AUTOMATIC FRAUD DETECTION ON EMPLOYMENT SCAM DATASETS USING MACHINE LEARNING

LIU, HONG YANG

RESEARCH PROJECT SUBMITTED TO THE FACULTY OF COMPUTER SCIENCE AND INFORMATION TECHNOLOGY UNIVERSITY OF MALAYA, IN PARTIAL FULFILMENT OF THE REQUIREMENTS FOR THE DEGREE OF MASTER OF DATA SCIENCE

2021

UNIVERSITY OF MALAYA

ORIGINAL LITERARY WORK DECLARATION

Name of Candidate: **LIU,HONGYANG** (I.C/Passport No: **EA6641098**)

Matric No: **17201091/1**

Name of Degree: **MASTER OF DATA SCIENCE**

Title of Project Paper/Research Report/Dissertation/Thesis ("this Work"):

Automatic Fraud Detection on Employment Scam Datasets Using Machine Learning

Field of Study: **Machine Learning**

I do solemnly and sincerely declare that:

(1) I am the sole author/writer of this Work;
(2) This Work is original;
(3) Any use of any work in which copyright exists was done by way of fair dealing and for permitted purposes and any excerpt or extract from, or reference to or reproduction of any copyright work has been disclosed expressly and sufficiently and the title of the Work and its authorship have been acknowledged in this Work;
(4) I do not have any actual knowledge nor do I ought reasonably to know that the making of this work constitutes an infringement of any copyright work;
(5) I hereby assign all and every rights in the copyright to this Work to the University of Malaya ("UM"), who henceforth shall be owner of the copyright in this Work and that any reproduction or use in any form or by any means whatsoever is prohibited without the written consent of UM having been first had and obtained;
(6) I am fully aware that if in the course of making this Work I have infringed any copyright whether intentionally or otherwise, I may be subject to legal action or any other action as may be determined by UM.

Candidate's Signature LIU,HONGYANG Date: December 23, 2020

Subscribed and solemnly declared before,

Witness's Signature Dr. Maizatul Date: December 23, 2020

Name: Dr. Maizatul

Designation: Supervisor

AUTOMATIC FRAUD DETECTION ON EMPLOYMENT SCAM DATASETS USING MACHINE LEARNING

ABSTRACT

Online recruitment enables the process of sourcing suitable candidates to become effective and efficient, from which job seekers could post their resumes online and search for jobs with little face-to-face interaction. However, there is an emerging challenge that fraudulent job adverts have become a severe issue that may influence the credibility of websites, threaten the privacy information security of individuals, and damage the reputation of organizations. Some fraudsters pretend to become employers to recruit job seekers on online websites and post job advertisements to deceive victims into sending in their resumes, which might contain sensitive information such as name, telephone number, email, and address. Besides, online recruitment website managers are overwhelmed, as identifying the fraudulent recruitment postings is a rather challenging problem and it is almost impossible for website managers to distinguish the fraud job advertisements manually. This research objective is analyzing the Employment Scam Datasets and building a Fraud Detection model with the help of Python Libraries to automatically detect the online fraud recruitment scam. Multiple machine learning classification techniques were investigated using LR, DT, KNN, SVM, Naïve Bayes. Besides, we also applied ensemble learning methods such as Random Forest, AdaBoost, Gradient Boosting, Xgboost to improve the model performance. The results showed that Random Forest outperformed other algorithms and achieved accuracy at 93%. For comparison, the result is slightly higher than the previous one. This model is beneficial to recruitment websites to effectively detect fake job advertisements, protect job seekers from privacy information leakage, and prevent fraudsters from damaging the reputation of companies.

Keywords: Machine Learning Classification, Recruitment Fraud Detection

Pengesanan Penipuan Automatik pada Kumpulan Data Penipuan Pekerjaan Menggunakan Pembelajaran Mesin

ABSTRAK

Pengambilan dalam talian membolehkan proses pengambilan calon yang sesuai menjadi berkesan dan cekap, dari mana pencari kerja dapat menghantar resume mereka dalam talian dan mencari pekerjaan dengan sedikit interaksi tatap muka. Namun, ada tantangan yang muncul bahawa iklan pekerjaan palsu telah menjadi masalah berat yang dapat mempengaruhi kredibiliti laman web, mengancam keselamatan maklumat privasi individu, dan merosakkan reputasi organisasi. Sebilangan penipu berpura-pura menjadi majikan untuk merekrut pencari kerja di laman web dalam talian dan menyiarkan iklan pekerjaan untuk menipu mangsa untuk menghantar resume mereka, yang mungkin mengandungi maklumat sensitif seperti nama, nombor telefon, e-mel, dan alamat. Selain itu, pengurus laman web pengambilan dalam talian merasa kewalahan, kerana mengenal pasti catatan pengambilan palsu adalah masalah yang agak mencabar dan hampir mustahil bagi pengurus laman web untuk membezakan iklan pekerjaan penipuan secara manual. Objektif penyelidikan ini adalah menganalisis Kumpulan Data Scam Pekerjaan dan membina model Pengesanan Penipuan dengan bantuan Python Libraries untuk secara automatik mengesan penipuan pengambilan penipuan dalam talian. Beberapa teknik klasifikasi pembelajaran mesin disiasat menggunakan LR, DT, KNN, SVM, Naïve Bayes. Selain itu, kami juga menerapkan kaedah pembelajaran ensemble seperti Random Forest, AdaBoost, Gradient Boosting, Xgboost untuk meningkatkan prestasi model. Hasil kajian menunjukkan bahawa Random Forest mengatasi algoritma lain dan mencapai ketepatan pada 93%. Sebagai perbandingan, hasilnya sedikit lebih tinggi daripada yang sebelumnya. Model ini bermanfaat untuk merekrut laman web untuk mengesan iklan pekerjaan palsu secara berkesan, melindungi pencari kerja dari kebocoran maklumat privasi, dan mencegah penipu merosakkan reputasi syarikat.

Kata kunci: Pengelasan Pembelajaran Mesin, Pengesanan Penipuan Pengambilan

ACKNOWLEDGEMENTS

I would like to thank my supervisor, Dr. Maizatul Akmar Binti Ismail, for her patient guidance through all stages of my research. I have received enormous support from my supervisor and her profound knowledge and valuable opinions are helpful to me during my writing process. I also get prompt responses and suggestions even at times of most inconvenience.

I would like to give thanks to all the teachers during my study at the University of Malaya. I have acquired knowledge and obtained skills from their professional teaching skills.

I would express my gratitude to my family and classmates. They accompanied me in the past days and helped me a lot when I am facing difficulty in my study and life.

I also would like to thank the staff at the Faculty of Computer Science and Information Technology. They are very patient and helpful when I am facing technical troubles. Thanks for them providing me this platform to finish my study.

TABLE OF CONTENTS

LIST OF FIGURES

LIST OF TABLES

LIST OF SYMBOLS AND ABBREVIATIONS

ORF	ONLINE RECRUITMENT FRAUD
NLP	NATURAL LANGUAGE PROCESSING
KNN	K-NEAREST NEIGHBOR
SVM	SUPPORT VECTOR MACHINES
LR	LOGISTIC REGRESSION
DT	DECISION TREE
NB	NAÏVE BAYES
RF	RANDOM FOREST
EMSCAD	EMPLOYMENT SCAM AEGEAN DATASETS

CHAPTER 1:INTRODUCTION

1.1 Background of Research

Nowadays, online recruitment becomes an increasing trend to recruit candidates. The online recruitment takes place online and using tools to gather resumes or conduct interviews and give feedback to applicants. A survey conducted by Brandão shows that 46% of job seekers had the experience that used OR to find jobs (Brandão et al., 2019). Meanwhile, there are many online recruitment websites, such as LinkedIn, Monster, or Indeed, emerging to simplify the recruitment processes. These websites are quite familiar for job seekers to find job vacancies. Human Resources in different organizations or companies will put job postings with required skills on these websites, whereas job seekers are required to upload their resumes containing their personal contact information to find jobs. The websites offer the opportunity for matching recruiters and candidates according to their demands. Considering the advantages of online recruitment methods, there is a tendency that more and more job seekers tend to use an online recruitment techniques to find their desired jobs.

Although online recruitment helped job seekers to reduce time and cost of finding the desired job and assist employers to recruit suitable workers(Lal et al., 2019), researchers found that the credibility of the e-recruitment job advertisements is under question, due to some fraudulent recruitment posts(Banerjee & Gupta, 2019). While vacancies posts on websites are aiming to help applicants finding jobs, as business expands, the volume of data on these sites increases rapidly as well. Some job seekers faced issues in which during the job search procedures, job advertisements may embed some fraudulent information in it. Job seekers were likely to be deceived by fraudulent postings, as it is

difficult for them to distinguish the real job publishers from these impostors. Besides that, the release of fraudulent information may also pose security threats to these job seekers, who might mistakenly disclose their personal information in resumes to imposters(Vidros et al., 2017).

Here, Online Recruitment Fraud(ORF) is defined as fraudsters offering fraudulent job advertisements postings to steal personal information and even money, which is taken as a type of cybercrime. ORF had brought not only financial losses to job candidates, it also influenced the reputation of the websites, as job seekers would give their negative reviews on these websites. Researchers have stated that the ORF had been a challenging problem for both genuine employers and job seekers (Lal et al., 2019). Consequently, automatic fraud detection on employment scam plays an important role in helping administrators to distinguish fraud postings on recruitment websites.

1.2 Problem Statement

Online recruitment fraud problems have become a severe issue, as it influences the credibility of websites, threatens the privacy information security of individuals. Besides, it also harms the reputation of organizations(Lal et al., 2019; Vidros et al., 2016). Job posting fraudsters might utilize the online recruitment platform to aggregate personal data from victims. The information may contain people's addresses, email details, and contact numbers. The scammers can utilize the victim's data for further transactions in the black market to earn money or conduct cybercrime activities (Vidros et al., 2017).

