AI & SOCIETY (2023) 38:1013–1024
https://doi.org/10.1007/s00146-022-01469-0

STUDENT FORUM

A machine learning approach to detecting fraudulent job types

Marcel Naudé[1] · Kolawole John Adebayo[2,3] · Rohan Nanda[4,5]

Received: 30 July 2021 / Accepted: 13 April 2022 / Published online: 25 May 2022

Abstract
Job seekers find themselves increasingly duped and misled by fraudulent job advertisements, posing a threat to their privacy, security and well-being. There is a clear need for solutions that can protect innocent job seekers. Existing approaches to detecting fraudulent jobs do not scale well, function like a black-box, and lack interpretability, which is essential to guide applicants' decision-making. Moreover, commonly used lexical features may be insufficient as the representation does not capture contextual semantics of the underlying document. Hence, this paper explores to what extent different categorizations of fraudulent jobs can be classified. In addition, this paper seeks to find what type of features are most relevant in classifying the type of fraudulent job. In this paper, we develop and validate a machine learning system for identifying identity theft, corporate identity theft and multi-level marketing amongst fraudulent job advertisements. We utilized four classes of features: empirical rule set-based features, bag-of-word models, most recent state-of-the-art word embeddings and transformer models for various machine learning classifiers. The machine learning models were validated by evaluating them on a publicly available job description dataset. Our results indicate that the word embeddings and transformer-based features consistently outperformed the handcrafted rule-set based features class. Ultimately, a Gradient Boosting classifier with a combination of empirical rule-set based features, parts-of-speech tags and bag-of-words vectors achieved the best performance with an F1-score of 0.88.

Keywords Fraud detection · Online recruitment fraud · word2vec · Transformers · Machine learning · Natural language processing

1 Introduction

Developments in technology and big data have ushered in a new era of Human Resources where services increasingly take place online (Sivathanu and Pillai 2018; Baykal 2020). Particularly, the use of Applicant Tracking Systems (ATS) to conduct e-recruitment has gained popularity among HR practitioners and recruiters. Job seekers find themselves increasingly duped and misled by fraudulent job advertisements, posing a threat to their privacy, security and well-being (Mahbub and Pardede 2018). Fraudsters are able to obtain sensitive information such as full names, addresses, contact information, social security numbers, and so forth (Vidros et al. 2016). Likewise, legitimate organizations are at risk of facing harm to their reputation and a smaller applicant pool as a consequence of such activities (Vidros et al. 2017).

Given these costs to society, there is a clear need for solutions that can protect the safety of stakeholders in the recruitment process. So far, this has been treated as a binary classification problem where researchers attempted to build a solution that can detect fraudulent job postings. However, Vidros et al. (2016) pointed out that there is room for gaining deeper knowledge of the different *characteristics* of the online recruitment fraud (ORF) problem. Beyond the simple

✉ Marcel Naudé
ma.naude@alumni.maastrichtuniversity.nl

Kolawole John Adebayo
kolawole.adebayo@adaptcentre.ie

Rohan Nanda
r.nanda@maastrichtuniversity.nl

[1] School of Business and Economics, Maastricht University, Maastricht, The Netherlands

[2] Dublin City University, Dublin, Ireland

[3] ADAPT Centre, Dublin, Ireland

[4] Maastricht Law and Tech Lab, Maastricht University, Maastricht, The Netherlands

[5] Institute of Data Science, Maastricht University, Maastricht, The Netherlands

detection, organizations are in need of transparent and interpretable explanations to improve decision-making throughout the job syndication process. Practitioner and job seeker access to tools that can filter out categories of fraudulent job postings per category may inform proactive and reactive measures to minimizing the negative social impact of ORF, such as loss of privacy, money, talent or reputation (Vidros et al. 2017).

In this paper, we develop and validate a machine learning system for identifying different categories of fraudulent job advertisements. Three distinct types of fraudulent jobs were conceptualized after a thorough analysis of the literature: identity theft, corporate identity theft, and pyramid schemes or multi-level marketing. We go beyond the existing handcrafted and bag-of-words features used by existing ORF works (Vidros et al. 2017; Mahbub and Pardede 2018; Alghamdi and Alharby 2019) and introduce most recent state-of-the-art word embeddings and transformer models features for various machine learning classifiers. The feature classes were validated by evaluating it on the publicly available dataset. The code is available at the listed repository. Our research questions are as follows:

1. What lexical, syntactic, semantic or contextual features can distinguish different types of recruitment fraud in the EMSCAD dataset?
2. Which features are most informative in predicting the type of fraudulent job?
3. What machine learning algorithm performs the best for fraudulent job type classification?

The paper is structured as follows. First, a literature review is conducted that explores the extant research stream on ORF classification. Second, the methodology and research design is introduced in the context of the research questions, and the dataset is presented. Third, we present the research framework which has been implemented and experiments carried out on. The results of different classification models are compared, evaluated and interpreted. Fourth, the findings are discussed with a focus on the limitations, and subsequently the appropriate implications are outlined.

2 Literature review

The first mention of ORF was in the 2016 paper by Vidros, Kolias and Kambourakis. It was introduced as an emerging, pressing matter that is worthy of research efforts; they called attention to how scammers increasingly exploit job posting platforms to harvest personal information for malevolent purposes. The authors attributed the increase in incidence of such activities to a more digitized and cloud-based hiring process, and suggested that further uptake of digital recruitment will coincide with even more ORF.

