Maastricht University

MASTER THESIS

A Machine Learning Approach to Detecting Fraudulent Job Types

by
Marcel Adriaan Naudé
i6108892

Maastricht University, School of Business and Economics
MSc Business Intelligence & Smart Services
Regular thesis
Maastricht, 08 August 2021

Thesis supervisor: Dr. Rohan Nanda, Dr. Kolawole John Adebayo
Second reader: Dr. Niels Holtrop

ABSTRACT

Human resources services, particularly recruitment and job-hunting, increasingly take place online (Sivathanu & Pillai, 2018; Baykal, 2020). Online job platforms and applicant tracking systems (ATS) can afford greater accessibility and efficiency to stakeholders in the recruitment process, especially in a post-COVID age where organisations have had to adapt to stringent distancing measures (Ptel, 2020). At the same time, malicious actors take advantage of such platforms to disseminate fraudulent job advertisements, with the goal of harvesting private information or scamming money (Underwood, 2020; Mahbub & Pardede, 2018). This is known as online recruitment fraud (ORF), and has been the subject of several research publications in the past decade, where the goal has been to detect fraudulent jobs amongst a corpus of real life job advertisements (Vidros, Kolias, Kambourakis & Akoglu, 2017; Mehboob & Malik, 2020).

This thesis seeks to extend the prior work on online recruitment fraud by exploring to what extent fraudulent job advertisements can be classified according to their type, and what type of features are most suited for this task. Three different non-exhaustive types of fraudulent jobs were conceptualized after a review of the literature and analysis of the Employment Scam Aegan Dataset (EMSCAD; Vidros et al., 2017): identity theft, corporate identity theft, and pyramid schemes or multi-level marketing. Four classes of features were extracted from the text using natural language processing methods: empirical rule set-based features, bag-of-word models, word embeddings and transformer models. A range of classifiers were fitted on the feature sets and the results were validated by evaluating it on EMSCAD.

All features classes were able to distinguish the different types of online recruitment fraud. Word embeddings, bag-of-words and transformer models consistently outperformed the

handcrafted rule-set based features class. The exception was when handcrafted rule-set features were augmented with bag of words representation of the text: ultimately, the results indicated that the Gradient Boosting classifier with rule-set based features class combined with the bag-of-words vector representation achieved the best performance with an F1-score of 0.88. In all cases, corporate identity fraud was the most difficult to classify correctly, with a significantly lower recall compared to other types. The length of company profile, part-of-speech tags, presence of company logo, presence of questions and presence of certain keywords were particularly informative rule-set based features.

The findings of this thesis was limited by choices in methodology, such as lack of cross-validation and consistent hyper-parameter tuning, as well as the inherent noisiness and homogeneity in the job dataset. Future work in this area could focus on building a more robust typology of the different fraudulent jobs, expansion of the publicly available dataset with up-to-date job advertisements, as well as different methods of extracting or learning features from the advertisements.

Keywords: fraud detection, online recruitment, machine learning, natural language processing

TABLE OF CONTENTS

1 INTRODUCTION

Developments in technology and big data have ushered in a new era of Human Resources where services increasingly take place online (Sivathanu & Pillai, 2018; Baykal, 2020). Particularly, the use of Applicant Tracking Systems to conduct e-recruitment has gained popularity among HR practitioners and recruiters (Vidros, Kolias & Kambourakis, 2016). While job seekers and recruiters benefit from increased accessibility and efficiency, such e-recruitment platforms are subject to malicious abuse by actors who find opportunities to scam and conduct fraudulent activities. In this chapter, the scope of the thesis with regard to this research problem is introduced. First, a brief background on the issue of online recruitment fraud is given. Next, the current research is motivated accordingly. Thereafter, four research questions are formulated to guide the direction of this thesis, and the relevant contributions are outlined.

1.1 Introduction

Job seekers find themselves increasingly duped and misled by fraudulent job advertisements, posing a threat to their privacy, security and well-being (Mahbub & Pardede, 2018). Fraudsters are able to obtain sensitive information such as full names, addresses, contact information, social security numbers, and so forth (Vidros, Kolias & Kambourakis, 2016). Likewise, legitimate organisations are at risk of facing harm to their reputation and a smaller applicant pool as a consequence of such activities (Vidros, Kolias, Kambourakis & Akoglu, 2017). This problem has grown even more pronounced due to the COVID-19 pandemic, as online platforms became more important for business operations (Herath & Herath, 2020; Ptel, 2020). Given these costs to society, there is a clear need for solutions that can detect whether or not a job posting is fraudulent.

This problem has raised the interest of several researchers who attempt to build a solution that can detect fraudulent job postings. So far, this has been treated as a binary classification problem: selecting informative features from the corpus and applying machine learning techniques to build a model capable of accurately classifying a particular job posting as either legitimate or fraudulent. Vidros et al. (2016) kick-started this stream of research with their seminal paper highlighting the timeliness and urgency of the issue, establishing a baseline classification model using a Random Forest classifier later in 2017. They noted that text mining and metadata-based classification only builds the foundation, and subsequent approaches should also consider user behaviour, company and network information. They consequently published their dataset (EMSCAD) to allow further work on this topic. Since then, several authors have sought to further strengthen the classification of the EMSCAD dataset jobs. The focus of much of this prior research lays primarily on two areas: (1) which features can strengthen the performance of classification algorithms, and (2) what classification algorithm yields the highest performance.

In terms of feature extraction, researchers have applied Natural Language Processing techniques. In most cases, the dataset is initially pre-processed to prepare for text mining (Alghamdi & Alharby, 2019). Next, researchers build a ruleset typically based on empirical analysis of the corpus in combination with NLP. For example, Mahbub and Pardede (2018) developed a classifier with higher accuracy by adding contextual features that mimic user behaviour. In the case of Mehboob and Malik (2020), identified a large swath of features and subsequently selected the most significant features for inclusion in the model through a combination of information gain, correlation coefficient and gain ratio. Features are then

extracted and fed to the classification models using bag of words modelling (Vidros et al., 2017) and CoreNLP (Mahbub & Pardede, 2018).

In terms of classification models, researchers have attempted to build upon prior research by employing different machine learning classifiers. Experimentation with both single and ensemble classifiers has seen an increase of accuracy from the baseline 89.5% established by Vidros et al. in 2017, to 98.3% by Dutta & Bandyopadhyay (2020). These authors had employed a Random Forest ensemble classifier and concluded ensemble classifiers outperform single classifiers in this binary classification task. This finding was concurred by Alghamdi and Alharby (2019) and Mehboob and Malik (2020) who likewise found an XGBoost classifier demonstrated the highest performance.

1.2 Research Motivation and Questions

While the focus has been on improving predictive performance of classification models, there is still room for developing this ORF research stream. Vidros et al. (2017) had pointed out the publicly available EMSCAD dataset provides room for gaining deeper knowledge of the *characteristics* of the online recruitment fraud problem. They noted that there are two general approaches to employment fraud: the first concerns the large-scale collection of private information as submitted through applications on the job platform, while the second approach goes beyond surface-level information collection by engaging with potential candidates through fake interviews or aptitude tests, typically with the goal of obtaining highly sensitive information. There is also recognition in public discourse that there are different *types* of fraudulent jobs shared on job-post platforms. For example, in their article published in the popular press, Reynolds (2021) discerns between different types of fake job postings, ranging from pyramid marketing to rebate processing scams.

Despite awareness of employment fraud manifesting in different types with different characteristics, no research has been done yet that applies machine learning to classify fraudulent job ads according to their idiosyncratic characteristics. Therefore, this thesis seeks to answer the following research questions:

(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?
    (a) Do novel word embeddings or transformer models outperform handcrafted features?

(3) What machine learning algorithm performs the best for fraudulent job type classification?

1.4 Research Contribution

This thesis proposes to make several contributions on both the academic and practical side. From an academic and theoretical perspective, this thesis carries on the work that had already been done with regards to detecting online recruitment fraud, particularly by deepening the understanding of how data mining can be used to draw conclusions from online job postings. From a research perspective, the efforts to conceptualise and classify different characterisations of fraudulent jobs extends prior work on detecting online recruitment fraud by testing prior conclusions in a novel setting, such as the assumption that ensemble classifiers are most suitable in this domain, or that fraudulent jobs can be distinguished with

an empirical ruleset. Furthermore, the ability to differentiate between different fraudulent job advertisements based on their underlying motive using structural, lexical or other semantic features should allow for better transparency and explainability. This is an important component of making data mining conclusions more accessible to a wider audience. From the practical perspective, the findings could better inform ATS providers of the type of malicious activity occurring on the platform, and allow them to allocate or redirect resources to prevent specific frauds. This may also be particularly relevant in the post COVID-19 era, where much recruitment activity has been transferred online to accommodate lockdowns and work-from-home orders. In a time of economic hardship and employment insecurity, it becomes evermore important to have a more comprehensive understanding of this costly phenomenon.

1.5 Outline

The rest of the thesis is structured as follows. First, a literature review is conducted that explores the extant research stream on online recruitment fraud classification. This is supplemented with a review of the machine learning and natural language processing techniques which are suitable for this domain. Second, the methodology and research design is introduced in the context of the research questions, with the goal to elucidate how the EMSCAD dataset is used to answer these questions. Third, the research design is implemented in Python and 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. Lastly, the thesis arrives at a conclusion.

2 LITERATURE REVIEW

In this section, the state of the art in online recruitment fraud research as well as the relevant machine learning techniques are reviewed. Firstly, a review of the current state of the art in online recruitment fraud research is conducted. Secondly, an overview of the different types of online recruitment fraud is presented.

