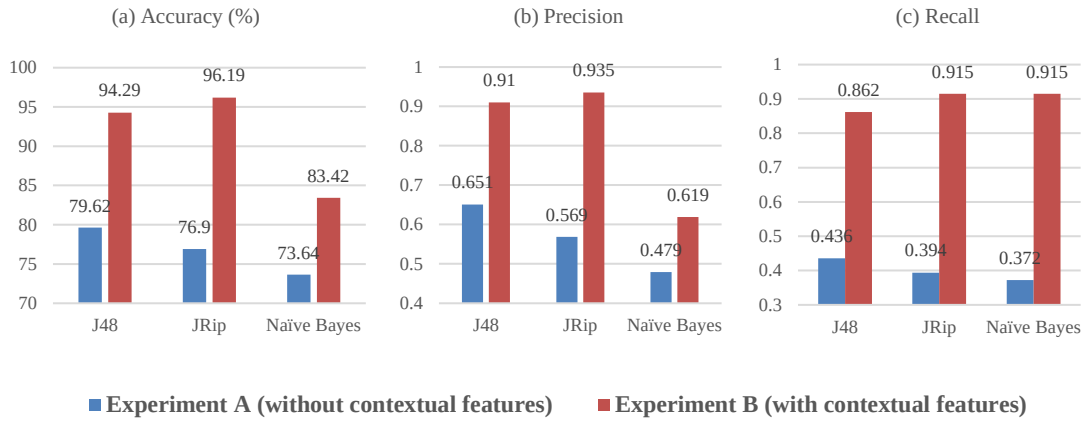

Fig. 2. Difference in performance measures for experiment A and B across three classifiers. (2a) Accuracy, (2b) Precision and (2c) Recall

For both JRip rule-based classifier and Naïve Bayes classifier the accuracy, precision and recall values also improved significantly from experiment A to experiment B. JRip performed the best among the three classifiers yielding an accuracy of **96%** for experiment B and Naïve Bayes showed the least impressive results for experiment B yielding an accuracy of **83%**, although it is a major improvement over **73%** of accuracy in experiment A. The precision values for class label 't' improved from **0.569** to **0.935** and **0.479** to **0.915** for JRip and Naïve Bayes, respectively. Similarly, the recall value for class label 't' increased from **0.394** to **0.915** and **0.372** to **0.915** for JRip and Naïve Bayes, respectively.

The performance measures of classifiers used in this research cannot be directly compared with the performance measures of the similar research done by Vidros, et al. [16], as the experiments use different subsets of the EMSCAD dataset and the pre-processing techniques are bound to be different. Hence, we evaluate our model based on the performance measures of two different experiments (A and B) as the differences clearly demonstrates the importance of contextual features in the feature space.

As mentioned earlier in the research methodology section, the purpose of conducting experiments using different classification algorithm was not to identify a superior one for the problem domain, rather to present a scenario that demonstrates the consistent improvement of performance of classifiers, when the contextual features are included in the feature space. The results obtained by conducting the incremental feature space expansion from experiment A to experiment B clearly manifests our claim that the contextual feature space improves the detection performance of online recruitment fraud.

5. Conclusion and future work

The cloud-based recruitment platforms are one of the most used platforms on the internet. With the ever-increasing number of users of these platforms, personal and financial risks are increasing as well, making the platform vulnerable to threats. Although extensive work has been done in similar areas of research to study fraudulent user behavior (as discussed in the related work section), the domain of ORF needs considerable attention from the research community due to the gravity of impacts it bears on privacy and security. The limited information within the scope of a recruitment advertisement itself, also poses a challenge. New directions need to be explored to enrich the collection of features that are considered by a learning algorithm to successfully detect a fraudulent job advertisement.

In this paper, we have proposed a novel feature space to improve the detection accuracy of ORF. Previous works [15,16] in the domain did not address any contextual features outside the scope of a job advertisement itself. Our proposed feature space considered these contextual features which resulted in a significant increase in terms of accuracy, precision and recall of multiple classifiers. The proposed feature space has been well-structured into different feature

classes which makes the future extensions easy and manageable. The introduction of such contextual features can potentially pave way for a new dimension of research in the domain of ORF detection.

The methodology proposed in this paper can have a major impact on numerous job advertisement portals that exist in the present era of internet. If more intelligent backend filtering systems can be implemented on these portals, based on the research methodology proposed in this study, more fraudulent jobs will be detected in the filtering stage. This will prevent regular users from accidentally revealing their personal information by applying for jobs that do not actually exist. Hence, to mitigate the risk of privacy and security, it is essential to develop intelligent systems embedded within these cloud-based recruitment platforms. Our research takes a step closer towards such intelligent filtering system.

The future improvement scopes for our research includes the extension of feature space with more classes of features. However, the primary objective for our future extension would be to automate the extraction process of contextual features by designing custom search engines. Once the contextual feature extraction process is automated, it will then be feasible to validate the importance of contextual features more accurately using the entire EMSCAD dataset. Future automation process will also address the threats of validity of our proposed feature space. Network information and user behavior analysis can add more value in the future works as well. Moreover, the generation of large scale data sets with network and user level information can facilitate further research, which is essential to tackle this relentlessly increasing problem.