In 2017, Vidros, Kolias, Kambourakis and Akoglu published a follow-up to the earlier paper, wherein they demonstrated the performance of a machine learning classifier on a corpus of fraudulent job ads. They compared a bag-of-words model and a handcrafted binary rule-set based features model, achieving 91% accuracy in predicting fraudulent jobs with a Random Forest classifier using the latter model. To spur further research on this issue, they made public the Employment Scam Aegean Dataset (EMSCAD). The authors noted that future work could be focused on expanding the dataset, expanding the feature space, and using different techniques to model the data.

In the subsequent years, several researchers have responded to this call for more research by expanding on the work done by the original authors. The majority of this work was validated on the same EMSCAD dataset, and focused on two key issues: improving the prediction performance by experimenting with different classifiers (Lal et al. 2019; Dutta and Bandyopadhyay 2020; Anita et al. 2021), and investigating whether the selection of features can be improved or expanded to build more robust models (Mahbub and Pardede 2018; Alghamdi and Alharby 2019; Mehboob and Malik 2020).

2.1 Online recruitment fraud types

Thus far, research has only treated the problem of ORF as a dichotomy: either a job is fraudulent, or it is not. While the classification of a fraudulent job on its own can be valuable for practitioners or ATS stakeholders, it is in the interest of the public to also better understand what the fraudster might be trying to achieve, as such awareness can better direct precautionary and/or reactive measures to fraud that has been detected.

Vidros et al. (2016), who differentiated between two groups of fraudulent jobs. The first group comprises advertisements for non-existing jobs which aim to harvest personal information such as names, phone numbers, and e-mail addresses. Such information may then be sold to third parties or used as targets for spam emails and spam calls. The second group of fraudulent jobs consists of attempts to social engineer either highly sensitive information out of the job seeker, such as social security numbers or passports, or lure the job seeker into depositing sums of money. A malicious actor may engage in behaviors that imitate a legitimate job hunting process, including adopting the identity of a legitimate employer or scheduling interviews and assessments. The key difference between these two groups is the severity of actions or steps taken by the fraudster to exploit the job syndication process.

Despite the scarcity of research, various public sources can be consulted to form a better perspective on what sets fraudulent job advertisements apart. Reynolds (2021) wrote about nine different types of job search scams, which differ in the type of actions it demands from the job seeker. JobHunt, a career advice website, gives an example of 'corporate identity theft', which are fraudulent jobs that claim to be from a real employer (Joyce 2021). Without being able to verify the true identity of the job poster, it becomes even more challenging for the job seeker to ensure their online safety.

Governmental organizations have also taken interest in informing the public about different recruitment fraud. Scamwatch, a service by the Australian government, highlights a form of job fraud where job seekers are requested to receive and/or send money in exchange for a commission payment (Scamwatch 2015). Similarly, the United States (US) Securities and Exchange Commission points out multi-level marketing programs that are disguised as legitimate business opportunities in job advertisements; such programs, which stress the earning of passive income and high returns, and emphasizes recruitment or referral of others, can be illegal and highly costly to the victim (SEC 2013). The US Federal Trade Commission outlines reshipping scams, reselling merchandise scams, mystery shopper scams and more, all of which are recruitment fraud that manifest in different forms (FTC 2021).

3 Research methodology

In this section, the methodology of the paper is explained. First, the research design is outlined. Next, the relevant dataset is briefly introduced. Thereafter, the data annotation and data preparation processes are described. Lastly, the feature extraction and classification methods are thoroughly explained.

3.1 Research design

The higher-level research design of this paper is rooted in the Knowledge Discovery from Data (KDD) model (Alghamdi and Alharby 2019), which consists of several consecutive processes that structure the extraction of information from data. Firstly, the target dataset is acquired. Second, the data are anonymized, annotated and pre-processed to prepare for analysis and transformation. Third, feature extraction and selection is performed for four classes of features. For each model developed per feature class, several multi-class classification algorithms are trained on the model to classify the type of fraudulent job, and the results are evaluated with respect to the research objectives (Fig. 1).

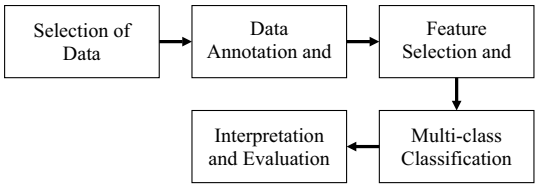

Fig. 1 Research design

3.2 Selection of data

The data used to fulfill the research objectives are based on the Employment Scam Aegean Dataset. (EMSCAD) dataset made publicly available by Vidros et al. (2017), which consists of 17,880 job vacancies collected between 2012 and 2014. These vacancies were collected by Vidros et al. from and in co-operation with the job advertisement platform Workable, with annotation done by specialized Workable employees according to company policy and meticulous analysis (Vidros et al. 2017). The personal information in job advertisements was anonymized to conform to GDPR regulations.

3.3 Data annotation

A new categorical numeric variable *'type'* was created. There are four possible values of this variable: real job (0), identity theft (1), corporate identity theft (2), and multi-level marketing (3). These categories are based upon the discussion in the literature review. Type one and two both aim to harvest sensitive personal information from the job seeker, while the latter aims to be indistinguishable from real job advertisements by adopting the identity of a legitimate job provider. Although the literature review had highlighted the existence of more than three types of fraudulent job advertisements, an empirical analysis of the EMSCAD dataset revealed a relatively low diversity and number of fraudulent jobs, which did not allow for a very high granularity in distinguishing between these job types.