2.1 Online Recruitment Fraud

The first mention of online recruitment fraud (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. These authors conceptualised ORF as 'malicious behaviour aiming to steal personal information' by exploiting the 'job syndication process', which ultimately results in economic and/or reputation damage to applicant tracking system stakeholders. In other words, 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 digitised and cloud-based hiring process, and purported that further uptake of digital recruitment will coincide with even more ORF. To explore whether this is a topic relevant to machine learning, the authors collected a repository of real and fake jobs, and established an initial empirical rule-set denoting fraudulent jobs. Each job in a balanced sample was assigned a 'fraud score' based on the number of rule-set based conditions it met. The authors found support for their rule set-based approach through imposing a decision threshold on the fraud score space. Hence, the timeliness and severity of the issue, as well as its relevance to machine learning was already demonstrated.

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

fake 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. In order 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. On the one hand, scholars attempted to improve the prediction performance by experimenting with different classifiers, such as ensemble and deep learning methods. Lal et al. (2016) tested an ensemble-based classifier composed of J48, Random Forest and Logistic Regression classifiers, achieving 95.4% accuracy using the same rule-set based features proposed by Vigros et al. (2017). Dutta and Bandyopadhyay (2020) compared single and ensemble classifiers, and affirmed that ensemble classifiers are better for this particular fraudulent job classification task. Mehboob and Malik (2020) achieved 97.9% accuracy using an XGBoost classifier, slightly outperforming the Random Forest. Most recently, Anita et al. (2021) applied a Bidirectional LSTM network and reached 98% accuracy.

On the other hand, scholars investigated whether the selection of features can be improved or expanded to build more robust models. In 2018, Mahbub and Pardede expanded the feature space by considering additional contextual features, such as information about the focal organisation's legitimacy and internet footprint. They used web search to manually determine whether the provider of the job ad had a valid web page, for example. The inclusion of these

features resulted in significant increases in predictive performance, achieving 96% accuracy using a JRip classifier. In that same year, Alghamdi and Alharby (2019) demonstrated 97% accuracy in predicting fraudulent jobs using Support Vector Machine as a feature selection method and Random Forest as a classifier. They found the most informative features were company logo, presence of company profile and industry. Similarly, Mehboob and Malik (2020) used a two-step feature selection strategy using information gain, correlation coefficient and gain. In addition to company profile, they also found that salary range and organisation type are effective indicators of whether a job is fraudulent. Goyal, Sachdeva and Kumaraguru (2021) used a fact-validation approach to incorporate external knowledge-based features based on information present within the job postings, and reached an F1-score of 0.98.

Past conclusions should be viewed in the frame of the limitations posed by the dataset. Since the rule set-based features were constructed on the single EMSCAD dataset, it is possible that the generalisability is undermined by dataset-specific artifacts. Goyal et al. (2021) noted that handcrafted features such as those proposed by Vidros et al. (2017) might not scale well with larger datasets, while Kim, Kim and Kim (2019) pointed out that such rule-based methods are more difficult to generalise. At the same time, the data is highly unbalanced, noisy and lacks potentially informative contextual metadata about the job posting. Kim et al. wrote that this is a reality of many job advertisement platforms, and that potential fraud detection solutions should be able to manage without meta-information. Hence, this stream of research stands to benefit from an updated and more expansive repository of real-life job postings, as well as more research into different supervised and/or unsupervised learning techniques.

In summary, the current state of the art on online recruitment fraud research holds that ensemble-based classifiers are best suited for this task (Dutta & Bandyopadhyay, 2020). Deep learning methods also show promise in this area (Anita et al., 2021). Furthermore, informative features include job post metadata such as industry and function (Alghamdi & Alharby, 2019), knowledge features based on a fact-checking algorithm (Goyal et al., 2021), contextual features such as maturity of company website (Mahbub & Pardede, 2018) and structural features such as presence of company profile or logo (Vidros et al., 2017).

2.2 Online Recruitment Fraud Types

Thus far, research has only treated the problem of online recruitment fraud as a dichotomy: either a job is fake, 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.

Not all online recruitment fraud is equal in its methods or goals; this notion was already raised in the seminal paper by 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, more severe 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. Here, the fraudster may engage in behaviours 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 in order to exploit the job syndication process; the extent to which the fraudulent job seems indistinguishable from a legitimate advertisement in the latter group may pose additional challenges for the job seeker who wishes to avoid a scam.

As this is a relatively contemporaneous and novel research stream, not much scientific research has investigated the characteristics of online recruitment fraud beyond the paper by Vidros et al. (2016). Despite this, various sources can be consulted to form a better perspective on what sets fraudulent job advertisements apart. Given the social impact of the online recruitment fraud issue, several public organisations and other interested stakeholders have written articles and/or created resources that aim to help the job seeker identify fraudulent job advertisements or deal with the aftermath thereof. On behalf of a job search platform, Reynolds (2021) wrote about nine different types of job search scams, including pyramid marketing, stuffing envelopes, data entry scams, and online re-shipping. These scams differ in the type of actions it demands from the job seeker, and are typically more focused on procuring some form of payment from the job seeker as opposed to the information harvesting fraud discussed by Vidros et al. (2016). Job-Hunt, 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 organisations 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 emphasises 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).

From this, two general areas of online recruitment fraud start to manifest and can be summarised as follows. First, there are information harvest scams which aim to dupe job seekers into providing sensitive information. Second, there are job scams which solicit services from the job seeker for malicious means, such as money laundering or pyramid schemes. To build upon prior research, it is relevant to consider to what extent these different categories or types of online recruitment fraud can be automatically detected within real-life online job postings.

2.3 Summary

In this section, background work in online recruitment fraud was studied. The exploration of literature in the ORF detection domain showed that this is a binary classification task which has undergone optimization as different authors experimented with different methods of feature selection and feature extraction, and with different machine learning algorithms. The state of the art holds that fraudulent jobs can be accurately detected with ensemble classifiers fitted on a variety of linguistic, contextual and structural features. However, gaps in background work point to room for exploring characteristics of fraudulent jobs on a more granular level. Academic and public sources were consulted for an understanding of the

different types of fraudulent jobs, and pointed towards the existence of several distinct expressions of fraudulent activity within online job advertisements.

3 METHODOLOGY

In this section, the methodology of the thesis is explained. First, the overarching research design is outlined. Second, the relevant dataset is introduced and steps to maintain anonymity are explained. Third, the data annotation process is presented alongside the criteria for distinguishing within the selected typology. Fourth, steps taken to adequately prepare the data for processing are explained. Fifth, the selection and extraction of feature classes are outlined. Sixth, an overview of the selected machine learning methods, including choices of classification and evaluation methods, is presented. Lastly, the relevant tools and programs used to fulfil the research objectives are briefly discussed.

3.1 Research Design

The higher-level research design of this thesis is rooted in the Knowledge Discovery from Data (KDD) model (Alghamdi & 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 is anonymised, annotated and pre-processed to prepare for analysis and transformation. Third, feature extraction and selection is performed for four classes of features: bag-of-words, rule set-based, word embeddings and transformer. For each model developed per feature class, several multi-class classification algorithms are trained on the model to classify the type of fraudulent job. Lastly, the output of the data mining models is compared, interpreted and evaluated with respect to the research objectives.

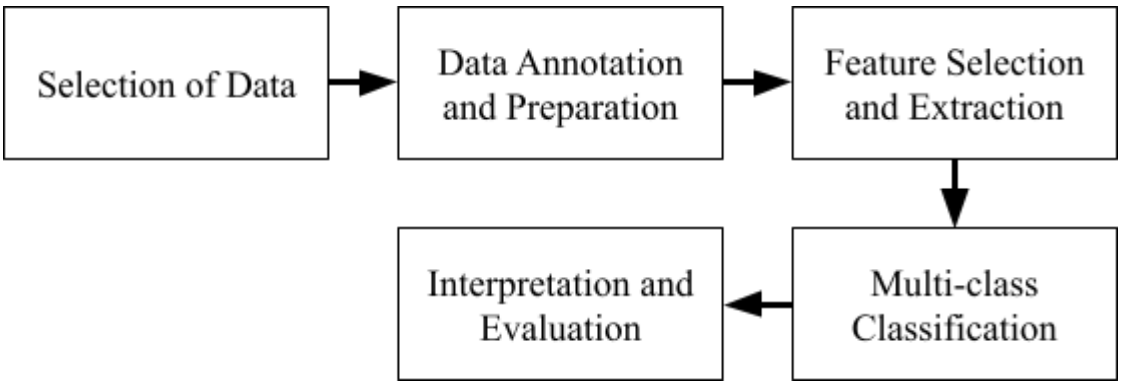

Figure 1. Research design.

3.2 Selection of Data

The data used to fulfil the research objectives is based on the Employment Scam Aegean Dataset (EMSCAD) made publicly available by Vidros et al. (2017), which consists of 17,880 job vacancies collected between 2012 and 2014. Each entry of the dataset comprises a single job vacancy, with its fields supplemented with a variable denoting whether or not the job is fraudulent. See Table 1 for a list and description of the variables in the dataset.These vacancies were collected by Vidros et al. from and in co-operation with the job advertisement platform Workable, with annotation done by specialised Workable employees according to company policy and meticulous analysis (Vidros et al., 2017). For this reason, it is assumed that the annotation is correct and trustworthy, and no additional steps will be taken to verify the annotation.