References

1. Scamwatch Jobs & employment (2015), https://www.scamwatch.gov.au/types-of-scams/jobs-investment/jobs-employment. Accessed April 12, 2018
2. Scammers target jobseekers on recruitment websites – ABC News (2016), http://www.abc.net.au/news/2016-07-24/hold-sun-job-seekers-being-scammed-on-legit-websites/7654804. Accessed April 13, 2018
3. Stringhini, G., Kruegel, C., Vigna, G.: Detecting spammers on social networks. In: Proceedings of the 26th Annual Computer Security Applications Conference, pp. 1-9. ACM, Austin (2010)
4. Boykin, P., Roychowdhury, V.: Leveraging social networks to fight spam. Computer. 38 (4), pp. 61-68 (2005)
5. Yeh, C.Y., Wu, C.H., Doong, S.H.: Effective spam classification based on meta-heuristics. In: Proceedings of the 2005 IEEE International Conference on Systems, Man and Cybernetics, pp. 3872-3877. IEEE, Waikoloa (2005)
6. Dinakar, K., Reichart, R., Lieberman, H.: Modeling the Detection of Textual Cyberbullying. The Social Mobile Web. 11 (2), pp. 11-17 (2011)
7. Zhang, Y., Hong, J. I., Cranor, L. F.: Cantina: a content-based approach to detecting phishing web sites. In: Proceedings of the 16th international conference on World Wide Web, pp. 639-648. ACM, Banff (2007)
8. Wang, W., McKeown, K.: Got you!: automatic vandalism detection in Wikipedia with web-based shallow syntactic-semantic modeling. In: Proceedings of the 23rd International Conference on Computational Linguistics, pp. 1146-1154. Association for Computational Linguistics, Beijing (2010)
9. Chen, Y., Zhou, Y., Zhu, S., Xu, H.: Detecting offensive language in social media to protect adolescent online safety. In: Proceedings of the Privacy, Security, Risk and Trust (PASSAT), 2012 International Conference on and 2012 International Conference on Social Computing (SocialCom), pp. 71-80. IEEE, Amsterdam (2012)
10. Dadvar, M., Jong, D., Ordelman, R., Trieschnigg, D.: Improved cyberbullying detection using gender information. In: Proceedings of the Twelfth Dutch-Belgian Information Retrieval Workshop (DIR 2012), pp. 23-25. Ghent University, Ghent (2012)

11. Dadvar, M., Trieschnigg, D., Ordelman, R. de Jong, F.: Improving cyberbullying detection with user context. In: Proceedings of the European Conference on Information Retrieval, pp. 693-696. Springer, Berlin (2013)
12. Cheng, J., Danescu-Niculescu-Mizil, C., Leskovec, J.: Antisocial Behavior in Online Discussion Communities. In: Proceedings of the 9th International AAAI Conference on Web and Social Media, pp. 61-70. AAAI, Oxford (2015)
13. Lee, K., Webb, S., Ge, H.: Characterizing and automatically detecting crowdturfing in Fiverr and Twitter. Social Network Analysis and Mining. 5 (2), pp. 1-16 (2015)
14. Xu, C., Zhang, J., Chang, K., Long, C.: Uncovering collusive spammers in Chinese review websites. In: Proceeding of the 22nd ACM international conference on Information & Knowledge Management, pp. 979-988. ACM, San Francisco (2013)
15. Vidros, S., Kolias, C., Kambourakis, G.: Online recruitment services: another playground for fraudsters. Computer Fraud & Security. 2016 (3), pp. 8-13 (2016)
16. Vidros, S., Kolias, C., Kambourakis, G., Akoglu, L.: Automatic Detection of Online Recruitment Frauds: Characteristics, Methods, and a Public Dataset. Future Internet. 9 (1), p. 6 (2017)
17. Stay Smart Online (2016), https://www.staysmartonline.gov.au/alert-service/fake-job-ads-designed-steal-your-identity-and-money. Accessed: April 13, 2018
18. Australian Business Register (2015), https://abr.gov.au/. Accessed: April 13, 2018
19. EMSCAD Dataset (2017), http://emscad.samos.aegean.gr/. Accessed: April 12, 2018
20. Workable (2016), https://www.workable.com/. Accessed: April 12, 2018
21. Google Custom Search (2012), https://developers.google.com/custom-search/. Accessed: April 13, 2018
22. Manning, D., Christopher, S., Mihai, B., John, F., Jenny, J., Steven, B., David, M.: The Stanford CoreNLP Natural Language Processing Toolkit. In: Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics, pp. 55-60. ACL, Vancouver (2014)
23. Jsoup HTML parser (2010), https://jsoup.org/. Accessed: April 13, 2018
24. Domain age tool (2017), https://www.webconfs.com/web-tools/domain-age-tool/. Accessed: January 11, 2018
25. Scam employment website targets jobless miners – ABC News (2016), http://www.abc.net.au/news/2016-11-30/jobless-miners-targeted-in-online-scam/8079304. Accessed: April 13, 2018
26. WAScamNet: Mining Recuirtment Scan – Government of Western Australia (2012), http://www.scamnet.wa.gov.au/scamnet/Scam_types-Jobs__Investment-Jobs__Employment-Mining_recruitment_scam.htm. Accessed: April 13, 2018
27. Frank, E., Hall, M. A., Witten, I.H.: The WEKA Workbench. In: Online Appendix for Data Mining: Practical Machine Learning Tools and Techniques, Morgan Kaufmann (2016)