Since the research objectives ordinarily concern the types of *fraudulent jobs*, all 866 fraudulent jobs from the EMSCAD dataset were isolated and individually annotated. To determine which category a given job advertisement belongs to, the following criteria were developed in collaboration with a Human Resource Management (HRM) industry expert and subsequently applied to annotate the job's company profile, description, requirements and benefits text fields.

1. **Identity theft**

(a) Explicit requests of personal information (e.g., full name, address, phone number) outside of typical submission of CV
(b) Referral to an external web-page or e-mail address to finish application

2. **Corporate identity theft**

(iii) Fraudulent job ad which claims to be from a legitimate, reputable company
(iv) It implies that the legitimate company is somehow involved with the posting of the job ad (e.g., through partnership or association)

3. **Multi-level marketing**

(e) Job ad incentivizes or requests applicant's involvement in a referral/commission-based scheme (get paid for successfully sharing the job ad with someone else)
(f) Content of the job advertisement describes activities where the benefit that the targeted candidate accrues is a result of the number of additional candidates they recruit, sign up or refer

3.4 Data preparation

A random sample of 866 real jobs were merged into the annotated fraudulent job dataset and assigned a *type* value of 0. The choice of 866 follows from sampling the same amount of real jobs as there are fraudulent jobs in the dataset. There were only 72 of type-3 jobs compared to 556 type-1 jobs. To mitigate class imbalance, type-3 jobs were randomly upsampled using resample with replacement method. The end result of the annotation process is a dataset with 556 real jobs, 556 type-1 jobs, 234 type-2 jobs, and 150 type-3 jobs.

The process of cleaning the text consisted of the following steps. First, the four text fields of the EMSCAD dataset—'company_profile', 'requirements', 'benefits' and 'description'—were merged into a single column called 'text' to simplify processing and transformation. Subsequently, the four columns were dropped from the dataset as they are no longer needed. Thereafter, a processing function was applied to the text fields for tokenization, stopword, removal of numbers and punctuation, removal of non-English words, and lower case conversion.

3.5 Feature extraction and classification

Four distinct classes of features are investigated: bag-of-words (BOW) model, empirical ruleset, word embeddings and transformer models. The dataset was split into an 80/20 training and testing set with stratified sampling. With the target variable of 'type', each feature class configuration is evaluated against the test set on several metrics: precision, recall, F-score and Matthews correlation coefficient (MCC).

This paper will use a selection of standard supervised classification algorithms, both single and ensemble, to test all the feature classes: Logistic Regression (LR), Stochastic Gradient Descent (SGD), k-Nearest Neighbors (KNN), Decision Tree (CART), Support Vector Machine (SVM), Random Forest (RF), AdaBoost (AB), and Gradient Boosting (GB). These classifiers were implemented with the scikit-learn library on Python.

3.5.1 Feature class 1: bag-of-words models

The first set of features in this class is a representation of the job advertisements simply based on the occurrence count of its words. This BOW transformation is implemented on the cleaned and pre-processed text. The total size of the vocabulary and hence the feature space is 6356 words after stopword removal and other processing steps.

The second set of features in this class, term frequency-inverse document frequency (tf-idf), follows a similar approach to the bag-of-words model. Here, the model also incorporates the importance of each word in a job advertisement as it relates to the entire corpus of text. A Tf-Idf vectorizer is fit on the job advertisements; this results in a feature space of 6356 words which forms the importance-weighted word representation of the job advertisements.

3.5.2 Feature class 2: empirical ruleset

This class of features is derived from prior work done by Vidros et al. (2017). Through conducting empirical analysis of a balanced dataset of real and fraudulent jobs, those authors determined several contextual, linguistic and metadata features that might be informative of the legitimacy of a job advertisement. The three metadata features were already included in the dataset and were not extracted manually. Due to steps taken to preserve the anonymity of the personal information in the dataset, features related to HTML analysis were not included in this step. A list of the included features can be found in Table 1. All features are based on the ruleset by Vidros et al. with the exception of url_in_text.

Additionally, the rule-set-based features model was augmented with part-of-speech (POS) tag counts. The idea here is that these POS tags may confer some additional information about the linguistic composition of a job description that was not captured by other features or variables.

Table 1 Rule set-based features Adapted from Vidros et al. (2017)

Category	Name	Description
Linguistic	contains_spamwords	Job text contains a spam word such as 'online', 'extra', 'cash'
	consecutive_punct	Number of consecutive punctuation in the job text
	money_in_title	Title contains money symbols
	money_in_description	Description contains money symbols
Contextual	url_in_text	Text contains an e-mail address, phone number or link to an external website
	external_application	Text contains phrases such as 'apply at' or 'send resume'
	addresses_lower_education	Text contains phrases such as 'High School' or 'No degree'
	has_incomplete_extra_attributes	Attributes such as industry, function, required education or employment type are empty
	has_no_company_profile	Profile is empty
	has_short_company_profile	Profile is less than 10 words
	has_no_long_company_profile	Profile is more than 10 words but less than 100 words
	has_short_description	Description is less than 10 words
	has_short_requirements	Requirements are less than 10 words
Metadata	Telecommuting	Job marked as a telecommuting job
	has_no_company_logo	No company logo
	has_no_questions	Screening questions are missing

3.5.3 Feature class 3: word embeddings

In this class of features, word embeddings are learned using the gensim implementation of word2vec. First, the text is cleaned by removing non-letters, converting to lowercase and split at whitespace. Stop words are not removed at this stage to preserve context. Second, the text is split into sentences using the nltk punkt tokenizer. Third, the word2vec model is used to transform the sentences to its vector representation. The parameters were subsequently set, resulting in a word vector dimensionality of 300, context window size at 10, minimum word count of 40, and downsampling of 0.001. Once the model has been trained on the 2,075,540 raw words, average feature vectors are calculated for each job description.