In order to be compliant with ethical and legal regulations, such as the General Data Protection Regulation (GDPR), careful attention was directed towards achieving a fully anonymised version of the dataset, before conducting any data processing or handling. The research objectives of this thesis do not hinge on the availability of any personally identifiable information, hence the anonymisation or deletion of such information should not have any significant impacts on the outcomes of the research. The original authors of the

EMSCAD dataset had already taken steps to ensure that most e-mails, phone numbers and URLs had been pseudonymised. However, there were instances where some of such information still remained. Hence, aggressive approaches were taken in Python to ensure all names, phone numbers, e-mail addresses and URLs were completely anonymised. First, entity recognition with the Stanza NLP pipeline as well as a dictionary of names was used to remove person entities. Next, the Presidio Anonymizer python module was used as an additional method to ensure the dataset was anonymous. Finally, a custom regex pattern was run over the dataset to filter out persistent phone numbers.

Variable	Description
title	The title of the job advertisement.
location	The location of the job.
department	The organisational department of the job.
salary_range	The range of salary of the job.
company_profile	A description about the company who posts the job advertisement.
description	A description about the job itself.
requirements	The requirements of the job.
benefits	The benefits of the job.
telecommuting	Whether the job allows for work-from-home.
has_company_logo	Whether the job advertisement has a company logo.
has_questions	Whether the job advertisement has questions.
employment_type	Whether the job is full-time, part-time or other.
required_experience	The prior experience required for this job.
required_education	The prior education required for this job.
industry	The industry of the job.
function	The function type of the job.
fraudulent	Whether the job is fraudulent or legitimate.

Table 1. Relevant Variables in the EMSCAD dataset.

During the thesis process, the possibility to collect additional sources of data to augment the current repository of job ads was raised and explored. This was motivated by two reasons: the small number of fraudulent jobs may negatively influence the predictive power of the machine learning model, and the platform- and time-specific nature of the EMSCAD dataset may deliver some contextual artifacts which jeopardise the generalisability of the potential findings. Several sources were considered, from downloading publicly available datasets on Kaggle such as Reed UK and eMedCareers, to manually scraping job advertisements from a job platform such as JobSpider. Ultimately, it was decided to not collect any additional datasets but to proceed solely on EMSCAD; this was due to complications with obtaining formal and legal authorisation to use the datasets, inconsistencies with data objects, and difficulties with GDPR compliance. Despite that, conducting analysis on the existing dataset still proves to be a worthwhile endeavour as it has been a source of insight in several prior publications.

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. Type three concerns such job advertisements where certain actions or services are solicited from the job seeker, such as face-to-face marketing of products for a commission-based income. Although the literature review had highlighted the existence of more than 3 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. Hence, this thesis's fraudulent job typology is directly based on the availability of data from EMSCAD, and is therefore non-exhaustive.

Since the research objectives concern only types of *fraudulent jobs*, all 866 fraudulent jobs from the EMSCAD dataset were then isolated and individually annotated. See Appendix A for a sample of the unprocessed job text per each type. In order to determine which category a given job advertisement belongs to, the following criteria were applied to the job's company profile, description, requirements and benefits text fields. These criteria were verified by an industry expert through annotated samples.

(1) Identity theft

(a) Classified in the EMSCAD dataset as fraudulent

(b) No reference to legitimate, reputable companies

(c) No company/employer identity, or fake company name

(d) Explicit requests of personal information (e.g. full name, address, phone number) outside of typical submission of CV

(e) Referral to an external web-page or e-mail address to finish application

(f) Does not belong to category 2 or 3

(2) Corporate identity theft

(a) Fake job ad which claims to be from a legitimate, reputable company

(b) It contains a description of the legitimate company

(c) 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

(a) Job ad incentivises or requests applicant's involvement in a referral/commission-based scheme (get paid for successfully sharing the job ad with someone else)

(b) Content of the job advertisement describes activities that are involved in MLM/referral schemes, i.e. the purported benefit that the targeted candidate accrues is a result of the number of additional candidates they recruit, sign up or refer

(c) Makes use of language like 'refer, 'referral bonus'

Whether or not a company or organisation mentioned in a given job advertisement is deemed as legitimate was informed by a Google search of the named entity. A legitimate and/or reputable company is one that is confirmed by additional research to be a provider of jobs, and lacks reviews or public notices which deride the company for job scams. This bears similarity to the approach used by Mahbub and Pardede (2018), who had included several features derived from manually searching named organisations or companies.

3.4 Data Preparation

After each job vacancy was manually annotated and a set of annotations were reviewed by an industry expert for validation, a random sample of 866 real jobs were merged into the annotated fake 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 fake jobs in the dataset. There were only 72 of type 3 jobs compared to 556 type 1 jobs. To mitigate some of this class imbalance, type 3 jobs were randomly upsampled using resample with replacement. 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 data transformation. Subsequently, the four columns were dropped from the dataset as they are no longer needed. Thereafter, a processing function which makes use of the 'en_core_web_lg' spaCy pipeline 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 Selection

Four distinct classes of features are investigated: bag-of-words model, empirical rule set, word embeddings and transformer models. The first two classes bridge to prior research done in this area as they have demonstrated efficacy in detecting fraudulent jobs, while the latter two are more novel and state-of-the-art approaches to such a natural language processing task. In the class of empirical rule set features, the extraction is completed prior to the cleaning and preprocessing of data. This is because many such explored features are based on elements of the text such as punctuation, capitalisation, pronouns and URLs, which are removed in the cleaning stage. The other classes of features are extracted on the cleaned dataset.

3.6 Classification and Evaluation

The target variable of this study is a categorical variable with four classes, hence the choice of machine learning algorithm should be suited for such a multi-class classification task. In

addition, the selection of classifiers can be informed by prior work in this field. Generally, ensemble learning methods showed better performance for fraudulent job classification (Dutta & Bandyopadhyay, 2020). Besides such ensemble classifiers, this thesis will also explore a set of single classifiers to validate and confirm prior findings in this area.

Therefore, based on prior state of the art results in this research area, as well as accessibility and usability of Python implementations, thesis will use the following classification algorithms 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). The classifiers will be fit on a training set using hold-out validation.

The classifiers are evaluated on several metrics: precision, recall, F-score and Matthews correlation coefficient (MCC). Accuracy is typically used as a key metric in prior work done in this field, however due to the class imbalance and the importance of not misclassifying a fake job, particular attention will be paid to a weighted average F-score. This metric, which is a combination of precision and recall, will be representative of the classifier's ability to balance precision and recall. The overall F-score for a particular classifier is calculated by weighing the F-scores for individual class predictions per the number of instances.

$$F\ score\ =\ \frac{2 * precision * recall}{precision + recall}$$

3.7 Tools and Programs

Given that this is a problem that requires extraction of information from a text corpus, the approach will largely be defined by concepts and theories of Natural Language Processing

(NLP) and machine learning. The dataset will be inspected, pre-processed, cleaned and analysed in Python using publicly available libraries such as pandas, spaCy, and scikit-learn. The program of choice is Python due to its large availability of NLP resources as well as extensive documentation. Moreover, scikit-learn provides for the easy implementation of a wide variety of classification algorithms.

3.8 Summary

In this chapter, the research methodology was presented. This thesis follows a process rooted in the Knowledge Discovery from Data model (Alghamdi & Alharby, 2019) in order to answer the research questions. A publicly available dataset containing 17,880 job advertisements is cleaned and anonymised to be used in this thesis. All fraudulent jobs were manually annotated with the type of fraudulent job according to criteria based on empirical analysis and literature review. Four classes of features were outlined and the relevant machine learning methods were subsequently introduced.

4 RESULTS

In this chapter, the research design is executed and the results are reported accordingly. The chapter is structured into four subchapters, each one pertaining to the feature extraction, classification and evaluation of a particular feature subclass. The first two feature classes are rooted in previous work done in this research area, while the latter two are feature classes which have not been thoroughly explored in previous work. First, the bag-of-words and tf-idf feature sets are reported. Second, the empirical rule-set features are explored and interpreted. Third, word2vec word embeddings are tested. Lastly, transformer-based models such as BERT, roBERTa and XLNet are explored.

4.1 Bag-of-Words

The first class of features is a simple bag-of-words (BOW) representation of the corpus, wherein a job advertisement is represented as an unordered collection of its constituent words. The first approach (bow model) entails the unigram word count vectorisation of each job advertisement. The resultant model comprises a matrix of words where the value of each word is equal to the amount of times that word is present in a given job advertisement. The second approach (tf-idf model) incorporates the importance of each word in a job advertisement as it relates to the entire corpus of text. The logic behind this second approach is that words which occur throughout many job advertisements may not be effective discriminators and should thus be discounted (Zhang, Yoshida & Tang, 2011). Hence, both approaches are conducted on the dataset and subsequently compared to explore which approach is more effective in the classification task at hand.

The first configuration in this feature class is a representation of the job advertisements based on the occurrence count of its words. This BOW transformation is implemented on the cleaned and pre-processed text using the CountVectorizer module from scikit-learn. This representation of the job advertisements is subsequently used as a feature for a machine learning algorithm. The total size of the vocabulary and hence the feature space is 6356 words after stopword removal and other processing steps. The dataset was split into an 80/20 training and testing set with stratified sampling. A variety of classifiers from the sklearn package were fitted on the training data. With the target variable of ‘type’, the bow model was evaluated against the test set. Stochastic Gradient Descent (SGD) performed the best, with a weighted average F1-score of 0.854 and MCC of 0.790. Table 2 shows the classification report for the SGD classifier, and demonstrates satisfactory recall for each class.