Regardless of previous works such as email spams, cyberbullying, and fake news were studied to solve the fraud detection problems (Allcott & Gentzkow, 2017; Bhardwaj & Sharma, 2019; Rosa et al., 2019), online recruitment fraud detection problems, especially recruitment scam detection, is relatively a new area and has not yet got proper study(Vidros et al., 2017)

Furthermore, the solutions conducted by previous research are not adequate in handling the recruitment scams(Vidros et al., 2017). In the job advertisements scam realm, relevant information for jobs, e.g. company name, company logo might be difficult to relate with previous research works in email spams, cyberbullying, and fake news. Consequently, to solve the job advertisements scam problems, it is a need to combine relevant information to build the employment scam classifier model.

Alghamdi and Alharby failed to take company logo as main attributes for detection purpose and have not yet compared various different machine learning algorithms(Alghamdi & Alharby, 2019). In this thesis, I will propose a model to help automatically find out fraudulent job postings. To build this model, it is needed to research company information attributes and compare different machine learning techniques. EMSCAD will be utilized to achieve the research goals. Besides that, different metrics will be applied to evaluate the performance of the models and the model will be used for detecting recruitment scam.

1.3 Research Objective

This research aimed to explore recruitment scam datasets and automatically detect the recruitment scam with machine learning methodologies. The research objectives are:

i. To apply machine learning algorithms (LR, DT, KNN, SVM, NB) for online recruitment detection on publicly recruitment scam datasets and categorize the job postings.

ii. To investigate the effects of the job-related company information in genuine or fraudulent job postings, if it influences the fraud detection model.

iii. To evaluate the performance of ensemble methods i.e. bagging and boosting.

1.4 Research Question

Q1 - Which are the best performance machine learning techniques for ORF detection on the chosen datasets?

Q2 - Does the company information affect the performance of the ORF detection model?

Q3 - Does ensemble learning methods produce better performance compared to other machine learning techniques (LR, DT, KNN, SVM, NB)?

1.5 Research Scope

The scope of this thesis is as follows:

1. The data sets were collected by the Aegean University and are designed to solve the problem of employment scams. The datasets contain around 180000 job ads posted from 2012 to 2014 and manually labeled as legitimate and fraudulent.

2. Recruitment scam detection is one example of the text classification problems, which aims to detect and prevent cybercrime activities. In this article, we would mainly use company information as the main attribute and find out the impact of model performance.

3. The machine learning approach is beneficial to achieve the goal that automatically detects patterns of ORF problems. In this thesis, the main contribution is to propose fraud detection models to classify recruitment into real or fake by utilizing supervised machine learning techniques.

1.6 Research Significance

In this thesis, an automatic ORF detection model is proposed to help recruitment websites to automatically and effectively detect fake job postings based on machine learning approaches. This model will consider attributes of the recruitment postings and different machine learning techniques to improve its performance. After applying the ORF model, the website administers can save a lot of cost and energy on detecting fake job advertisements. Besides that, individuals and organizations on this website will be able to get protected without money loss or privacy information leakage, as illegal users would be found out to prevent them from conducting cybercrime. In addition, the detection model can also prevent fraudsters from damaging the reputation of companies.

1.7 Research Organization

This research work is composed of five chapters. In chapter 1, I will introduce the background information regarding Online Fraud Detection problem. Related works about cybercrime detection techniques are discussed in the second chapter. Chapter 3 represents methodology process and machine learning models that we would utilize in detecting the fake job postings. Chapter 4 shows the results of findings and compares the performance of machine learning approaches based on certain benchmarks. Finally, chapter 5 presents the conclusion and highlights the direction of future work.

CHAPTER 2: LITERATURE REVIEW

The literature review parts mainly introduce several papers related to the online fraud detection field. To the best of our knowledge, employment scam detection is a new field and there are not many papers related to this area(Lal et al., 2019). However, recruitment fraud detection has similarity in other works, such as email spam filtering, identification of fake news, and cyberbullying(Dutta & Bandyopadhyay, 2020). These problems all belong to fraud detection problems, aiming to classify text into two categories. Besides that, these problems have similar solutions with the combination of NLP and machine learning techniques. Consequently, we would review the existing papers about cybercrime activities and compare their machine learning models in this section.

This chapter is divided into five sections. Section 2.1 mainly introduces online fraud detection problems such as email spam detection, fake news detection, and cyberbullying detection. Section 2.2 compares different classification methodologies for solving the above detection problems, and the author, proposed techniques, metrics, and limitations are displayed in one table. Section 2.3 describes machine learning classifiers for building the online recruitment fraud detection model. The ensemble approach is also introduced in Section 2.4. Finally, Section 2.5 summarizes the literature review part.

2.1 Online Fraud Problems

2.1.1 Email Spam

Radhakrishnan in his research workstudied machine learning algorithms for filtering emails as spam or ham. A comparison was conducted between two data mining algorithms: J48 Decision Tree and Naïve Bayes. A publicly available email dataset which contained 3672 ham emails and 1500 spam emails have been utilized in this thesis. The researchers applied pre-processing methods to remove incorrect values and missing values and then use the machine learning algorithms to classify them into spam or ham with the help of Weka. The efficiency and accuracy of the classification classifiers had been taken as metrics to compare the performances of different algorithms. The results of experiments showed that the J48 Decision Tree had a better performance compared to Naïve Bayes. Email classification classifier using J48 Decision Tree also achieved an accuracy of 96.5971% with the size of 400 attributes in 0.06 seconds. However, there is still room for improvement on the efficiency of the classifier by using a combination of classifiers, which can be used for further work (Radhakrishnan & V, 2017).

Rusland compared performances of Naïve Bayes algorithm for e-mail spam filtering on two different datasets: Spam Data and SPAMBASE. The Spam data was collected from different email servers containing 9324 emails with 500 attributes, while the SPAMBASE is much smaller and contains only 4601 e-mail messages with 58 features. The two datasets were compared based on four metrics: accuracy, precision, recall and F-measure. The results of the experiment showed that Naïve Bayes algorithm performed better on SPAMBASE dataset as compared with the Spam Data, although Spam Data had a larger scale. The writer also indicated that the total number of attributes would impact

performances of Naïve Bayes classifier(Rusland et al., 2017). Consequently, the quality of the datasets had a significant impact on the performance of Naïve Bayes model when filtering spam e-mails.

2.1.2 Fake news

The researches (Ajao et al., 2018) in their paper proposed hybrid CNN and RNN models on Twitter posts to detect and classify fake news messages. The methodologies of the work mainly focused on deep learning LSTM models as compared with other two models: LSTM with dropout regularization layers and LSTM with CNN model. The three different models were applied on approximately 5800 tweets, which were labeled as 'rumors' and 'non-rumors'. Finally, the proposed LSTM model achieved 82% accuracy without previous knowledge of the domain and the performance of the LSTM model is better than the other two proposed models. However, Ajao et al. indicated that a novel framework using neural networks required a large amount of training datasets to improve its performance.

Lakshmanarao discussed four machine learning techniques in solving fake news problems in social media by classifying the posts into real or fake ones. In the primary step, NLP techniques were utilized for handling the news data and transforming textual data into vectorized formats. The datasets were derived from Kaggle, containing only four attributes: id, title, text and label. The label feature represents whether or not the instances are fake. Then, a comparison among supervised machine learning models were executed, which were SVM, KNN, Decision Tree and Random Forest, to compare models performance on the cleaned Twitter news data sets. Results of the experiment showed

that Random Forest Classification achieved classifier accuracy at 90.7%, which outperformed accuracy of other models, which were SVM(75.5%), KNN(79.2%), Decision Tree(82.7%). The results indicated that Random Forest algorithm had better accuracy among all other three models in classifying Fake News messages(Lakshmanarao et al., 2019).

2.1.3 Cyberbullying

Hani et al. had proposed two text classifiers: SVM and Neural Network for detecting and preventing social media cyberbullying on different n-gram language models. The cyberbullying dataset was collected from Kaggle website and manually labelled into bullying or non-bullying by the authors. After applying text preprocessing methods with two feature extraction approaches: TF-IDF and sentiment analysis, 92.8% accuracy was yielded using Neural Network with 3-gram model, which outperformed 90.3% accuracy using SVM with 4-grams model. The results of the experiments indicated that NN has better performance than SVM in classifying cyberbullying dataset . However, the size of training data is limited and this would have negatively influenced the performance of the NN classifier(Hani et al., 2019).

Reynolds et al. discussed five machine learning techniques including J48, JRIP, IBK1, IBK3, SMO to automatically detect sentences related to cyberbullying. The datasets were collected from Formspring.me website which allows users to anonymously answer questions. This website contents consist of many bullying words. The questions and answers were crawled and manually labelled into ‘containing’ or ‘not containing’

bullying words by paid workers. To develop the classification model, they utilized the Weka tool to apply machine learning algorithms onto two different processed data sets (NUM: number of bullying words; NORM: the density of bullying words). The results of the experiments showed that the model trained using NORM data set outperforms the model trained by NUM dataset, whereas models using J48 and KNN were capable of achieving true positives with 78.5% accuracy. One of the main limitations in this thesis was that imbalanced dataset (less than10% data contains cyberbullying) gave inadequate evaluation using standard metrics, such as accuracies. This is because the false negative rate was quite high, which caused the learning algorithms to classify the instances into non-cyberbullying which contained over 90% of the total training dataset(Reynolds et al., 2011).

2.2 Comparison of Text Classification Techniques

A comparison table is built to compare different methodologies.