3.5.4 Feature class 4: transformers

To explore whether the contextual word embeddings or transformers can offer improved performance on the fraudulent job type classification task, a selection of standard transformer models are trained on the corpus. This paper uses the SimpleTransformers implementation of the HuggingFace Transformers library. It was chosen as it allows researchers to achieve state of the art results without compromising due to complexity of code. This module is run on the raw job advertisement text, since SimpleTransformers applies tokenization automatically.

Table 2 Classification report for SGD fitted on bag of words feature set

Class	Precision	Recall	F1-Score
0	0.898	0.866	0.881
1	0.850	0.820	0.835
2	0.722	0.830	0.772
3	0.935	0.966	0.951
Weighted avg	0.857	0.853	0.854

Table 3 Classification report for SVM fitted on tf-idf feature set

Class	Precision	Recall	F1-Score
0	0.820	0.973	0.890
1	0.908	0.802	0.852
2	0.895	0.723	0.800
3	0.935	0.967	0.951
Weighted avg	0.876	0.870	0.868

4 Results

4.1 Bag-of-words and TF-IDF

Stochastic Gradient Descent (SGD) performed the best, with a weighted average F1-score of 0.854 and MCC of 0.790.

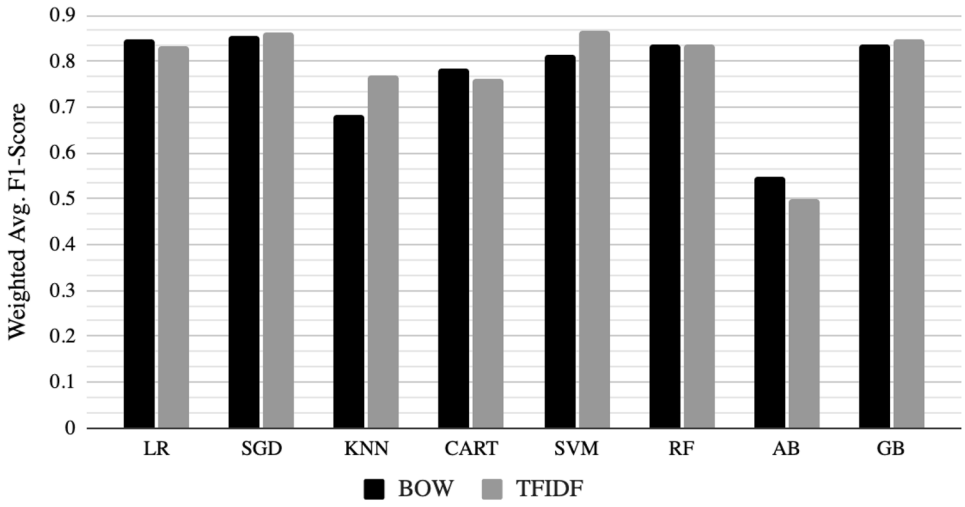

Fig. 2 Comparison of weighted average F1-score between BOW and Tf-Idf feature sets

Table 4 Classification report for RF fitted on rule-set based features

Class	Precision	Recall	F1-Score
0	0.759	0.732	0.745
1	0.790	0.747	0.769
2	0.530	0.553	0.542
3	0.737	0.933	0.824
Weighted avg	0.733	0.730	0.730

Table 2 shows the classification report for the SGD classifier, and demonstrates satisfactory recall for each class.

The tf-idf transformation resulted in the Support Vector Machine (SVM) classifier outperforming all other classifiers, with a weighted average F1-score of 0.868 and MCC of 0.814. Despite the SVM-tfidf outperforming the SGD-bow classifier on the average metrics, comparison of the two classification reports in Tables 2 and 3 shows the SVM-tfidf has lower recall on fraudulent job types 1 and 2, hence the performance difference could be attributed to the SVM model being able to capture real jobs more effectively.

A summary of the results for feature class 1 can be seen in Fig. 2.

4.2 Empirical ruleset

The highest results were achieved with a Random Forest Classifier, with a weighted average F1-score of 0.730. While this is superior to a random classification, this result indicates the rule-set-based features alone are significantly less able to differentiate between the types of job fraud, compared to differentiating between real and fraudulent jobs. The classification report in Table 4 demonstrates how the