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

The second set of features in this class, tf-idf, follows a similar approach to the bag-of-words model. 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. On the training and test set, the tf-idf model is tested against the variety of classifiers. 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.

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

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

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

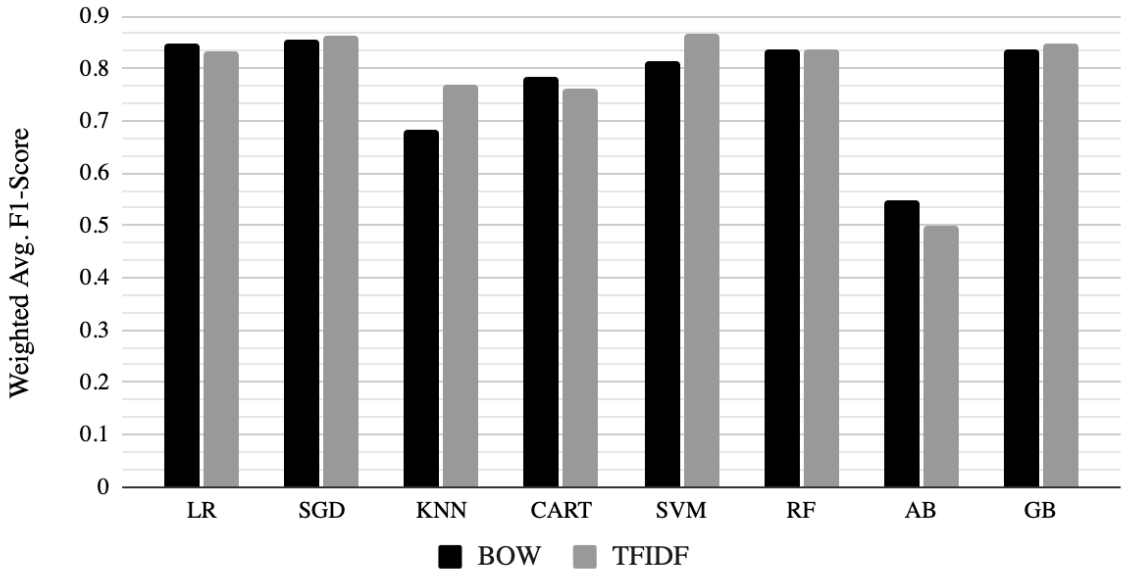

Figure 2. Comparison of weighted average F1-score between BOW and Tf-Idf feature sets

4.2 Empirical Ruleset

This class of features has its origins in 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. From these previously identified features, this thesis applied methods in Python to extract, for each job advertisement, their binary and count values. Three metadata features were already included in the dataset and were not extracted manually. Due to steps taken to preserve the anonymity of the dataset, features related to HTML analysis were not included in this step. A list of the extracted features with their description can be found in table 4. All features are based on the rule-set by Vidros et al., with the exception of url_in_text.

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	Job title contains money symbols
	money_in_description	Job description contains money symbols
Contextual	url_in_text	Job text contains an e-mail address, phone number or link to an external website
	external_application	Job text contains phrases such as ‘apply at’ or ‘send resume’
	addresses_lower_education	Job 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	Company profile is empty
	has_short_company_profile	Company profile is less than 10 words

	has_no_long_company_profile	Company profile is more than 10 words but less than 100 words
	has_short_description	Job description is less than 10 words
	has_short_requirements	Job requirements are less than 10 words
Metadata	telecommuting	Job is marked as a telecommuting job
	has_no_company_logo	There is no company logo
	has_no_questions	Screening questions are missing

Table 4. Rule Set-Based Features adapted from Vidros et al. (2017)

Since the rule set-based features were derived from the original authors' work (Vidros et al., 2017), it was important to compare the current thesis's attempts at extracting such features to the performance attained by those original authors. In order to validate the extraction of these features, a Random Forest classifier was trained on a balanced dataset of 450 genuine and 450 fraudulent jobs, after being split into 80/20 training and testing set through hold-out validation. The sample of jobs was determined by the variable 'in_balanced_dataset', which totals 900 jobs in total. The choice of Random Forest follows from the fact that Vidros et al. obtained their highest performance with this classifier. The input variables are the binary features, and the outcome variable is whether the job is fraudulent or not. The Random Forest classifier reached an accuracy of 88.9% and weighted average F-score of 0.889; the results per class can be seen in table 5. This is comparable to the findings of the original authors: Vidros et al. (2017) had achieved 90.5% accuracy and 0.906 F-score with their Random Forest rule set-based feature model.

Class	Precision	Recall	F1-Score
0 (legitimate)	0.907	0.867	0.886
1 (fraudulent)	0.872	0.911	0.891

Table 5. Rule set validation

The marginally lower performance of this thesis's rule-set based feature model can be attributed to omissions in the working dataset that rendered the extraction of certain features impossible. Due to aggressive steps taken in anonymising the dataset, HTML tags were removed from the text fields of job description, benefits, requirements, and company profile. Vidros et al. had found from a Pearson's correlation feature analysis that certain features derived from these HTML tags, such as 'has_no_html_lists_in_requirements' and 'has_no_html_lists_in_benefits' were effective predictors of fraudulent jobs. Despite the non-inclusion of these features, the achieved performance is still satisfactory and confirms the features were extracted appropriately.

4.2.1 Rule Set-Based Features

After having validated the rule set-based features on the original balanced dataset and target variable, these features are used as input to predicting the *type* of fake job. First, all the rule set-based features were included in the model, with the exception of variables that showed evidence of multicollinearity. It was observed through a correlation matrix that the following pairs: (1) telecommuting and has_company_logo, and (2) has_no_company_profile and has_short_company_profile, have a perfect correlation of 1. Telecommuting and has_no_company_profile were thus removed from the feature space. See Appendix B for the correlation matrix.

The dataset was split into a 80/20 training and test set with stratified sampling. The aforementioned classifiers were then evaluated on the test set.

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 6. Classification report for RF fitted on rule-set based features

The lowest results are attributed to the Stochastic Gradient Descent (SDG) and Support Vector Machine (SVM) classifiers; while these had performed exceptionally well in the bag-of-word model feature class, they failed to achieve sufficient performance on the handcrafted feature class. Instead, 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 as seen in table 6 demonstrates how the rule set-based model struggled to effectively predict corporate identity fraud, with relatively low recall of 0.553 and low precision of 0.530.

4.2.2 Rule-set Based Features with Part-of-Speech Tags

Next, 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.

The updated configuration with the POS tag counts included was re-trained and evaluated. 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 to 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 figure 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.

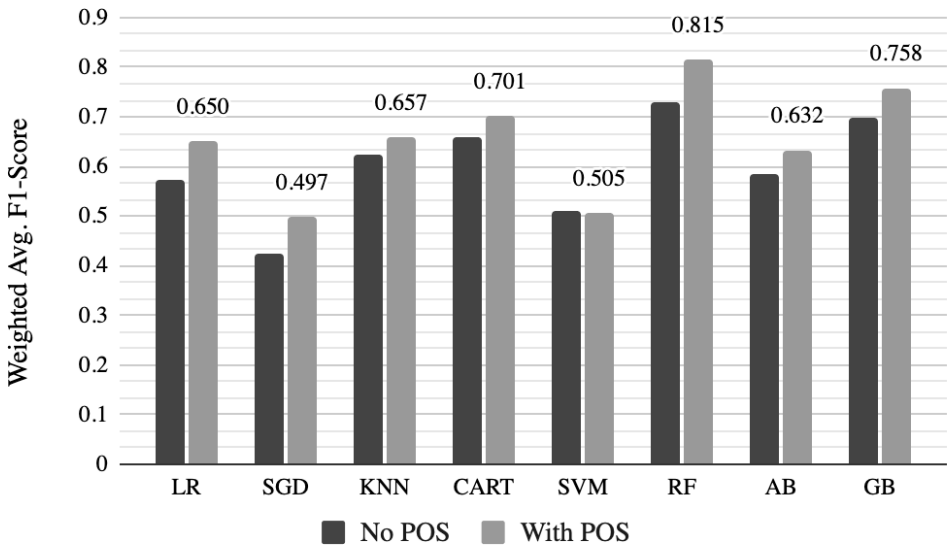

Figure 3. Comparison of Rule-set Based Features with Part-of-Speech Tags

4.2.3 Rule-Set Based 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 figure 4.

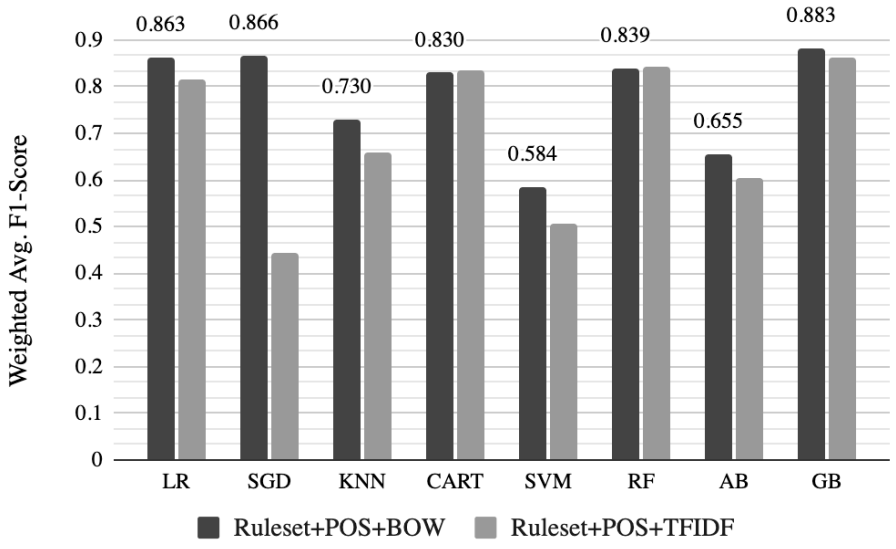

Figure 4. Comparison of F1-scores for feature class 2 combined with feature class 1

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.