Table 2.1: Comparison of Different Methodologies

Autor	Dataset	Detection problems	Proposed Techniques	Compare Algorithms	Metrics	Limitations
(Radhakrishnan & V, 2017)	Enron Email dataset	Email Spam	Naïve Bayes, J48 Decision Tree	Naïve Bayes, J48 Decision Tree	Accuracy & Efficiency	The efficiency of the classifier has plenty of rooms for improvement.
(Rusland et al., 2017)	Spam Data, SPAM BASE	Email Spam	Naïve Bayes	No compared	Accuracy, Precision, Recall, F-Measure	The quality of the datasets would influenced performance of Naïve Bayes classifier

(Ajao et al., 2018)	Twitter posts	Fake News	Hybrid CNN and RNN models	Long-Short Term Memory(LSTM), LSTM with dropout regularization, LSTM with CNN	Accuracy, Precision, Recall, F-Measure	The number of instances in training datasets is small
(Lakshmanarao et al., 2019)	Emergent Dataset in Kaggle	Fake News	SVM, KNN, Decision Tree, Random Forest	SVM, KNN, Decision Tree, Random Forest	Accuracy	There are no adequate evaluation metrics for evaluating the performance of machine learning techniques
(Hani et al., 2019)	Social Media Dataset in Kaggle	Cyberbullying	Neural Networ, SVM	Neural Network, SVM, Logistic Regression	Accuracy, Precision, Recall, F-Measure	The size of training data is limited
(Reynolds et al., 2011)	Formspring.me website data	Cyberbullying	J48(Decision Tree), JRIP, IBK1(KNN), IBK3(KNN), SMO(SVM)	J48(Decision Tree), JRIP, IBK1(KNN), IBK3(KNN), SMO(SVM)	TP	The feature selection step is subjective and will be reviewed in future work. There are no adequate evaluation metrics for evaluating the performance of machine learning techniques

2.3 Simple Classifiers

In this part, we would review several machine learning classifiers for recruitment fraud detection.

2.3.1 LR

Logistic Regression(LR) is a form of machine learning. It is similar to multiple linear regression, while LR is used to obtain binominal response values(Sperandei, 2014). LR is used to predict variables and classify the values into binary classes instead of continuous values. Hence, LR is usually used for the classification problem used to solve binary classification problem(0 or 1). Besides, LR is easy to implement and very efficient to train in the classification problem.

2.3.2 DT

Decision Tree(DT) could handle both prediction and classification problems. Both continuous and discrete values can be used as independent and dependent variables in DT algorithm (Song & Lu, 2015). DT model is a tree-like structure and is composed of nodes and branches. Each internal node in the DT tree represents a judgment on an attribute, and each branch in DT represents the output of the judgment, and finally, the leaf node represents the classification result. Besides, Song and Lu stated that DT is robust to outliers(Song & Lu, 2015).

2.3.3 KNN

KNN is an algorithm for both regression and classification usage. The KNN classifier classifies the unlabeled instances based on the k surrounding neighbors(Kowsari et al., 2017). by assigning them to its neighbors(Zhang, 2016). For example, given a training dataset, for a new input instance, calculate its k closed instances in the training dataset, and if the most closed k instances belong to a class. Then, the input instance belongs to this category.

2.3.4 SVM

Support Vector Machine (SVM) is a power classifier formally defined by a decision boundary to predict labels based on feature vectors (Huang et al., 2018). It is another commonly used supervised learning algorithm for text classification and regression problems. In classification tasks, it is often used to predict the class label. The SVM has the advantages of solving high-dimensional feature problems; solving machine learning algorithms with small samples and handling non-linear features. However, SVM also has some weaknesses. For example, the efficiency of the algorithm is not high when there are many observations, and it is difficult to find a suitable kernel function for text data.

2.3.5 NB

Naïve Bayes(NB) is a machine learning algorithm and is mainly used to categorize textual data(Abbas et al., 2019). It is designed based on the *Bayesian theorem*. Bayesian theorem assumes the independence of given variables to other feature variables. The NB classifier will calculate the posterior probability of features and select the highest probability as the outcome of the prediction. Hence, an NB classifier is very easy and fast to implement in

comparison with other algorithms. However, NB classifiers also have some drawbacks. For example, an NB classifier assumes independence among predictors, whereas in reality, this assumption is not always correct. Besides, when the number of attributes increases and becomes huge, the correlation between attributes also becomes larger. Hence, this makes the model become complex and the performance of the model would also decrease.

2.4 Ensemble Approach

Ensemble approach is a very powerful tool in machine learning and it combines multiple models together and produces a solution to the computational problem(Muller & Muller, 2012). Dong stated that ensemble methods have better performance than any other single classifier(Dong et al., 2020). There are several commonly used techniques for the ensemble approach, which are bagging and boosting(Breiman, 1996).

Breiman stated the *Bagging* is rather easy to implement as an ensemble algorithm, with a good performance. In this method, different training datasets are randomly selected(without replacement) is made from entire training datasets(Anaissi et al., 2016), and the sub-sampled are used to train the basic model individually in parallel. Finally, an individual classifier would be integrated to take a majority vote of decisions. RF is an ensemble learning algorithm that combines decision trees and follows the bagging technique. (Breiman, 2001). It is composed of a set of decision trees and randomly assigned them to the subset of training data sets. Each decision tree in the forest would perform classification tasks separately and get the individual results. Finally, take the majority votes to decide the final class of the test object. Random forest has several advantages. For instance, it can produce high-dimensional features without pre-

preprocessing and feature selection of the input data; Not easy to overfitting; Train speed is quite fast(Yaman & Subasi, 2019).

Boosting is also an ensemble learning technique. In Boosting, multiple sequential models are created and the latter model depends on the previous one. The difference is that Boosting is iterative. It makes use of models that are influenced by the performance of models that were made before. Each model corrects the errors from the last model(Žižka et al., 2019). We would propose several most commonly used boosting algorithms such as *AdaBoost, Gradient Boost and XGBoost.* AdaBoost is a simple algorithm that follows the Boosting technique and it trains very quickly for the weak learners which are misclassified(Dong et al., 2020). Gradient Boost randomly samples the data and uses the sub-samples to train the individual learner to reduce the residuals from the previous learner and then to form a strong learner that is close to real value(Chen & Guestrin, 2016). XGBoost is a very effective tree Boosting technique in ML algorithms, and the most advantages is its scalability in different scenarios(Kowsari et al., 2019). Besides, the implementation of this algorithm is to reduce overfitting and minimize prediction loss.

.

2.5 Summary

Online Fraud happened with the advent of the Internet, in which people are aiming to trick victims into stealing their personal information or money. In this chapter, 6 different papers which had researched on text classification in different online fraud fields have been discussed. The research mainly focuses on the processes of building the text classification models for online fraud detection problems. Different machine learning techniques have been compared and discussed in the performance evaluation part using scores and metrics. Similarly, ORF is also one of the examples of using text classification, which uses similar solutions with other fraud detection problems that we had mentioned earlier. The process of building the ORF model mainly consists of four parts. This process is illustrated as follows:

i. In the preprocessing part, raw texts will be cleaned using text processing methods such as stop words removal, tokenization, stemming and lemmatization, to remove duplicate or populate missing values;
ii. In feature extraction step, textual datasets will be transformed into a suitable format such as vectors, which is capable of feeding into machine learning techniques for training;
iii. The next step is model building. Different classification algorithms would be utilized to train datasets and construct a suitable model to automatically detect ORF postings;
iv. Finally, the last step is model evaluation. Metrics will be used to measure the performance of different classifiers that was built, in order to distinguish which model is good or not;

Above is a summary of procedures for handling textual data and building machine learning models to automatically classify the postings to be fake or real(Kowsari et al., 2017).

From the literature review, these are two limitations occurred in most of the paper reviewed.

1. Ensemble approach has not yet got proper attention in fraud detection field.

2. The datasets for the text classification problem are often highly imbalanced, and the quality of the datasets dramatically influences the performance of detection models.

We have found that recruitment fraud problems have not yet get proper research.
Besides that, some authors utilized the limited measurement metrics to evaluate the text categorization models and there is still plenty of room for improvements of the recruitment fraud detection model(Vidros et al., 2016).

CHAPTER 3:METHODOLOGY

In this chapter, the specific experiment methodology process will be explained. A working flow framework is proposed to build the recruitment fraud detection model detecting and preventting recruitment fraud activities. The detection model starts with scam data collection, and the details shown in Figure 3.1.

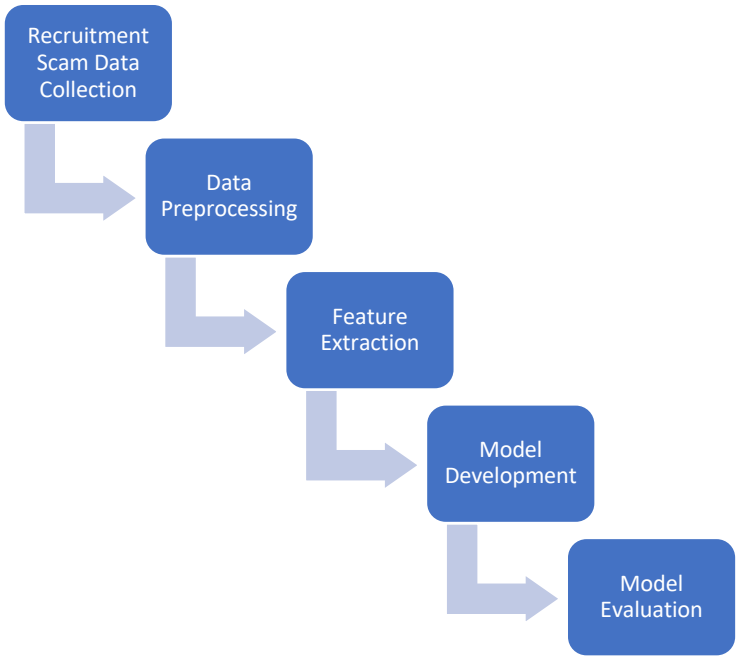

Figure 3.1: The Working Flow for Building Recruitment Fraud Detection Model

3.1 Introduction

In the first section, an overview of data origin will be presented for a better understanding of the 17 attributes details in the dataset. Then, after performing resampling methods, data preprocessing methods for removing unnecessary or noisy values will be explained to eliminate errors in further steps. Thirdly, text transformation methods will be introduced to transform textual data into vectors, which made textual data suitable to feed data into

estimators. Finally, after implementing all classification algorithms, an evaluation table will be built to compare the different models and evaluate the model performance using scores and metrics.