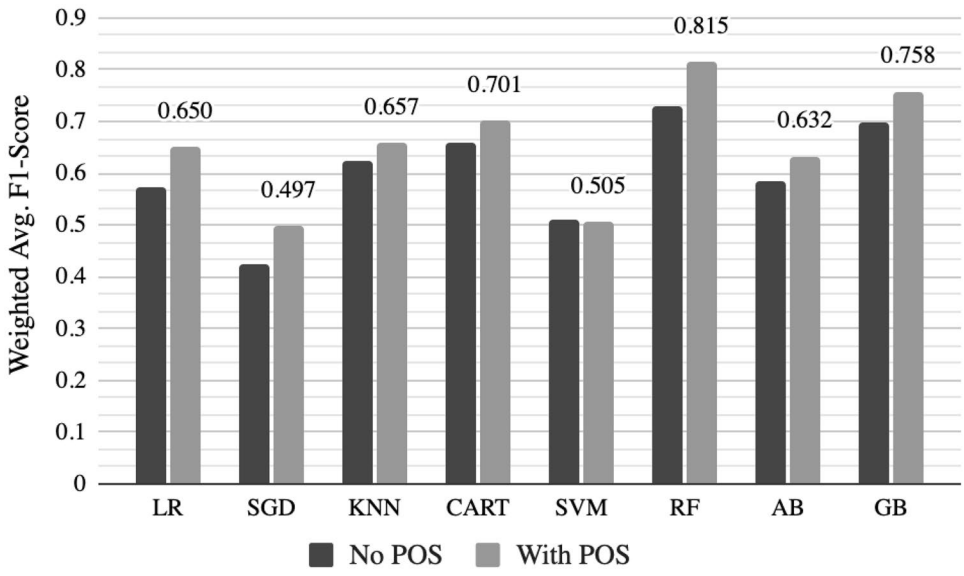

Fig. 3 Comparison of Rule-set based features with parts-of-speech tags

empirical rule set-based model struggled to effectively predict corporate identity fraud, with relatively low recall of 0.553 and low precision of 0.530.

4.2.1 Rule-set features with part-of-speech tags

The highest performing classifier was again a Random Forest classifier; in this instance, the inclusion of POS tags increased the weighted average F1-Score by 0.085–0.815. Likewise, improvements in precision and recall score for each class was observed. Similar improvements in both F1-score and MCC can be observed in Fig. 3 for most other classifiers, except SVM. Based on this, it can be inferred that POS tags are informative and effective features to classify fraudulent job types.

4.2.2 Rule-set features with part-of-speech tags and bag of words

Since the rule-set-based features performed worse on average in comparison to the vector representation of the job advertisements, the bag-of-words features are appended to the rule-set-based features with part-of-speech tags model to explore whether this will improve the classification performance. The results for these configurations are presented in Fig. 4.

It is evident that the combination of the two feature classes increased the overall performance of the tested classifiers. In the configuration with bag-of-words representation, Gradient Boosting achieved the highest performance yet, with an F1-score of 0.88. Similarly, for the configuration with tf-idf features, Gradient Boosting reached an F1-score of 0.86.

4.2.3 Feature explainability

The most important features were extracted from the highest performing rule-set-based classifier, the Random Forest with POS tags. It can be observed in Fig. 5 that consecutive punctuation, noun count, short company profile, company logo, and incomplete extra attributes are particularly important features. On the other hand, a short description, short company profile, money symbols in title and directing to external applications were less important features.

Figure 6 shows an example of the predictions, which is based on a Random Forest classifier of ruleset and pos tags, using Local Interpretable Model-Agnostic Explanations (LIME; Ribeiro, Singh and Guestrin, 2016). For an example instance of a type 1 job, a class probability of 0.64 was obtained. Figure 7 further shows to what extent certain features contributed to this classification probability. The presence of a short company profile and lack of company logo were important predictors of a type 1 identity theft job. On the other hand, the lack of money symbols in the title decreased the probability of being classified as type 1.

4.3 Word embeddings

The average word vectors are used as features for this multiclass classification task. With this model, a weighted average F-score of 0.766 with a Gradient Boosting classifier was achieved. The classification performance of corporate identity theft (type 2) job advertisements is the lowest, with a recall of 0.553. Despite this, the word embeddings model performed marginally better than the rule-set-based model and in the same range as the bag of words models. See