4.2.4 Feature Explainability

The most important features were extracted from the highest performing rule set-based classifier, the Random Forest with POS tags. This importance measure is based on the permutation of the training data set's feature values. It can be observed in figure 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. This gives an indication as to what characteristics within job descriptions are effective discriminators.

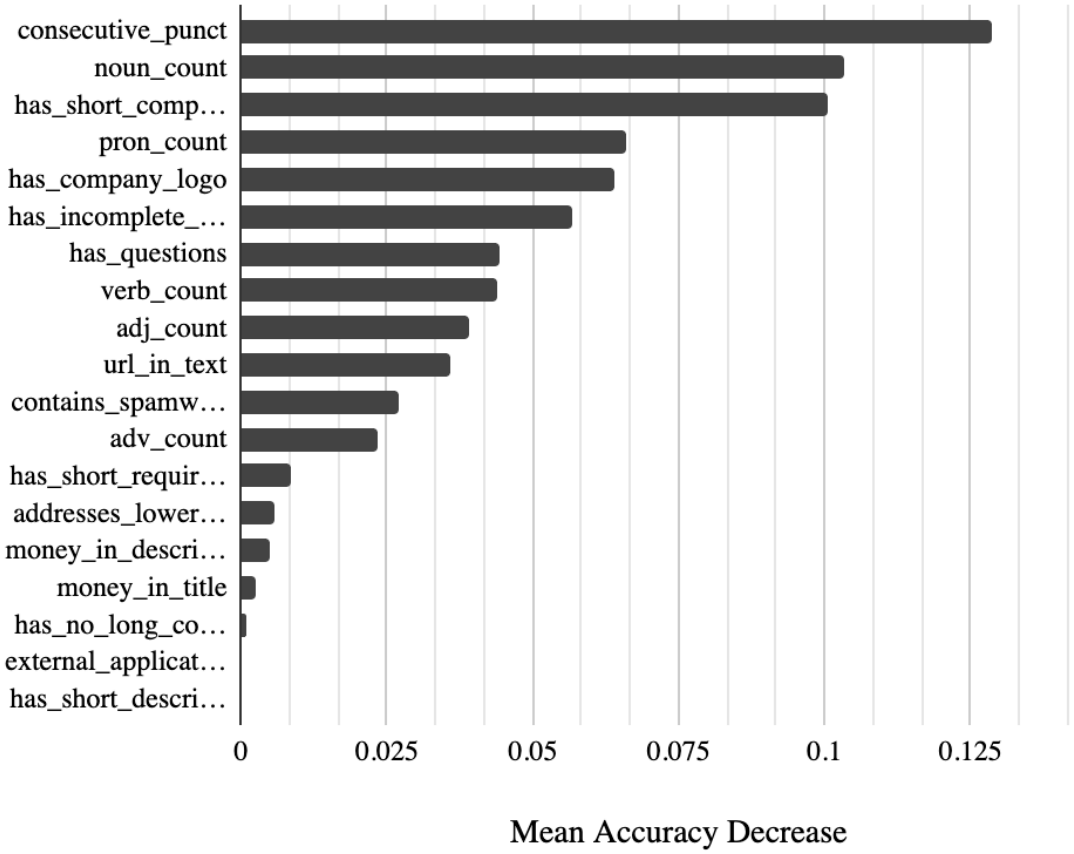

Figure 5. Ruleset Feature Importances using Permutation

Beyond Random Forest permutations, prediction explanations for different classes can be inspected in table 7 using Local Interpretable Model-Agnostic Explanations (LIME; Ribeiro, Singh & Guestrin, 2016) for a more transparent outlook on the rule-set classifiers. The discriminative features are those which contributed to being classified in the respective class. The identity theft example (type 1) is characterized by a short company profile, low amount of POS tags, lack of company logo and presence of spam words. The corporate identity theft (type 2) example is significantly longer in length, and in contrast to the identity theft

example, this prediction is also based on the lack of spam words. Further distinguishers of the type 2 example includes the presence of external links and relative lack of consecutive punctuation. Lastly, the multi-level marketing (type 3) example is similar to the type 2 prediction in the length, however particular discriminators in this case include a large number of consecutive punctuation and questions in the text. The interpretation derived from these explanations show that there are certain lexical features that allow one to differentiate between the types of fraudulent jobs.

Type	Example Text	Discriminative Features
1	Start Part-time, flexible opportunity. We are looking for people who are serious (and not just curious) I need someone who can post my ads on eBay. Laptop and internet connection. eBay with + rating £ per week. £ per listing.	● short_company_profile = 1 ● has_company_logo = 0 ● adj_count <= 3 ● noun_count <= 48 ● adv_count <= 3 ● contains_spamword = 1 ● verb_count <= 23 ● has_short_requirements = 0
2	NAME_MASKED Solutions is a global provider of products, systems and services to the oil and gas industry. Our engineering, design and technology bring discoveries into production and maximize recovery from each petroleum field. We employ approximately , people in about	● noun count > 149.25 ● consecutive_punct <= 4 ● has_company_logo = 0 ● contains_spamword = 0 ● url_in_text = 1 ● has_short_description = 0

	countries. Go to #URL_MASKED# for more information on our business, people and values [...]	• adj_count < 46 • money_in_title = 0
3	[...] Expert negotiations, maximizing total compensation package Signing bonus by NAME_MASKED in addition to client signing bonus (if applicable) 1 Year access to AnyPerk Relocation Services for out of town candidates. Continued education in your area of profession, seminars, workshops and other skill development events. Contract employees receive quarterly bonuses for the duration of their project Direct-Hire employees receive double bonues ($,) per referred/recruited candidate into their newly appointed company. All candidates are encouraged to participate in our Referral Bonus Program ; earn $ - $, per hired referral [...]	• consecutive_punct = 46 • noun_count > 149.25 • adv_count > 12 • has_questions = 1 • has_company_logo = 1 • has_short_company_profile = 0 • has_short_description = 0

Table 7. Classifier explanations per three different examples

Figure 6 further illustrates an example of the predictions, which is based on a Random Forest classifier of ruleset and pos tags. It can be observed that the lack of money symbols, as well as over four consecutive punctuation, in the title decreased the probability of being classified as type 1, in favor of being classified as a legitimate job advertisement. Similarly, figure 7 shows that the corporate identity fraud (type 2) job advertisement which did not address

lower education had a decrease in its probability for a type 2 classification. This can be indicative of corporate identity fraud targeting a different segment, despite appearing very similar to legitimate job advertisements on face-value. Lastly, figure 8 shows the LIME interpretation of a multi-level marketing (type 3) advertisement, where a long requirement section as well as not prompting for external application were detractors from a positive classification in this class.