3.2 Data Collection

The employment scam dataset was originated from Aegean University and it was collected from its official website named Emscad(Kowsari et al., 2019). The public dataset contains 17014 genuine and 866 fake job postings in a total of 17880 real-life ads collected from 2010 to 2014. Besides that, the dataset was stored in csv format file consisted of 17 different attributes.

For better understanding the datasets, the description table of dataset is built as follows:

Table 3.1: Dataset Description

Format	Name	Description
String	Title	Job title
	Location	Job location
	Department	Job department
	Salary range	Salary Range
	Company profile	Company description
	Description	Job description

Text	Requirements	The Requirement of the job
	Benefits	The benefits offered by recruiters
Nominal	Telecomputing	Is telecomputing positions or not(False, True)
	Company logo	Has company logo or not(False, True)
	Questions	Has questions or not(False, True)
	Fraudulent	Is fraudulent or not(False, True)
	Employment type	The job employment types(e.g. Full-time, Part-time)
	Required experience	The job required experience (e.g. Internship, Mid-senior level, Entry level)
	Required education	The job required education level of the employment position (e.g. Bachelor's Degree, Associate Degree, Certification)
	Industry	The job industry(e.g. Marketing, Computer software, Media)
	Function	The job function(e.g. Consulting, IT, Engineering

Table 3.1 shows that the employment dataset will be classified into three categories based on data format which are ‘String’, ‘Text’ and ‘Nominal’. ‘String’ format features such as ‘Title’, ‘Location’, ‘Department and Salary range’, were short phrases and described the specific attributes of the jobs.

The textual dataset such as Company profile, Description, Requirements and Benefits consisted of HTML tags and long sentences. Textual dataset is not suitable for automatic text classification using machine learning techniques. Hence, some preprocessing and transformation steps are needed before feeding data into algorithms. These procedures on how to process textual dataset will be explained in further sections.

The last data set category is nominal type attributes, such as company logo, fraudulent, employment type, etc. These attributes mainly described the company information and specific requirements of the job postings, which can be grouped into categories for calculating the percentage of instances. However, as nominal data type can not be ordered and measured, it is also impossible for machine learning techniques to directly apply to nominal data. Hence, preprocessing methods for nominal data features are proposed for the next step.

Figure 3.2 and Figure 3.3 show the examples of fraudulent job postings and genuine job postings respectively. The examples include all three categories of data format, in the process of designing the recruitment detection model. Regardless of the effort and support website managers have provided some methods to identify the recruitment scam, there are always cases of failing due to the high similarity between legitimate and fraudulent postings. The high similarity between legitimate and fake job postings may negatively influence the accuracy of the model.

title	location	company_...	description	has_comp...	fraudulent
IC&E Technician	US, , Stocton, CA	<img src="#URL_a...	<img src="#URL_ae07dc35dfe86ebc1101b48eec77574871f2ba0daaf971ad3f2e2246bee5b31d#"> <p...	t	t
Forward Cap.			The group has raised a fund for the purchase of homes in the Southeast. The student on this proje...	f	t
Technician Instrument & Controls	US	<img src="#URL_044fce3aa43cecf7fd7f1fd790ab4a1f937f268e543cd5f08b2dbd6e52143406#"> Edison...	Technician Instrument & Controls Location Deweyville, TX Location Name...	t	t

Figure 3.2: Fraudulent Data Overview

title	location	company_...	description	has_comp...	fraudulent
Marketing Intern	US, NY, New York	<h3>We're Food52, and we've created a groundbreaking and award-winning cooking site. We support, con...	Food52, a fast-growing, James Beard Award-winning online food community and crowd-sourced and cur...	t	f
Customer Service - Cloud Video Production	NZ, , Auckland	90 Seconds, the worlds Cloud Video Production Service. 90 Seconds is the worlds Cloud V...	Organised - Focused - Vibrant - Awesome! Do you have a passion for customer service? Slick...	t	f
Commissioning Machinery Assistant (CMA)	US, IA, Wever	Valor Services provides Workforce Solutions that meet the needs of companies across th...	Our client, located in Houston, is actively seeking an experienced Commissioning Machinery Assist...	t	f

Figure 3.3: Genuine Data Overview

3.3 Data preprocessing

Many datasets contain noise and unnecessary instances, such as missing values, misspelling values, stops words, and many more. Data preprocessing phase took place in cleaning data instances and preprocessing data into suitable machine learning algorithms. Since noisy data might bring adverse effects to the performance of machine learning models, in this section, different data preprocessing methods are proposed for handling textual and nominal data respectively.

3.3.1 Textual data preprocessing

Table 3.2: Data Pre-Processing Techniques for Textual Data

Techniques	Explanation
Word Tokenization	The text paragraphs or sentences are broke into a list of smaller tokens such as words.
Stop words removal	The steps of removing stop words means filtering out the tokens from a set of stop words
Stemming	The steps of stemming means reducing the words to their root words.
Lemmatization	The steps of lemmatization is another approach to remove inflection by utilizing vocabulary and morphological analysis.

Table 3.2 introduces some data pre-processing techniques for textual data. The first preprocessing step for textual data is word tokenization. In this step, paragraphs or sentences are tokenized into a list of small chunks. The smaller chunks might be words, phrases or symbols which were called tokens. Next, since stop words did not contain important meaning in the machine learning cycle, they will be removed in this step. The main goal of stop words removal phase is to filter out noisy words, such as {“the”, “then”, “not”, “we”, “a”} from tokens. This process will be achieved with the help of a stop words resource list. Last and not least, stemming and lemmatization are both linguistic normalization for words, aiming to reduce the words into their root words. The difference between lemmatization and stemming was that stemming normalizes single word without knowledge of content relationships, whereas lemmatization is more complicated, which transforms root words based on structure of the words and context.

3.3.2 Nominal data preprocessing

Many machine learning algorithms could not directly learn from nominal data. It is required for nominal input data sets to be numeric. Hence, in the data preprocessing part for nominal data, the categorical data are required to be converted into numerical form. There are two different techniques for nominal data into a suitable format feeding into machine learning algorithms:

- Integer Encoding:

In integer encoding, the nominal data will be assigned into an integer value to label the nominal category. For example, assigning 0 to people who answered questions and 1 to people who did not answer questions. The same procedure can also be applied to other nominal features. One nominal data will correspond to an integer value. However, this method also has weakness. There is a natural ordered relationship between the integer variables. However, machine learning algorithms might also learn the data without considering the nominal data containing order relationship or not.

- One-Hot Encodings

One-hot encoding is another preprocessing method which transforms nominal features into one-hot numerical array. Then, data will be fed into machine learning estimators. The one-hot encoding creates a binary column for each category. For example, in the question feature, there will be two answers(True or False). Therefore 2 binary columns are needed: {1,0}, representing the True and {0,1} represents the False. Integer encoding might cause low performance of results in machine learning model, as nominal data do not have ordinal relationship. Therefore, we can use one-hot encoding to replace the integer encoding.

3.4 Feature Extraction

In general, texts are unstructured datasets. Feature extraction used in this thesis illustrated a transformation model in the way of converting text into structured features. Firstly, the texts had to be cleaned. Unnecessary characters and words need to be omitted. Then, feature extraction model will be applied to transform features words into numeric types and feed them into machine learning algorithms(Jones, n.d.). Some common techniques of feature extraction will be reviewed, which are *Bag of Words(BOW), TF-ID*.

- BOW

BOW is a very simple way of extracting features from text data. Texts are converted into matrices with the count of word occurrence describing the frequency of unique words appearing in the individual document. In BOW, every row represents the job advert and the column represents the term from the corpus. The cell number represents the frequency count of a particular term. Figure 3.4 shows the matrix of job adverts using BOW methods.

		2	3	5	ability	ability work	able	apply	company	competitive	develop	excellent	experience
0	45	0	0	0	1	5	1	3	0	0	0	0	1	0
1	35	0	0	0	0	2	0	0	0	0	0	0	1	1
2	8	0	0	0	0	0	0	0	0	0	0	0	0	0
3	19	0	0	1	0	0	0	0	0	0	0	0	0	0
4	23	0	1	0	0	1	1	0	0	0	0	2	1	2

5 rows × 533 columns

Figure 3.4: The Matrix of Job Adverts Using Bow

This process can be conducted with the help of CountVectorizer in python. However there are several limitations in building the matrix of job adverts. Huge vectors in BOW

are needed to represent the frequency of words in the individual job advert. This will consume a lot of memory when the length of documents increases. Besides that, vectors representing each job advert may contain many zeros, which increases the complexity of the matrix. On the other hand, the BOW ignores the semantic relationship between words and does not contain the context information. This means that whether the order of sentence changes or not, it gives no impact to the frequency in the BOW model, but in reality, this may greatly influence the meaning of sentences.

- TF-IDF

TF-IDF is another method for feature extractions for text, reducing the impact of common words in the corpus(Jones, n.d.). Figure 3.5 shows the matrix of job adverts using TF-IDF technique. TF-IDF is the multiplication of TF and IDF. TF-IDF highlights the signature word containing high frequency in one document and those containing lower frequency in all the other documents. The signature word contains a higher weight in the TF-IDF algorithm.

			1	2	3	5	ability	ability work	able	apply	benefits	company
0	0.537041	0.0	0.0	0.000000	0.000000	0.039984	0.137416	0.040602	0.114185	0.0	0.0	0.0
1	0.575904	0.0	0.0	0.000000	0.000000	0.000000	0.077469	0.000000	0.000000	0.0	0.0	0.0
2	0.430115	0.0	0.0	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.0	0.0	0.0
3	0.353557	0.0	0.0	0.000000	0.064547	0.000000	0.000000	0.000000	0.000000	0.0	0.0	0.0
4	0.501688	0.0	0.0	0.071235	0.000000	0.000000	0.049209	0.072698	0.000000	0.0	0.0	0.0

5 rows × 665 columns

Figure 3.5: The Matrix of Job Adverts Using Tf-Idf

3.5 Model Evaluation

After building the online fraud detection models using machine learning techniques, model evaluation parameters are crucial to measure the performance of the models. The evaluation parameters are calculated based on the confusion matrix, consisting of four element, such

as True Positive(TP), True Negative(TN), False Negative(FN) and False Positive(FP) (Figure 3.6). In this evaluation step, four different model evaluation metrics will be used to compare performance of tasks. In general, the evaluation models were needed to measure the performance of the machine learning techniques. In this project, four different machine learning metrics will be proposed, and the equation formulas are shown in Table 3.3.