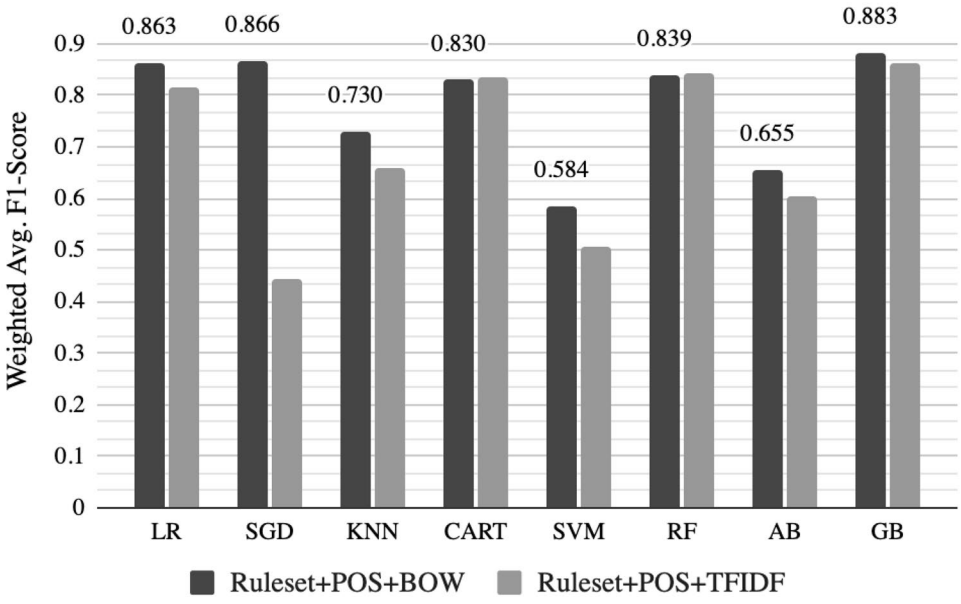

Fig. 4 Comparison of F1-scores for feature class 2 combined with feature class 1

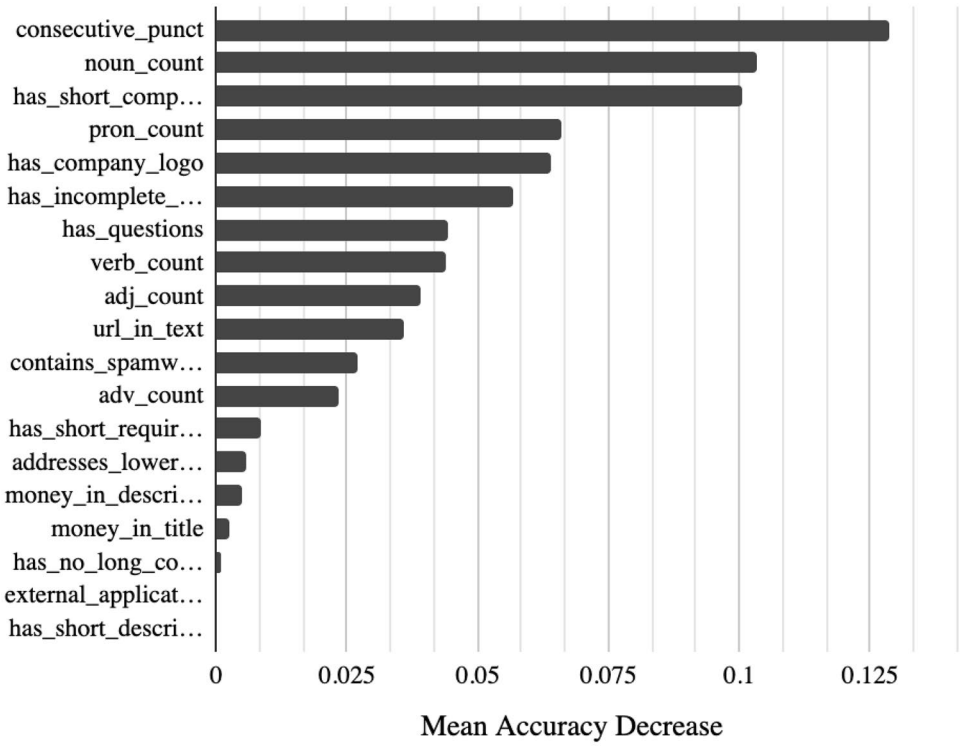

Fig. 5 Ruleset Feature Importances using permutation

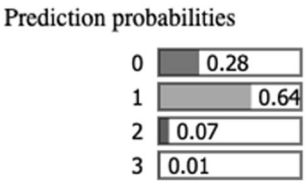

Fig. 6 LIME prediction probabilities for an example instance of class 1

Table 5 for the classification report of the Gradient Boosting classifier.

4.3.1 Word2vec embeddings combined with rule-set features

The word embeddings model was combined with the rule-set-based features class to explore whether augmenting the

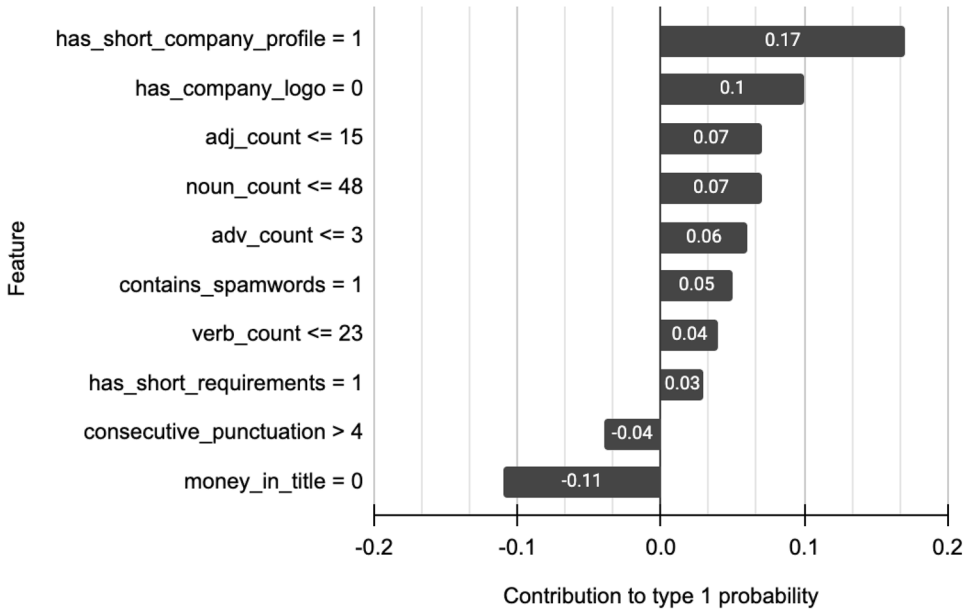

Fig. 7 Adapted LIME output for an example instance of class type 1

Table 5 Classification report for Gradient Boosting fitted on word-2vec embeddings

Class	Precision	Recall	F1-Score
0	0.713	0.866	0.782
1	0.814	0.712	0.760
2	0.765	0.553	0.642
3	0.879	0.967	0.921
Weighted avg	0.775	0.770	0.766

Table 6 Classification report for Random Forest fitted on combined embeddings and ruleset features

Class	Precision	Recall	F1-Score
0	0.832	0.929	0.878
1	0.837	0.784	0.810
2	0.775	0.660	0.713
3	0.935	0.967	0.951
Weighted avg	0.835	0.837	0.834

Table 7 Evaluation results for 3 transformer models

Transformer model	Weighted avg. F1-score	Matthews corr. coeff
BERT (bert-base-cased)	0.780	0.694
RoBERTa (roberta-base)	0.806	0.723
XLNet (xlnet-base-cased)	0.839	0.768

Table 8 Classification report for XLNet

Class	Precision	Recall	F1-Score
0	0.883	0.875	0.879
1	0.908	0.802	0.852
2	0.702	0.851	0.769
3	0.706	0.800	0.750
Weighted avg	0.846	0.837	0.839

empirical approach with the vector representation of words can yield improved performance. Simply put, the data frame containing the average feature vectors was concatenated with the rule-set based features. The full set of classifiers were fit on the training set and evaluated on the testing set. The classification report in Table 6 shows the addition of rule-set-based features improved the classification performance, and resulted in a Random Forest classifier with an F1-score of 0.83.