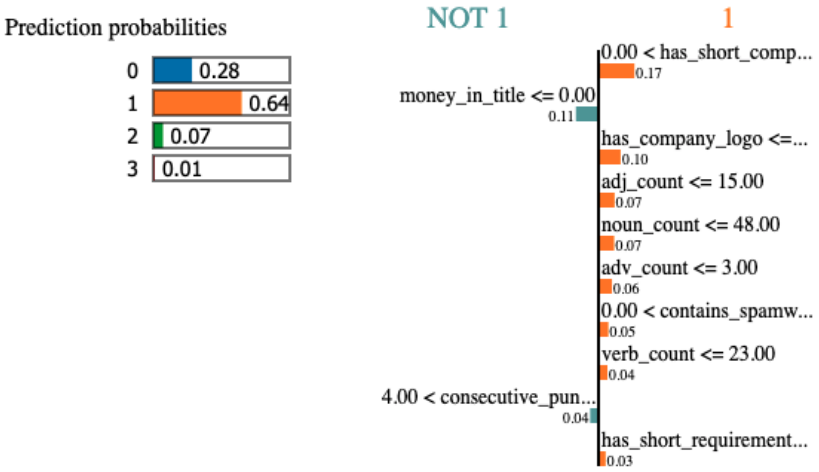

Figure 6. LIME output for an example instance of class 1

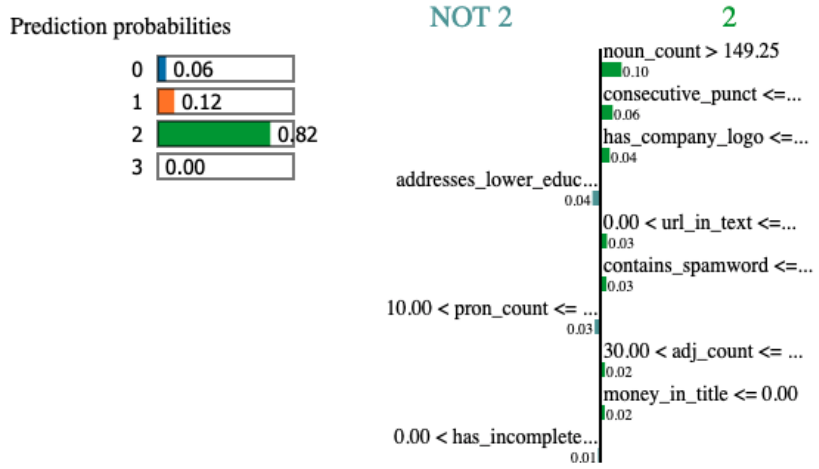

Figure 7. LIME output for an example instance of class 2

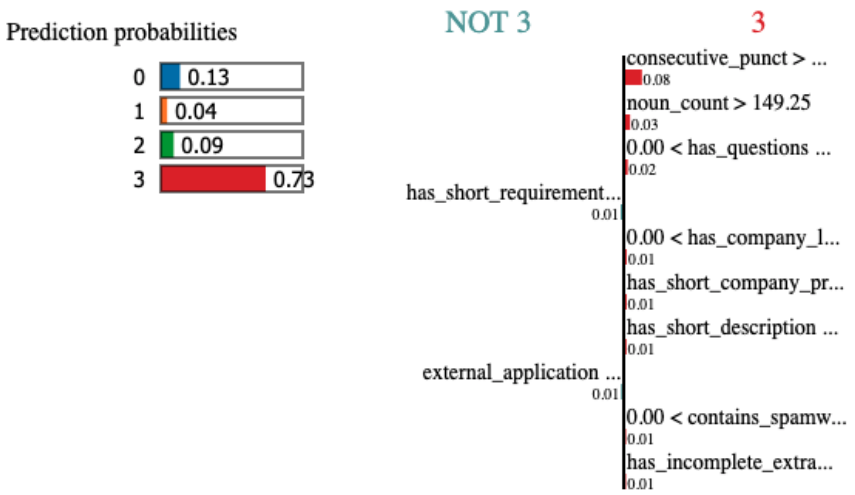

Figure 8. LIME output for an example instance of class 3

4.3 Word Embeddings

In this class of features, word embeddings are learned using the gensim implementation of word2vec. Word embeddings have become increasingly popular in a variety of natural language processing (NLP) tasks, including text classification (Li, Drozd, Guo, Liu, Matsuoka & Du, 2019). Vector based word representations are able to learn semantic or syntactic information contained in unannotated text (Sienčnik, 2015). Word2vec in particular uses either a continuous bag of words model or a skip-gram model in order to derive these word vectors (Ma & Zhang, 2015), and achieves state-of-the-art results in a variety of NLP tasks (Li et al., 2019).

First, the text is cleaned by removing non-letters, converting to lowercase and split at whitespace. Stop words are not removed at this stage in order 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.

4.3.1 Word2Vec Embeddings

The average word vectors are used as features for a multi-class classification task. This model was trained and tested on a 80/20 split of the data using the full set of classifiers. 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 but generally worse than the bag-of-words models. See table 8 for the classification report of the Gradient Boosting classifier.

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 8. Classification report for Gradient Boosting fitted on word2vec embeddings

4.3.2 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 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 9 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.

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 9. Classification report for Random Forest fitted on combined embeddings and ruleset features

4.4 Transformers

In recent years, transformer-based models have gained attention for their advantage over traditional word embeddings. Through language modelling, such transformers are able to learn contextual representations of words and capture additional higher-level semantics. Pre-trained transformer models are trained on generic corpora and consequently fine-tuned for specific downstream tasks (Wolf et al., 2019), e.g. the most commonly used transformer-based model, BERT, is trained on millions of words from Wikipedia and BooksCorpus (Devlin, Chang, Lee & Toutanova, 2018). BERT uses mechanisms such as Masked Language Modeling to learn bi-directional contexts (Acheampong, Nunoo-Mensah & Chen, 2021).

In order to explore whether the contextual word embeddings can offer improved performance on the fraudulent job type classification task, a selection of standard transformer models are trained on the corpus. This thesis 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.

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 10. XLNet yielded better performance, with a weighted average F-score of 0.84, which is comparable to the word2vec embeddings combined with the rule-set features.

Transformer Model	Weighted Avg. F1-Score	Matthews Corr. Coeff.
BERT (bert-base-cased)	0.758	0.657
RoBERTa (roberta-base)	0.830	0.757
XLNet (xlnet-base-cased)	0.844	0.778

Table 10. Evaluation results for 3 transformer models

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 11 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.

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

Table 11. Classification report for XLNet

4.5 Summary

In this chapter, the research design was executed for each of the feature classes and the multi-class classification results were evaluated. A summary of the best classification results per each class of features is collated in table 12. 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.

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

Table 12. Summary of classification results

5 DISCUSSION

In this chapter, the results are interpreted and preliminary conclusions are drawn. First, the research questions are revisited and answered according to results obtained in the prior chapter. Second, the main limitations of this thesis are discussed to add context to the findings. Lastly, implications of the findings are explained.

5.1 Research Questions

What lexical, syntactic, semantic or contextual features can distinguish different types of recruitment fraud in the EMSCAD dataset?

Four classes were explored: bag of words, ruleset-based, word embeddings and transformers. All features classes were able to distinguish the different types of online recruitment fraud. 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.

The first class, bow and tf-idf feature sets, classified the fraudulent job types by the lexical composition of the job description. Transforming the bag of words per the relative importance of each word increased the classification performance in some cases. Overall, these vector representations of the job descriptions achieved high performance relative to other feature classes. This supports the notion that fraudulent job types can be distinguished by the content of their job descriptions.

The second class consisted of handcrafted features derived from the work by Vidros et al. (2017); these features were originally developed through empirical analysis of the dataset with the goal to distinguish between legitimate and fraudulent jobs. Despite being developed for a different classification task, these features still showed marginal propensity to predict fraudulent job type. The relatively low performance of the handcrafted features when tested in isolation indicates that there is room for more work towards expanding the empirical ruleset.

The next two classes of features were not explored in previous works in this research area. The third class consisted of word2vec embeddings trained on the EMSCAD dataset; this incorporated semantic information into the model. The word embeddings performed worse than the bag-of-words models, and this could be attributed to the fact that either the word2vec parameters used were not optimal, or the bag-of-words models overfit the data and delivered inflated performance metrics. The fourth class of features extended the semantic information of word embeddings by introducing additional context. BERT, RoBERTa and XLNet achieved satisfactory performance.

Which features are most informative in predicting the type of fraudulent job?

On its own, the SVM classifier fitted on the tf-idf representation of the job description performed best. Ultimately, it was demonstrated that better performance can be obtained by building a model from a combination of feature classes: a Gradient Boosting classifier fitted on a model 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 observed that the rule set-based features were not sufficient on their own to achieve satisfactory classification results. In all cases, augmenting the ruleset feature space with additional

features such as 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.

Do novel word embeddings or transformer models outperform handcrafted features?

Word embeddings and transformer models consistently outperformed the handcrafted rule-set based features class. The exception was when handcrafted rule-set features were augmented with bag of words representation of the text. The better performance was reflected in the classification reports of the third and fourth feature classes, where configurations which consisted of word embeddings vectors or transformer-based models achieved a higher weighted average F1-score. 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.

What machine learning algorithm performs the best for fraudulent job type classification?

Several classifiers achieved similarly good performance: Random Forest, Gradient Boosting, Logistic Regression, Support Vector Machine (SVM) and Stochastic Gradient Descent (SGD. However, this was dependent on the class of features that were used to build the model. SVM and SGD performed better with the bag-of-words representation. Ensemble learning methods, such as Random Forest and Gradient Boosting, performed better in the ruleset and word2vec models.

5.1 Limitations

The approach pursued by this thesis has a few limitations, particularly in the dataset and methodology.

First, in terms of the dataset, steps taken during the anonymisation process may have altered the data in a way that modified original linguistic relationships or meanings in the job descriptions. For example, all personally identified information was masked with the same word. Hence, any model would not be able to distinguish between an e-mail address or a real name. At the same time, the time-bound and platform-specific aspect of the dataset might bring into question whether or not 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, particularly as they have potentially been able to adapt to prior research findings, might pose a challenge as it pertains to expanding the dataset with new entries. Applicant Tracking Systems may require the input of different attributes or fields today, which could complicate the expansion.

Another limitation as it pertains to the dataset is the diversity of fraudulent job advertisements and the consequences that this has on sample size and class imbalance. The large majority of fraudulent jobs were annotated as type 1, while type 2 and 3 fraudulent jobs were much more sparse. Additionally, many jobs in the type 2 and 3 classes were duplicates, as a result of the fraudster posting the same job at different points in time. This increases the likelihood that the classifiers are overfitting on the dataset. 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, which can ultimately have another negative impact on the generalisability of the findings.

In terms of methodology, the results could have benefited from a more exhaustive cross-validation. While this thesis applied hold-out validation to minimise concerns of over-fitting or selection bias, there still exists some uncertainty about the robustness of certain models due to the fact that hold-out validation does not average out performance measures over multiple runs. Because of this, there exists the possibility that conclusions about certain configurations are not valid. For example, the GB classifier having the highest score with a combination of feature sets might be a consequence of selection bias where different results would be obtained with a different partition of the data.

5.2 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 thesis 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.

6 CONCLUSION

This thesis developed and validated 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. The fraudulent job advertisements from the publicly available EMSCAD dataset were manually annotated into the three distinct types. Four classes of features were utilised: empirical rule set-based features, bag-of-word models, word embeddings and transformer models, which were trained with various machine

learning classifiers. The model was validated by evaluating it on a publicly available job description dataset. The results indicate that the Gradient Boosting classifier with empirical rule-set based features, part-of-speech tags and bow vector achieved the best performance with an F1-score of 0.88.