1) Accuracy: The percentage of correctly classified data points over total number of total recruitment records.

2) Precision: The percentage of correctly classified data points over the summation of true positive and false positive recruitment records.

3) Recall: The percentage of true positive classified data points over the summation of true positive and false positive recruitment records.

4) F-measure: Weighted average of precision and recall.

Using several metrics to compare the performance will give more accurate results as compared to using only one metric. For instance, if we only use accuracy as our performance measures, imbalance data will generate a high accuracy caused by false-

positive data points. Hence, using only accuracy may not be the best measure for assessing imbalance classification classes.

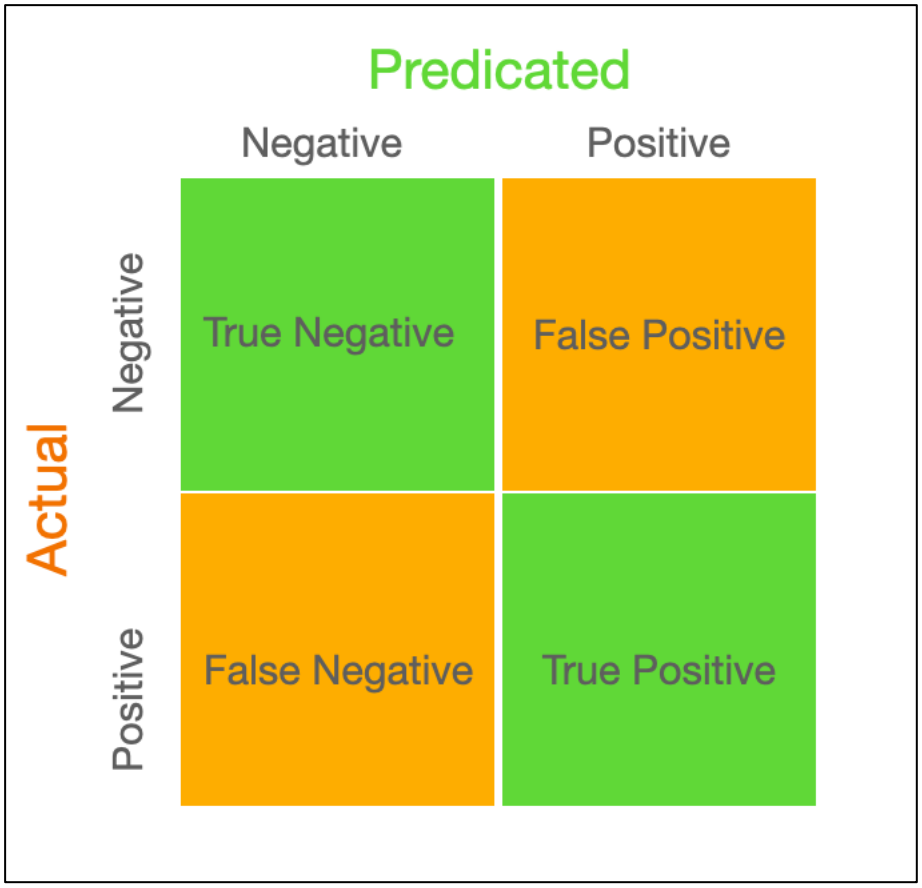

Figure 3.6: Confusion Matrix

Table 3.3: Evaluation Metrics for Machine Learning Models

Metrics	Equation
Accuracy	$\frac{true\ positive + true\ negative}{total\ examples}$
Precision	$\frac{true\ positive}{true\ positive + false\ positive}$

Recall	$\frac{true\ positive}{true\ positive + false\ negative}$
F-measure	$\frac{2 \times precision \times recall}{precision + recall}$

CHAPTER 4: RESULTS AND DISCUSSION

4.1 Introduction

To gain better insights from the dataset, we have applied multiple step experiments on the employment scam data set coming from AEGEAN laboratory which has been aggregated from the university's official website(http://emscad.samos.aegean.gr/). Due to the highly imbalanced data sets, we randomly select 800 fraudulent recruitment records and 800 genuine recruitment records individually to process dirty data containing missing values and duplicated values, which are usually occupied by blanks, 'NaN' or null .

This chapter has 6 sections. In section 4.2, we resample the datasets to balanced datasets and then clean the data to remove outliers, duplicated or missing values. We also transform the text or nominal datasets to suitable format feeding the machine learning classifiers.

In section 4.3, we evaluate the performance of machine learning techniques to determine the best algorithm for the recruitment detection model after removing English stop-words and punctuation, transforming the categorical features into numerical values, and utilizing TF-IDF to transform the text into vectorization.

In section 4.4, we compare the performance of different ensemble approach to observe the improvement. In section 4.5, we apply text analysis on company profile feature and visualize the distribution of company logo to find the insights. Finally, in section 4.6, it is crucial to compare the current research with previous research and find out the differences between them.

The experiments are conducted using Python Library on MacOS Catalina. The processor is Dul-Core Intel Core i5 with 2.7 GHz and the memory for running the Python code is 8 GB.

4.2 Data Preparation

Before the preprocessing parts, The original dataset is highly imbalanced and we need handle the imbalanced datasets to facilitate better results. We apply resampling methods on the original unbalanced datasets to create a new data frame by randomly choosing 800 fraudulent job postings and 800 genuine job postings using resampling methodology.

4.2.1 Resampling Datasets

The original dataset is highly imbalanced and we need to handle the imbalanced datasets to facilitate better results. As shown in Figure 4.1, the datasets are highly imbalanced, and the number of real datasets far exceeds false datasets. These unbalanced datasets may harm the performance model. Hence, we apply resampling methods on the original unbalanced datasets to create a new data frame by randomly choosing 800 fraudulent job postings and 800 genuine job postings using resampling methodology. Figure 4.2 shows the distribution of the dataset after sampling methods.

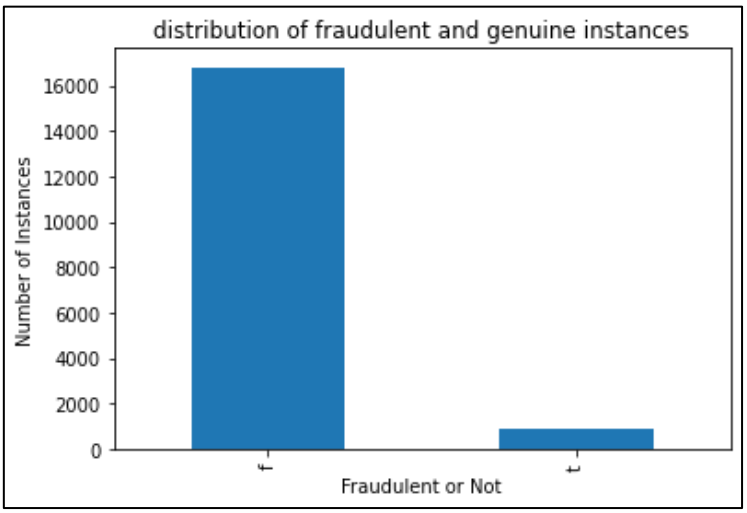

Figure 4.1: Distribution of Fraudulent and Genuine Instances in Original Datasets

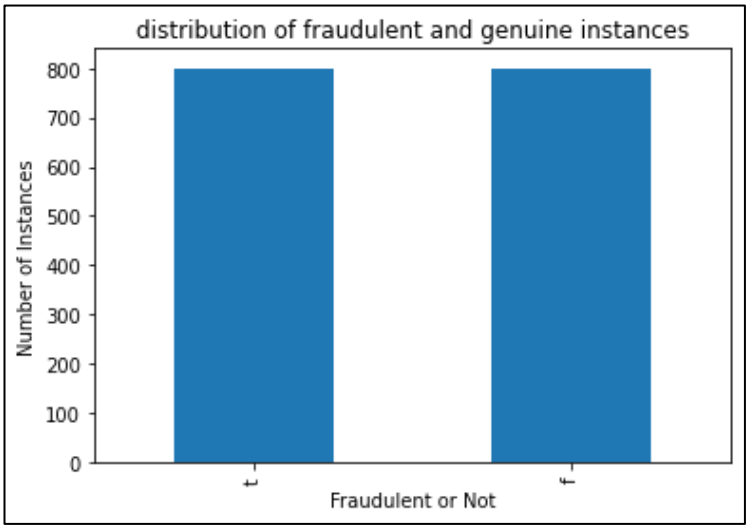

Figure 4.2: Distribution of Fraudulent and Genuine Instances After Resampling

4.2.2 Data Cleaning

In the beginning, we preprocess the datasets to remove dirty values with duplicated values and missing values. For duplicated values, they may reduce the accuracy of the machine learning modeling. To solve the duplicated values, we use Python libraries to remove the values and keep the first occurrence value that appears.

Figure 4.3 shows the dataset overview. Several categorical attributes and HTML fragments are containing missing values. For instance, the missing values NaN in the department and benefits columns in the table. The different types of data attributes bring complexity to preprocessing work. For missing values, it is not wise to directly delete the instances, and if we did that, it would cause the information to be lost. Besides, the proportion of missing values out of the total data is too high.