4.4 Transformers

Three different pre-trained transformer models are evaluated on the job advertisements: BERT, roBERTa, and XLNet. The results for these evaluations are reported in Table 7. XLNet yielded better performance, with a weighted average F-score of 0.84, which is marginally better than the word-2vec embeddings from feature class 3.

The results indicate that, for this classification task, these transformer-based models do not offer significant improvements in performance in comparison to configurations derived from other feature classes. However, analysis of the classification report in Table 8 shows that the transformer model achieves better recall for predicting corporate identity theft; this class has consistently underperformed in all prior models. Further fine-tuning of the XLNet model may yield even better performance.

4.5 Summary of results

A summary of the best classification results per each class of features is collated in Table 9. The best overall result was achieved with the Gradient Boosting classifier, in a model consisting of the empirical ruleset features, part-of-speech tags and the bag-of-words representation of the text.

5 Discussion

All feature classes were able to distinguish the different types of ORF. In all cases, corporate identity fraud was the most difficult to classify correctly, with a significantly lower recall compared to other types. This could be attributed to the fact that such types of fraudulent job advertisements are crafted more carefully to resemble a job opportunity from a legitimate job provider; some of these type 2 job advertisements may potentially be a direct duplicate of a real job advertisement.

Word embeddings and transformer-based features consistently outperformed the handcrafted rule-set-based features class. This may be attributed to the fact that such handcrafted features may not capture nuanced semantic or syntactic characteristics specific to the corpus which are otherwise picked up by word embeddings. Ultimately, it was demonstrated that the best performance was obtained by a Gradient Boosting classifier fitted on a configuration consisting of a combination of rule-set-based features, part-of-speech tags and bag-of-words vectors was the most effective in this text classification task. It was further observed that augmenting the ruleset feature space with additional features such as

Table 9 Summary of classification results

Feature Class	Model	Classifier	Weighted Avg. F1-Score
Bag-of-words	Bag of words	Stochastic gradient Descent	0.854
	TF-IDF	Support vector Machine	0.868
Rule set-based	Ruleset	Random forest	0.730
	Ruleset + POS	Random forest	0.815
	Ruleset + POS + Bag of Words	Gradient boosting	0.883
	Ruleset + POS + TF-IDF	Gradient boosting	0.863
Word embeddings	Word2Vec	Gradient boosting	0.766
	Word2Vec + Ruleset + POS	Random forest	0.834
Transformer	BERT	–	0.780
	RoBERTa	–	0.806
	XLNet	–	0.839

word embeddings or vector representation of text, resulted in significantly improved performance.

In terms of interpretability of the empirical features, it was found that the length of company profile part-of-speech tags, presence of company logo, presence of questions and presence of certain keywords were particularly informative of the fraudulent job type. More specifically, feature importance permutation revealed that consecutive punctuation, noun count, short company profile, presence of company logo and incomplete extra attributes were among the most important features. As per the feature correlations, type 2 corporate identity theft and type 3 multi-level marketing job advertisements are more likely to have longer company profiles than the type 1 identity theft advertisements. At the same time, type 1 advertisements are more likely to include money symbols in the description and title. Moreover, a higher amount of consecutive punctuation was more associated with predicting type 3 multi-level marketing vacancies compared to other types.

5.1 Implications

These are relevant findings in light of fraudulent job advertisements becoming more indistinguishable to the casual observer, and it may be more challenging to rely on stereotypical heuristics such as fraudulent jobs being short and heavy on spam words. The explainability and transparency afforded by better awareness of such knowledge can help inform ATS stakeholders in their policy-making to combat rising recruitment fraud, while potential use of such a classification system on job boards can better alert potential victims of the warning signs of job scams. Vidros et al. (2017) had formulated a goal of eventually creating an employment fraud detection tool for commercial purposes. The findings of this paper can contribute to the eventual development of such a tool, through highlighting what aspects of online job advertisements are indicative of its underlying type of fraud. For example, a job advertisement platform that wishes to add an additional layer of protection for its users, may implement a filter that automatically flags potentially fraudulent job advertisements for human review. An employee or entrusted stakeholder may then verify the legitimacy of the job advertisement, assisted by the output produced by the machine learning system. This output may consist of an explanation as to why the job advertisement was marked as fraudulent, in addition to what type of ORF is at hand. Being able to discern between multi-level marketing and corporate identity fraud, can better direct the potential victim as to their next steps: damage control on behalf of the affected organization may be more relevant in the case of corporate identity fraud, while reporting potential pyramid schemes to regulatory offices may be more relevant for multi-level marketing cases.

5.2 Limitations

Steps taken during the anonymization process (to be compliant with GDPR) may have altered the data in a way that modified original linguistic relationships or meanings in the job descriptions. At the same time, the time-bound and platform-specific aspect of the dataset might bring into question whether the findings will be generalizable. The data was compiled between 2012 and 2014, so changes in fraudsters' approach to job scams over the past decade might pose a challenge as it pertains to expanding the dataset with new entries.