The results of the classification, as well as feature correlations, indicate that the identified job types can be distinguished on certain contextual and/or linguistic features. Besides word embeddings, 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 and refer to money more often. On the other hand, multi-level marketing vacancies used more consecutive punctuation and asked questions.

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. Based on this social aspect, the major contributions of this thesis have been submitted to the Special Issue on AI for People, part of AI & Society - Journal of Culture, Knowledge and Communication (Springer).

6.1 Future Work

Given the costly nature of this issue, it is important to derive solutions for the general public that can combat such fraudulent recruitment behaviour, and this necessitates that machine learning models are sufficiently valid and generalisable. There is a need to form a more comprehensive and scientifically-validated typology of fraudulent jobs. A more thorough understanding of fraudulent job characteristics will not only facilitate more effective feature extraction and classification, but can also provide valuable explainability and transparency for the benefit of the public.

In addition, future work in this area can focus on exploring different approaches to learn contextual and semantic information about the different types of fraudulent jobs that go beyond the rule-set based approach. Future research might consider additional natural language processing techniques such as Latent Semantic Analysis, the ELECTRA transformer, or GloVe and fastText for word embeddings. When exploring different methods, attention should be paid towards delivering interpretable and explainable results.

Such approaches should be robust, dynamic and able to adapt to changes in fraudsters' activities; after all, publication of any results in the public domain will allow any actors to attempt to circumvent state-of-the-art fraud detection methods, especially if they are based on clear-cut rules. On that same note, there is ample room to explore more unsupervised approaches to this issue; for example, future work can reconceptualise or validate the types of fraudulent job advertisements identified in this thesis by applying clustering.

Above all, this thesis 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.

7 REFERENCES

Acheampong, F. A., Nunoo-Mensah, H., & Chen, W. (2021). Transformer models for text-based emotion detection: a review of BERT-based approaches. *Artificial Intelligence Review*, 1-41.

Alghamdi, B., & Alharby, F. (2019). An intelligent model for online recruitment fraud detection. *Journal of Information Security*, *10*(03), 155.

Anita, C. S., Nagarajan, P., Sairam, G. A., 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* (pp. 268-286). IGI Global.

Chin, S. C., Street, W. N., Srinivasan, P., & Eichmann, D. (2010, April). Detecting Wikipedia vandalism with active learning and statistical language models. In Proceedings of the 4th Workshop on Information Credibility (pp. 3-10).

Devlin, J., Chang, M. W., Lee, K., & Toutanova, K. (2018). Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*.

Dutta, S., & Bandyopadhyay, S. (2020). Fake Job Recruitment Detection Using Machine Learning Approach. *International Journal Of Engineering Trends And Technology*, *68*(4), 48-53. doi: 10.14445/22315381/ijett-v68i4p209s

FTC. (2021). *Job Scams*. Federal Trade Commission Consumer Information. Retrieved 20 March 2021, from https://www.consumer.ftc.gov/articles/job-scams.

Goyal, N., Sachdeva, N., & Kumaraguru, P. (2021). Spy the Lie: Fraudulent Jobs Detection in Recruitment Domain using Knowledge Graphs. *Knowledge Science, Engineering And Management*, 612-623. https://doi.org/10.1007/978-3-030-82147-0_50

Herath, T., & Herath, H. S. (2020). Coping with the new normal imposed by the COVID-19 pandemic: Lessons for technology management and governance. *Information Systems Management*, 37(4), 277-283.

Joyce, S. (2021). *5 Major Types of Scam Jobs and Job Scams Online - Job-Hunt.org*. Job Hunt. Retrieved 7 May 2021, from https://www.job-hunt.org/job-search-scams/.

Kim, J., Kim, H. J., & Kim, H. (2019). Fraud detection for job placement using hierarchical clusters-based deep neural networks. *Applied Intelligence*, *49*(8), 2842-2861.

Lal, S., Jiaswal, R., Sardana, N., Verma, A., Kaur, A., & Mourya, R. (2019, August). ORFDetector: ensemble learning based online recruitment fraud detection. In *2019 Twelfth International Conference on Contemporary Computing (IC3)* (pp. 1-5). IEEE.

Li, B., Drozd, A., Guo, Y., Liu, T., Matsuoka, S., & Du, X. (2019). Scaling word2vec on big corpus. *Data Science and Engineering*, *4*(2), 157-175.

Ma, L., & Zhang, Y. (2015, October). Using Word2Vec to process big text data. In *2015 IEEE International Conference on Big Data (Big Data)* (pp. 2895-2897). IEEE.

Mahbub, S., & Pardede, E. (2018). Using Contextual Features for Online Recruitment Fraud Detection. In B. Andersson, B. Johansson, S. Carlsson, C. Barry, M. Lang, H. Linger, & C. Schneider (Eds.), *Designing Digitalization (ISD2018 Proceedings)*. Lund, Sweden: Lund University. ISBN: 978-91-7753-876-9.

Mehboob, A., & Malik, M. S. I. (2020). Smart Fraud Detection Framework for Job Recruitments. *Arabian Journal for Science and Engineering*, 1-12.

Ptel, M. (2020). Social Posting in Covid-19 Recruiting Era-Milestone HR Strategy Augmenting Social Media Recruitment. *Dogo Rangsang Research Journal*, 10(6), 82-89.

Reynolds, B. (2021). 14 Common Job Search Scams and How to Protect Yourself | FlexJobs. Retrieved 25 January 2021, from

https://www.flexjobs.com/blog/post/common-job-search-scams-how-to-protect-yourself-v2/

Ribeiro, M. T., Singh, S., & Guestrin, C. (2016). "Why Should I Trust You? ": Explaining the Predictions of Any Classifier. *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, 1135–1144. https://doi.org/10.1145/2939672.2939778

Scamwatch. (2015). *Jobs & employment scams*. Australian Competition and Consumer Commission. Retrieved 20 March 2021, from 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. Retrieved 20 March 2021, from https://www.sec.gov/oiea/investor-alerts-bulletins/investor-alerts-ia_pyramidhtm.html.

Sienčnik, S. K. (2015, May). Adapting word2vec to named entity recognition. In *Proceedings of the 20th Nordic Conference of Computational Linguistics (NODALIDA 2015)* (pp. 239-243).

Sivathanu, B., & Pillai, R. (2018). Smart HR 4.0–how industry 4.0 is disrupting HR. *Human Resource Management International Digest*.

Underwood, C. (2020, March 9). *'This is really sick': Unemployed Calgarians fed up with online job scams.* CBC. https://www.cbc.ca/news/canada/calgary/unemployed-calgary-job-seekers-online-scams-1.5484202

Vidros, S., Kolias, C., & Kambourakis, G. (2016). Online recruitment services: Another playground for fraudsters. Computer Fraud & Security, 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.

Wolf, T., Debut, L., Sanh, V., Chaumond, J., Delangue, C., Moi, A., ... & Rush, A. M. (2019). Huggingface's transformers: State-of-the-art natural language processing. *arXiv preprint arXiv:1910.03771*.

Zhang, W., Yoshida, T., & Tang, X. (2011). A comparative study of TF* IDF, LSI and multi-words for text classification. *Expert Systems with Applications*, *38*(3), 2758-2765.

8 APPENDICES

8.1 Appendix A

Type	Job Advertisement Text
0	Telnet is New Zealand's largest privately owned contact centre outsource provider, servicing a diverse range of national and international blue chip clients from their contact centre in Auckland CBD. The team are looking for a Technical Writer to join their team to support their in-house development teams, compliance teams and general users. Reporting to the Compliance Manager, you will need to liaise with IT, Operations, Finance, HR and Client Services teams as well as upper-management. You will be responsible for: - Creation of user documentation of in-house developed software - Development of process flowcharts and process documents - Writing and editing press releases, marketing materials and internal corporate communications - Development of technical policies and procedures - Version control of technical documentation In carrying out this role you will also be expected to be a subject matter expert in the software tools you'll be documenting and the tools used to for these documents. You will need to take ownership of outcomes and negotiate to achieve timeframes and be capable of working well in a team. Skills and experience: - Proven technical writing and editing skills - Ability to meet robust deadlines - Proficient in Microsoft Office applications - Proven ability to create unique and relevant content - Experience with a variety of writing styles: technical, creative and marketing are preferred - Experience in web design and computer graphics preferred To apply please click 'apply' below and attach a copy of your CV. Please note, we will require a sample of your portfolio upon interview. We require all applicants to complete a Ministry of Justice criminal record check and drug screen. CallCentre People Recruitment is recognised as being specialists within the CallCentre industry. We provide permanent, temporary, contract and management staff for a number of large national and multi-national businesses in various industries.
1	If you have experience in financing for auto sales and a great attitude you can work in our Hazelcrest office. From $ top $ a week by contract. prior car sales exp prior car loan financing exp profit sharing car allowance company car