To reduce the negative impact on the final machine learning modeling and improve the accuracy, we decide to apply imputation methods to impute the corresponding data sets. For instance, the missing values in String attributes like ‘location’, ‘department’, ‘salary_range’ were used to be filled by ‘Unknown’. The missing values in HTML attributes such as ‘company profile’, ‘description’, ‘requirements’ were automatically filled by null string. The missing values in nominal attributes such as ‘employment_type’, ‘required_exprience’, ‘required_education’, ‘industry’, ’funciton’ were used to be filled by ‘Missing’.

	title	location	department	salary_range	company_profile	description	requirements	benefits	telecommutin
0	Marketing Intern	US, NY, New York	Marketing	NaN	<h3>We're Food52, and we've created a groundbr...	Food52, a fast-growing, James Beard Award-w...	\r\n Experience with content management...	NaN	
1	Customer Service - Cloud Video Production	NZ, , Auckland	Success	NaN	<h3>90 Seconds, the worlds Cloud Video Product...	Organised - Focused - Vibrant - Awesome! <br...	What we expect from you: \r\n Y...	What you will get from us \r\n<...	
2	Commissioning Machinery Assistant (CMA)	US, IA, Wever	NaN	NaN	\r\n Valor Services provides Workfo...	Our client, located in Houston, is actively...	\r\n Implement pre-commissioning and co...	NaN	
3	Account Executive - Washington DC	US, DC, Washington	Sales	NaN	Our passion for improving quality of life t...	THE COMPANY: ESRI – Environmental System...	\r\n \r\nEDUCATION: Bachelor's o...	Our culture is anything but corporate—we ha...	
4	Bill Review Manager	US, FL, Fort Worth	NaN	NaN	SpotSource Solutions LLC is a Global Human ...	JOB TITLE: Itemization Review Manage...	QUALIFICATIONS: \r\n \r\n R...	Full Benefits Offered	

Figure 4.3: Dataset Overview

4.2.3 Data Transformation

In the next part, as the machine learning techniques should be fed by numerical values, the text data and categorical features should be transformed into numerical values. Regarding the HTML text fragments, there are many HTML tags such as ' ', '<ul>', '<h3>' embedded in the sentences. We utilize regular expression operators to replace these HTML tags. Besides, to improve the accuracy of the detection model, stop words like 'that', 'a', 'of', 'in ' and punctuations like '!' '#' '.' '*' are removed from the English sentences with the help of NLTK and string libraries in python. Finally, the categories features were encoded into numerical vectors using one-hot encoding methods(Figure 4.4). Finally, text sentences were tokened and transformed into suitable formats to feed the machine learning algorithms.

	country_code_AE	country_code_AU	country_code_BE	country_code_BH	country_code_BR	country_code_CA	country_code_CH	country_code_CN	country_c
0	0	0	0	0	0	0	0	0	
1	0	0	0	0	0	0	0	0	
2	0	0	0	0	0	0	0	0	
3	0	0	0	0	0	0	0	0	
4	0	0	0	0	0	0	0	0	

Figure 4.4: One-Hot Encoding Results

4.3 Result of Experiment

After the data preprocessing part, several features of datasets have been already cleaned. Table 4.1 shows the data sets features are classified into two categories, and one is job information features and the other is company information features. It is hard to know if the company information variables play a significant role in recruitment fraud detection or not. Instead, we can test if the company information has an effect on the prediction model. If the performance of the prediction model has a significant change, it means the company details have an impact on prediction.

Table 4.1: Dataset Character

Category Features	Feature	Character
Job information Features	country_code	US, GB
	employment_type	Full-time, Part-time, Constract
	required_experience	Entry Level, Executive, Intern
	required_education	Doctorate, Master Degree, Bachelor
	industry	Automotive, IT, Health care, Real State
	function	Consulting, Engineering, Research, Sales
	telecommuting	True or False
	has_questions	True or False
	description	Details of job description
	requirements	Requirements of job information
	benefits	Benefits of job information
Company information Features	company profile	Company information profile
	has_company_logo	True or False

There are several features directly or indirectly affecting the experimental results using machine learning algorithms. To analyze the influence of features in the datasets, in this section, we trained the datasets using multiple classifiers such as LR, DT, KNN etc. For validating the impact of company details on the automatic detection model using machine learning algorithms, the training features have been split into two parts, which are features with company information and features without company information. The results would be showed in the table separately : (1) classification results with company information(CI) and (2) classification results without company information(WCI).

Table 4.2: Classification Results With Company Information(CI) and Without Company Information(WCI)

Classification Model	Accuracy		Precision		Recall		F-measure	
	CI	WCI	CI	WCI	CI	WCI	CI	WCI
LR	0.87	0.84	0.87	0.82	0.86	0.86	0.87	0.84
DT	0.87	0.70	0.85	0.70	0.90	0.71	0.87	0.71
KNN	0.80	0.79	0.78	0.78	0.86	0.81	0.82	0.80
SVM	0.86	0.88	0.86	0.88	0.85	0.88	0.86	0.88
NB	0.69	0.67	0.63	0.62	0.91	0.90	0.74	0.73

***CI** Features with company information in the datasets

***WCI** Features without company information in the datasets

Table 4.2 shows the results of machine learning classifiers including five different algorithms(LR, DT, KNN, SVM, NB). We have utilized the 10-fold cross-validation to validate the results and the total number of testing instances is 160 for each iteration. Compared with the performance among algorithms, LR-model and DT-model outperforms other machine learning techniques. LR achieves 87% accuracy with company information and 84% accuracy without company information. The accuracy of 87% means 139 out of 160 instances are correctly classified into right class categories(fraudulent or genuine) and the performance is rather better than the accuracy of features without company information, which means only 134 out of 160 records are correctly classified into right class categories. DT-model also performs well when classifying the recruitment postings, which perform 87% accuracy and 85% precision. Besides, from experimental results, it is essential to notice that the classification algorithms trained with company information have better performance than those algorithms trained without company information in almost all the listed classifiers.

4.4 Ensemble Learning

In this experiment, we also perform ensemble learning techniques aiming to validate whether there is an increment of the performance in the prediction of recruitment postings. The ensemble learning algorithms have two most commonly used techniques which are bagging and boosting. The idea of bagging is combing all the results of models and vote for the best performance model as the final result. Boosting means a sequential process, from which latter model tries to correct the errors from the previous model.

Table 4.3: Ensemble Learning Techniques With Company Information(CI) and Without Company Information(WCI)

Ensemble Model	Learning	Accuracy		Precision		Recall		F-measure	
		CI	WCI	CI	WCI	CI	WCI	CI	WCI
Bagging	Random Forest	0.93	0.89	0.92	0.86	0.95	0.93	0.93	0.89
Boosting	AdaBoost	0.84	0.72	0.86	0.78	0.81	0.61	0.83	0.69
	Gradient Boosting	0.83	0.78	0.85	0.77	0.80	0.79	0.83	0.78
	XGBoost	0.83	0.77	0.85	0.77	0.80	0.78	0.83	0.77

***CI** Features with company information in the datasets

***WCI** Features without company information in the datasets

Table 4.3 shows the experimental results of the ensemble learning model. The bagging model RF outperforms much better than any other boosting models and RF achieve 93% accuracy on features with company information. This means that 149 out of 160 instances are correctly classified. Besides, RF-model also has a higher precision(92%) and the rate

means that those true positive values account for a large part of actual positive values. Finally, the ensemble models with company information have better performance than those models without company information and it shows similar performance results with machine learning algorithms in table 4.2. Overall, the ensemble learning techniques, especially the RF-model, have performance improvement while classifying the recruitment ads.

4.5 Datasets Features Analysis

4.5.1 Binary Analysis

One of the objectives in this thesis is about detecting the impact of company information. The experimental results in section 4.4 show that the company information features have a great impact on the prediction performance of automatic fraud detection models. In this part, we would analyze the two company-related features using different techniques and understand how they influence the results. There are two features in company information, which are 'has_company_logo' and 'company profile'. Specifically, the feature 'has_company_logo' has nominal values including true or false , while 'company profile' contains HTML fragments and the contents in it are English sentences. Then, we would utilize the Python library to visualize the distribution of company logos and analyze company profiles.

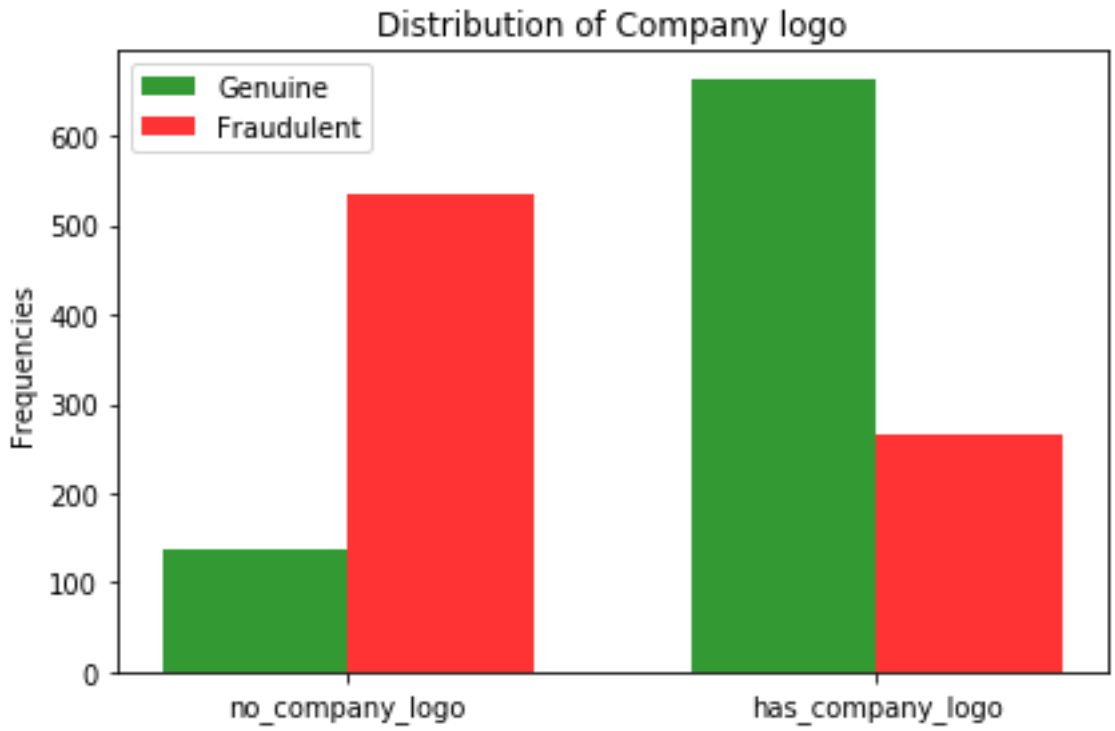

Figure 4.5: Distribution of Company Logo Feature

The diagram (Figure 4.5) is a bar chart that shows the distribution of company logos in recruitment postings datasets. The genuine instances are labelled using green color and the fraudulent instances are labelled using red color. There are two variables in x-axis and they are 'no_company_logo' and 'has_company_logo', which represent recruitment postings with company logos and postings without company logo individually. For the recruitment postings without company logos there are around 120 genuine instances and 530 fraudulent postings, while for the recruitment ads with company logos, the number of genuine instances is about 680, which is more than the number of fake job postings at around 250.