Another limitation is the homogeneity of fraudulent job advertisements and the consequences that this has on sample size and class imbalance. In an attempt to remedy the class imbalance, several type 3 jobs were upsampled through resampling with replacement. While this did address the class imbalance to some extent, it may have further compounded the issue of homogeneity within the dataset, and consequently the generalizability of the findings. Lastly, there remains uncertainty about the robustness of certain models due to the fact that hold-out validation does not average out performance measures over multiple runs.

6 Conclusion

This paper developed and validated a machine learning system for identifying different categories of fraudulent job advertisements. The results indicate that the Gradient Boosting classifier with empirical rule-set based features, part-of-speech tags and bag-of-words vector achieved the best performance with an F1-score of 0.88. Hence, the identified job types can be distinguished on certain contextual and/or linguistic features. The length of the job description and company profile was a particularly informative feature. For example, identity theft vacancies had a higher correlation with a short company profile. Meanwhile, corporate identity theft vacancies tended to be longer. On the other hand, multi-level marketing vacancies used more consecutive punctuation and asked questions.

Future work in this area can focus on exploring different approaches to learn contextual and semantic information from the job advertisements, and might consider additional natural language processing techniques such as Latent Semantic Analysis, the ELECTRA transformer, or GloVe and fastText for word embeddings. Above all, this paper and the research stream in general would greatly benefit from a public access database of contemporaneous job advertisements, as this would allow researchers to form more relevant and timely recommendations as it pertains to combating online recruitment fraud.

Acknowledgements Kolawole Adebayo has received funding from Enterprise Ireland's CareerFit-Plus Co-fund and the European Union's Horizon 2020 research and innovation programme under the Marie Skłodowska-Curie grant agreement No. 847402.

Author contributions Marcel Naudé contributed to implementation of the methodology, feature engineering, application of machine learning and language models, analysis of results and writing the original draft. Kolawole John Adebayo contributed to supervision, review of the data annotation and editing the original draft. Rohan Nanda conceived of the presented idea and contributed to the supervision and editing the original draft.

Funding N/A.

Data availability Employment Scam Aegean Dataset (EMSCAD) is a publicly available dataset containing 17,880 real-life job ads. The dataset is available online at http://emscad.samos.aegean.gr/.

Code availability https://github.com/fjtrepo/fraudulentjobtypes.

Declarations

Conflict of interest The authors declare that there is no conflict of interest in this research.

Consent for publication All authors declare their consent to have the research published.

Ethics approval N/A.

Consent to participate N/A.

References

Alghamdi B, Alharby F (2019) An intelligent model for online recruitment fraud detection. J Inf Secur 10(03):155

Anita CS, Nagarajan P, Sairam GA, Ganesh P, Deepakkumar G (2021) Fake job detection and analysis using machine learning and deep learning algorithms. Revista Geintec-Gestao Inovacao E Tecnologias 11(2):642–650

Baykal E (2020) Digitalization of human resources: E-HR. In: Tools and techniques for implementing international e-trading tactics for competitive advantage. IGI Global, pp 268–286

Dutta S, Bandyopadhyay S (2020) Fake job recruitment detection using machine learning approach. Int J Eng Trends Technol 68(4):48–53. https://doi.org/10.14445/22315381/ijett-v68i4p209s

FTC (2021) Job scams. Federal Trade Commission Consumer Information. https://www.consumer.ftc.gov/articles/job-scams

Joyce S (2021) 5 major types of scam jobs and job scams online—JobHunt.org. Job Hunt. https://www.job-hunt.org/job-search-scams/

Lal S, Jiaswal R, Sardana N, Verma A, Kaur A, Mourya R (2019) ORFDetector: ensemble learning based online recruitment fraud detection. In: 2019 Twelfth international conference on contemporary computing (IC3). IEEE, pp 1–5

Mahbub S, Pardede E (2018) Using contextual features for online recruitment fraud detection. In: Andersson B, Johansson B, Carlsson S, Barry C, Lang M, Linger H, Schneider C (eds) Designing digitalization (ISD2018 proceedings). Lund University, Lund

Mehboob A, Malik MSI (2020) Smart fraud detection framework for job recruitments. Arab J Sci Eng:1–12

Reynolds B (2021) 14 Common job search scams and how to protect yourself|FlexJobs. https://www.flexjobs.com/blog/post/common-job-search-scams-how-to-protect-yourself-v2/

Ribeiro MT, Singh S, Guestrin C (2016) Why should i trust you?: explaining the predictions of any classifier. In: Proceedings of the 22nd ACM SIGKDD international conference on knowledge discovery and data mining, pp 1135–1144. https://doi.org/10.1145/2939672.2939778

Scamwatch (2015) Jobs and employment scams. Australian Competition and Consumer Commission. https://www.scamwatch.gov.au/types-of-scams/jobs-employment/jobs-employment-scams.

SEC (2013) SEC.gov|Beware of pyramid schemes posing as multi-level marketing programs. Sec.gov. https://www.sec.gov/oiea/investor-alerts-bulletins/investor-alerts-ia_pyramidhtm.html

Sivathanu B, Pillai R (2018) Smart HR 4.0—how industry 4.0 is disrupting HR. Hum Resour Manag Int Dig 26:7

Vidros S, Kolias C, Kambourakis G (2016) Online recruitment services: another playground for fraudsters. Comput Fraud Secur 2016(3):8–13

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