	Looking for adventurous people to join a thriving industry. We offer training and competitive earnings. Find out why imports are the way to go and view our cars at our website.
2	Corporate overview NAME_MASKED Solutions is a global provider of products, systems and services to the oil and gas industry. Our engineering, design and technology bring discoveries into production and maximize recovery from each petroleum field. We employ approximately , people in about countries. Go to #URL_MASKED# for more information on our business, people and values. We are looking for individuals who are prepared to take a position. Not only a position within NAME_MASKED Solutions, but also a position on the exciting challenges the global oil and gas industry faces now and in the future NAME_MASKED Solutions is a leading global provider of engineering and technology, products and service solutions to the Oil ; Gas industry. At NAME_MASKED Solutions we offer an ocean of opportunities. Our people are our biggest asset and our business relies on their abilities to win projects and execute them to the highest standards. We are committed to developing our people's capabilities through challenging tasks supported by excellent training and development opportunities. All our major achievements are team efforts. We are looking for dedicated team players who like to be part of a winning team, who meet challenges head on to serve our customers' needs. Responsibilities and tasks •Act as Quality Manager for each project by supporting, aligning and communicating with the project team in meeting client quality related requirements•Ensuring that any changes in the requirements are adequately captured, communicated and implemented•Adhere to and support the quality management system compliant with the requirements of #URL_MASKED#.•Be the client point of contact for quality related matters•Undertake pre-PEM (Project Execution Model)/SEM (Service Execution Model) gate review audit with project /service team•Attend all project PEM/SEM gate reviews, ensuring compliance to the PEM/SEM process and procedures •Ensure compliance to the NAME_MASKED business risk process directly and by influence on the project team•Create metrics and reporting for and on behalf of the project team for both internal and external communication•Conduct Internal, Project, inter-company or external quality audits as required by the audit schedules•Facilitate project audits by clients or inter-company representatives•Review lessons learnt register to identify those lessons that are beyond project specific and communicate to others and produce Lessons Learnt bulletins as required•Support any new Quality Initiatives as required.•Identify and launch the "Just Care" approach where quality related events would benefit.•Support Project Manager to ensure that the quality culture is maintained throughout the project life-cycle.•Action, review and close out Project specific and general Quality SYNERGI cases.Qualifications ; personal attributes •A recognised Quality Assurance or Quality Management qualification or HNC/HND or equivalent in engineering discipline •Lead Assessors Course with examination pass (recognised by IRCA)•Formal training in the ISO StandardsCandidates/job-holders without the relevant formal qualifications above but possessing other academic or vocational qualifications or who can demonstrate a greater level of relevant practical experience with a proven track-record may be considered equally competent.•Create a culture of

	continuous improvement.•Encourage a high level of performance in self and others.•Be able to establish, maintain and develop customer relations.•Provide a high level of internal and external customer satisfaction.•Work as part of a team and exercise tolerance and consistency when dealing with others.•Be a self starter, capable of working on own initiative in order to achieve tasks and overcome problems as well as provide direction for others.•Proactive, flexible and decisive with the ability to be innovative and challenging in line with Company values.•Be accommodating and receptive to change.•Good time-management and organisational skills•Confident communicator - verbal and written.•Good contractual and commercial awareness.•Good presentation skills
3	We have the demand. We are looking for people that are quick learners, and are very efficient, to handle the demand. We have the best mortgage protection program in the business, and we have serious demand nationwide; especially in the // area. After all, we are in a market where % of all Americans, if they added up all the insurances that they hold, would not have enough to cover their mortgage. Can you imagine a product that everybody NEEDS (not just wants), and all you have to do is meet with people that have already requested our service. No Cold Calling or Door Knocking! We are looking for someone that is very professional and is able to learn quickly, because you can't make six figures in a year by moving slowly. We will train you in every aspect of the business, and will show you how to personally produce a monthly commission of at least 7,,. We have the tools, and we are looking to bring on someone that we can groom into management. You will learn this business, and eventually we plan to teach you how to build a strong staff, that you will train to move product the same way that you do. This combination should easily push you into a yearly compensation of k+ within your first 3 years in the business if you move at a steady pace. Industry experience is not necessary, but a track record of success is. Our company mails out over 1.5 million letters each week and our homeowners fill out a questionnaire and mail back the request for coverage to us. We simply call that exclusive lead and set up an appointment with them. We meet the customer in their home and go over what type of mortgage protection the homeowner needs and then write up the policy. We also market and promote fixed indexed annuities that solve most of the baby boomers retirement issues. We have many full time agents making over 6 and 7 figure incomes! We are primarily looking for those that desire to move into management, though, if you are looking for part time or full time warm sales, please send your resume as well. We have a serious demand all over the // area that consistently outweighs our number of agents. Typically our employees make k+ part time, k full time, k+ Management. Though, we design a specific plan to make sure you hit whatever compensation YOU seriously desire, and are willing to work for. What we are looking for: * A passion to honestly help families. * Positive attitude with a strong desire to become wealthy. * A person who can follow a step by step selling system. What we are not looking for: *Someone who isn't accountable to their word. *Someone who says they are a great salesman, but doesn't have a penny to their name. *Someone who has had a recent foreclosure or bankruptcy (you can't get licensed in those cases) We need you and you need us. Contact us with your resume (preferred), or call the Hiring Manager with

	your information for possible interview.• Compensation: k+ part time, k full time, k+ Management

Table A-1. Sample of Job Descriptions sorted by Type

8.3 Appendix B

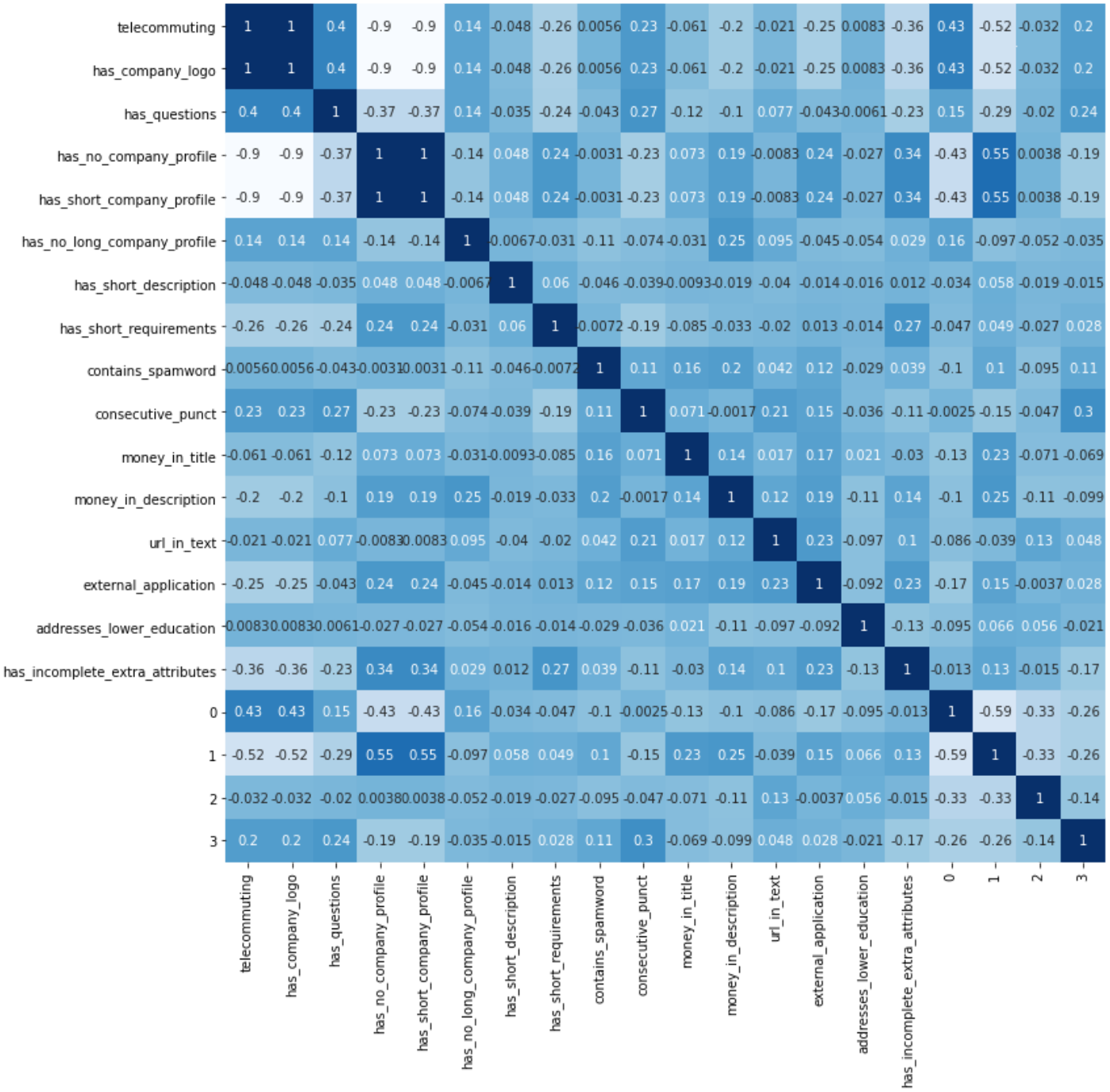

Figure A-1. Correlation matrix for ruleset-based features.

8.2 Appendix C

Maastricht University Leading in Learning!

Official statement of original thesis

By signing this statement, I hereby acknowledge the submitted thesis (hereafter mentioned as "product"), titled:

A Machine Learning Approach to Detecting Fraudulent Job Types

to be produced independently by me, without external help.

Wherever I paraphrase or cite literally, a reference to the original source (journal, book, report, internet, etc.) is given.

By signing this statement, I explicitly declare that I am aware of the fraud sanctions as stated in the Education and Examination Regulations (EERs) of the SBE.

Place: Maastricht, the Netherlands

Date: 08 August 2021

First and last name: Marcel Naudé

Study programme: Business Intelligence and Smart Services

Course/skill: EMTH1200 Master's Thesis / EBS4040 Writing a Master's Thesis

ID number: i6108892

Signature: Marcel Naudé