The company logo is often neglected by job seekers when they are finding jobs in the job markets. However, from the above analysis result, we could know that the company logos are an essential factor in determining the credibility of the job postings. The company logo is the first impression of job seekers, and most famous companies have their own company logos. People are easy to distinguish the job ads belong to which company, as the company logo is a brand identity to grab people's attention. However, from the distribution of company logo features, there is much number of fraudulent job posted without company logos. This means that the fraudsters tend to hide the companies' logo information and try to make the candidates neglecting this vital information. Instead, they often use other descriptive text to attract the candidates. Consequently, in the next step, we would also analyze the textual data to gain insights.

4.5.2 Text Analysis

Text analysis is a key part of analyzing English sentences, from which we can find more information or gain insights that are not easy to discover through traditional methods (such as information extraction). Before the text analysis part, we will preprocess the sentences, such as removing HTML tags, punctuation and stop words, converting words to lowercase, and identifying the lemma for each word.

Table 4.4: Most Common Words and Bigrams for Fraud and Genuine Company Profile

		No.	Word	Count	Bigram	Count
Fraud Company Profile		1	candidate	251	signing bonus	86
		2	service	177	leverage career	75
		3	bonus	172	refined resources	74
		4	client	167	aptitude staffing	70
		5	business	159	staffing solutions	70
		6	company	148	candidate enjoy	66
		7	experience	147	employee receive	66
		8	hire	137	represent candidate	53
		9	industry	134	follow perk	53
		10	recruiting	129	perk expert	53
		No.	**Word**	**Count**	**Bigram**	**Count**
Genuine Company Profile		1	company	502	full time	61
		2	work	433	around world	50
		3	team	378	business process	50
		4	service	370	high quality	50
		5	business	330	new work	42
		6	people	296	increase productivity	41
		7	provide	295	document communication	40
		8	customer	266	valor services	40
		9	client	250	long term	36
		10	help	238	medium large	33

Table 4.4 shows the occurrences of common words and bigrams(a two word phrase) in the fraud and genuine company profile individually. Bigrams such as ‘signing bonus’, ‘leverage career’, ‘refined resources’, ‘follow perk’ inside fraud company profile indicate the alarming scam company profile content. While bigrams such as ‘full time’, ‘around world’, ‘high quality’, ‘increase productivity’, ‘valor services’, ‘long term’, inside genuine company profile tend to appear in the genuine job postings. Counting the highest appearing number of words and bigram, fraud postings tend to provide candidates with bonus like money or goods to attract them, while the genuine job postings are more likely to offer these candidates with a good job opportunity and experience, which is full time, high quality and long term.

In addition to that, the word cloud visualization technique is used to better present the top words in the fraudulent and genuine company profile. In the fraudulent company profile, it is easily to find a word such as bonus, resources and while the ‘team’, ‘business’, ‘work’ are more probable to appear in the genuine company profile(Figure 4.6),

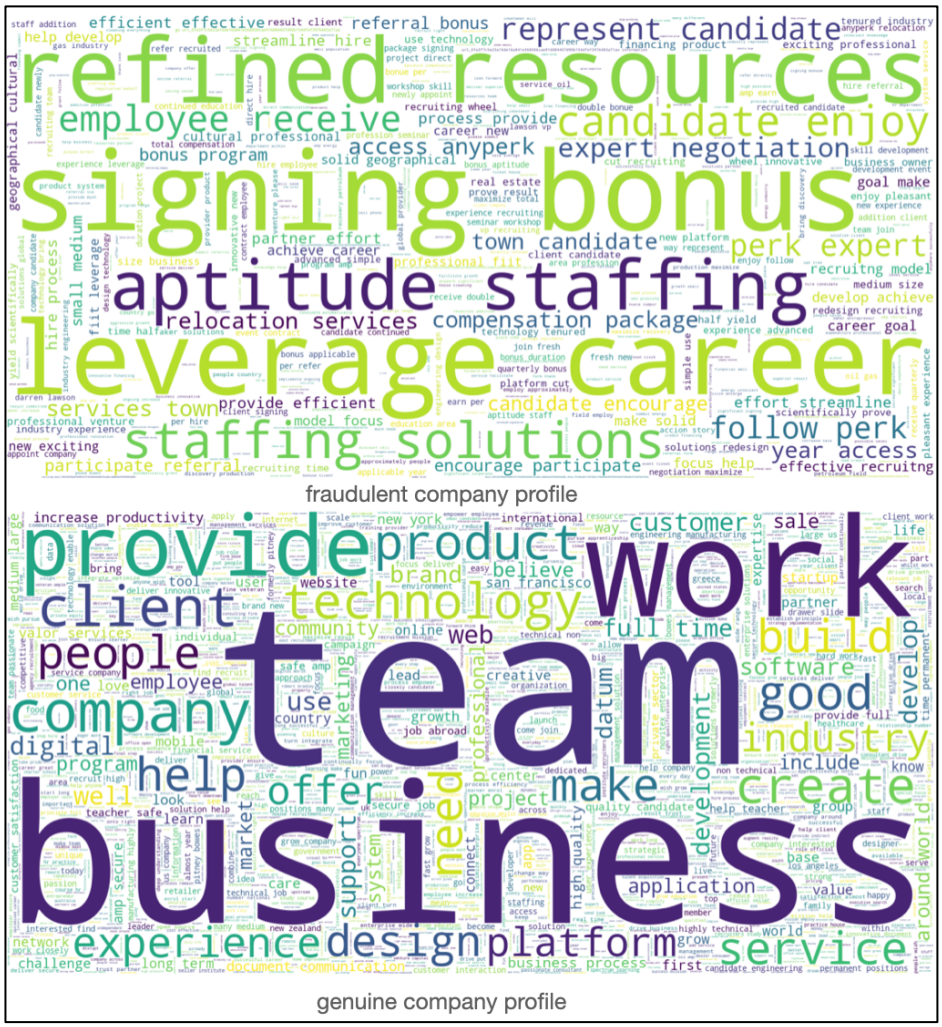

Figure 4.6: Wordcloud for Fraudulent and Genuine Company Profile

4.6 Comparison with previous research

In this section ,we would compare our results and findings with previous studies. Recruitment scam detection is one example of text classification problems, which aims to detect and prevent cybercrime activities. In this thesis, we mainly use company-related information: company logo and company profile as our main attributes to find its impact on the online recruitment model performance. Compared with the detection models to solve other cybercrimes activities such as email spam, fake news and cyberbullying in literature review parts, the online recruitment fraud detection model to solve recruitment scam is a new realm and our research enhances the study in recruitment scam field.

In comparison with previous research in the same field, we obtain the results using machine learning algorithms on the same datasets. Table 4.5 shows the model conducted from previous experiments and evaluated by the accuracy parameter.

Table 4.5: Experimental Result Evaluation With Past Researches

Experiments	Tools	Data sets	Algorithms	Accuracy
Vidros et al. 2017	Weka	EMSCAD datasets	ZeroR	0.50
			OneR	0.84
			Naïve Bayes	0.88
			LR	0.90
			J48	0.91
			Random Forest	0.91
Current research	Python Libraries(Scikit-Learn)	EMSCAD datasets	LR	0.87
			DT	0.87
			KNN	0.80
			SVM	0.86
			NB	0.69
			Random Forest	0.93
			AdaBoost	0.84
			Gradient Boosting	0.83
			Xgboost	0.83

Previous study conducted by Vidros utilized Weka as their experimental tool(Vidros et al., 2017). In this thesis, we have used different tools(Python) to perform on the same datasets. Table 4.5 shows the experimental results we have obtained using Python Libraries. Random Forest is the best performance classifier in both studies. Previous research using ZeroR as baseline algorithm which only has 0.50 accuracy, the other algorithms such as OneR, Naïve Bayes, LR and J48 significantly outperforms the baseline algorithm. RF achieved 91% accuracy in classifying the recruitment job postings.

In this research, we have conducted machine learning classifiers such as LR, DT, KNN, SVM, NB, AdaBoost, Gradient Boosting and XGBoost. The results in the current research show that LR and DT achieve a rather high accuracy while detecting the recruitment fraud postings. Besides, the current research also used ensemble approach algorithms to predict the recruitment fraud postings and they follow the bagging and boosting techniques. RF is an extension of bagging techniques and AdaBoost, Gradient Boosting and XGBoost belong to the boosting algorithms. The RF(with 93% accuracy)

using bagging techniques outperform other algorithms using boosting techniques in building fraud job detection models.

However, although the RF in this thesis has better performance, the accuracy of other models in this thesis (e.g. LR, DT, and NB) is lower than that of previous studies. Vidros's research has some similarities and differences in the analysis process with my work. We both use EMSCAD dataset, which is the only publicly dataset for recruitment fraud detection to my best of knowledge. However, there are still some differences. Vidros's research only randomly chooses 450 genuine and 450 fraudulent job ads individually to analyze, while we increase the datasets and have 800 legitimate and 800 fraudulent job postings individually as our research purpose. The scarcity of data may influence the results of models. On the other hand, analysis tools are different. In Python, we have done a lot of work on preparing the datasets, for example, we have filled the missing values and removed the duplicated values before our research. Besides, we also utilize the NLTK tools to preprocess the sentences and remove outliers.

Last but not least, the feature extraction parts are totally different. Feature extraction illustrated a transformation model in the way of converting text into structured features. We both utilize the feature extraction model applied to transform features words into numeric types and feed them into machine learning algorithms. Vidros's research uses the BOW model and in this thesis, we have utilized the TF-IDF model. The BOW ignores the semantic relationship between words and does not contain the context information, while TF-IDF model emphasizes the signature word and it contains a higher weight. The details between them have been illustrated in Chapter 3.

Besides, we also add another feature determining the recruitment fraud detection problems for comparison. The location of job postings is often neglected by candidates, while it is also one indispensable factor determining whether the job postings fake or not. In Vidros's research work(Figure 4.7), the geographic location of countries is highly biased, and the US holds the most fraudulent entries. In this thesis, Figure 4.8 shows 673 out of 800 fraudulent job adverts are from the US. The results are similar to previous research. Besides, we also found that Vidros's research and this thesis mainly target recruitment fraud in English-speaking countries.

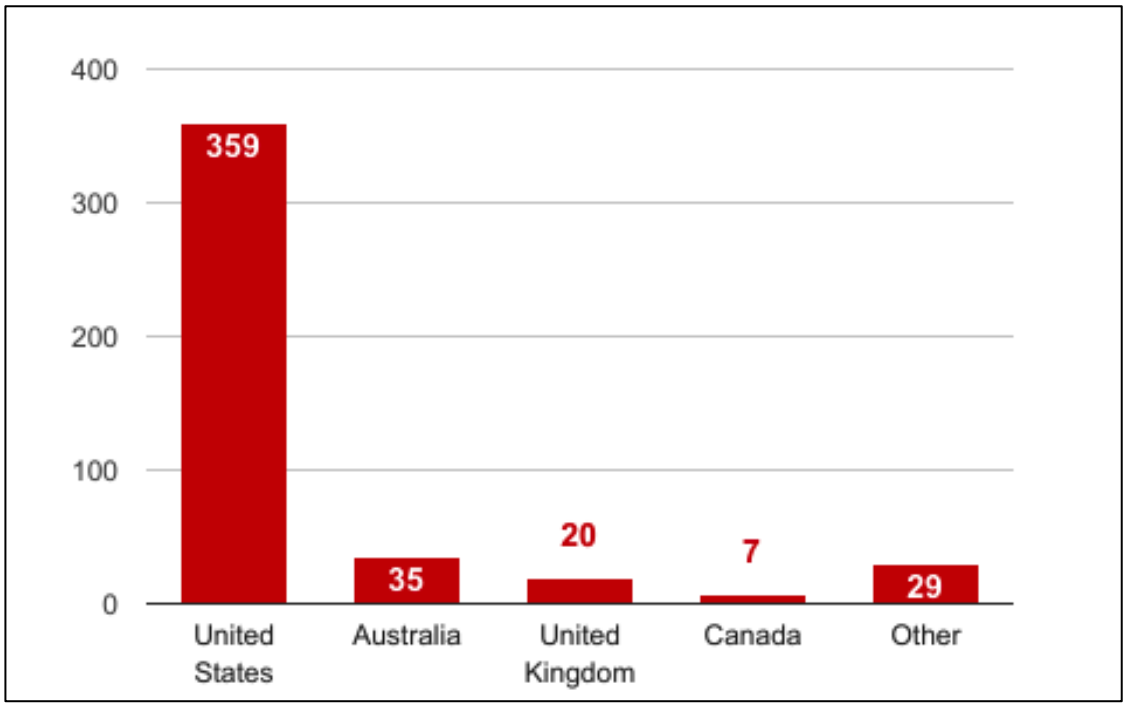

Figure 4.7: Top 4 Countries With Fake Job Ads by Vidros Et Al.(2017)

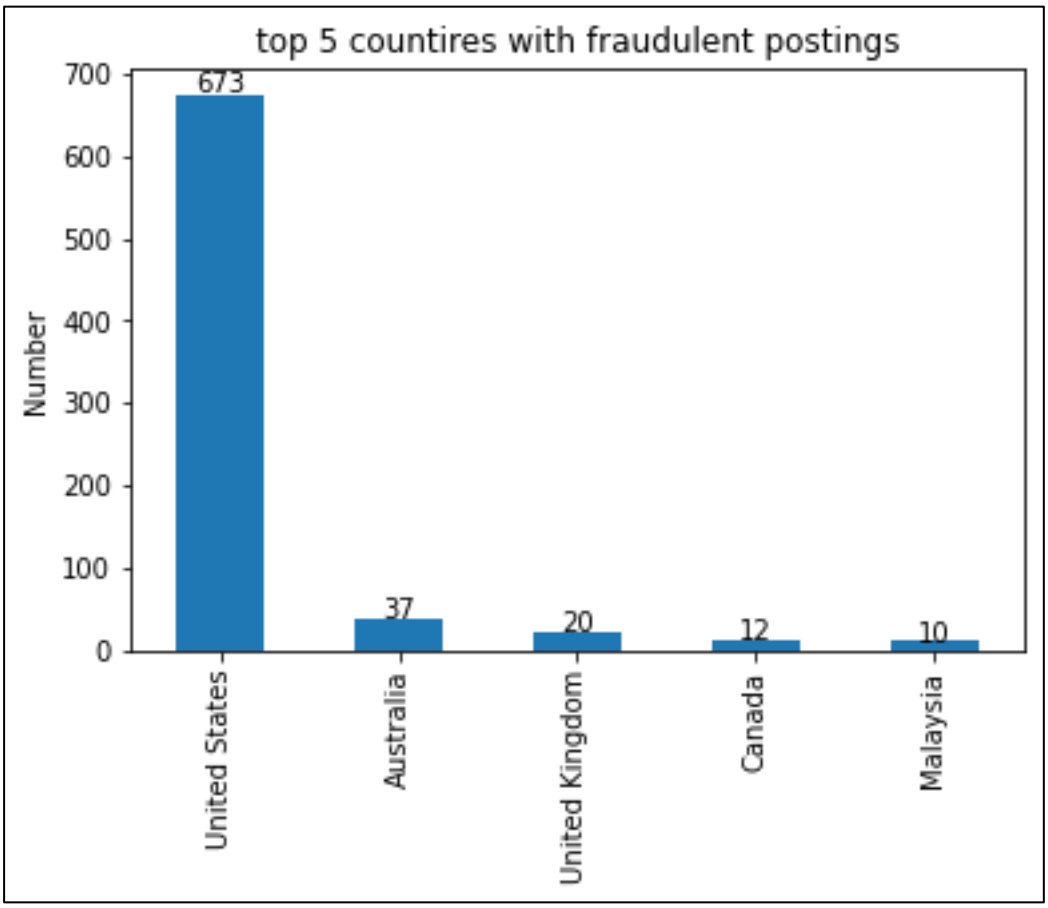

Figure 4.8: Top 5 Countries With Fake Job Ads

4.7 Summary

In this part, we mainly discuss the experimental results using different machine learning techniques. From the experiments showed in table 4.2 and table 4.3, RF outperforms other machine learning techniques. Besides, we test the company-related information variables' effect on the prediction model and found that company information has significantly impacted the results of the prediction model. The RF model trained by features without the company information achieves lower accuracy than the RF model trained by features with the company information. This means company information features are essential in the fraud detection model.

Then, to better understand how the company information influences the performance, we analyze the company information features which are company logo and company profile. In Figure 4.5, bar charts show the distribution of company logos in recruitment postings datasets. The graph shows that there are around 120 genuine postings and 530 fraudulent postings without company logos, as compared with 680 genuine postings and 250 fraudulent postings with company logos. From this data, it is easy to observe that genuine job ads are more likely with company logos, while fraudulent job ads tend to be posted without company logos. Then we justify how the company logo feature determines the credibility of the job ads and find that the fraudsters tend to hide the companies' logo and try to make the candidates neglecting this vital information.

Finally, we analyze the context in company profile using NLTK tools. The company profile only describes the company, while text analysis methods help us mining more information from this plain text. The occurrences of common words and bigrams for fraud and genuine company profile were extracted and showed in table 4.4. From the table, bigrams such as 'signing bonus', 'refined resources', 'follow perk' inside fraud company profile indicate the scam company profile which tends to provide candidates with bonus like money or goods to attract them. On the other hand, Bigrams such as 'high quality', 'increase productivity', 'valor services', 'long term', inside genuine company profile tend to appear in the genuine job postings and provide candidates with a good job experience. Besides, the word cloud visualization technique is used to present the company profile, and the representation of words also achieve similar results.

CHAPTER 5: CONCLUSION AND FUTURE RESEARCH

5.1 Conclusion

The main challenging in our research is the high similarity between legitimate and fake job postings may negatively influence the accuracy of the model and previous research are not adequate in handling recruitment scam. In this research, we build the fraud detection model to help automatically find out fraudulent job postings. The main contribution in our research is taking company features as main attributes for the detection purpose and compared different supervised machine learning techniques to classify recruitment into real or fake. EMSCAD datasets have been utilized to achieve research goals. Besides that, different metrics also were applied to evaluate the performance of the models. This research prevents fraud activities, especially online recruitment fraud problem.

The experimental results show that Random Forest that follows bagging methods outperforms any other algorithms in predicting the employment scam datasets. Besides, we obtained a slightly higher accuracy with previous research conducted by Vidros in classifying the recruitment scam ads.

In this thesis, we have three objective, which are:

i. To apply machine learning algorithms (LR, DT, KNN, SVM, NB) for online recruitment detection on publicly recruitment scam datasets and categorize the job postings.

ii. To investigate the effects of the job-related company information in genuine or fraudulent job postings, if it influences the fraud detection model.

iv. To evaluate the performance of ensemble methods i.e. bagging and boosting.

For objective one, we first preprocessing the data using including one-hot encoding to transform the categorical and HTML fragments into numerical values. Then, we utilize LR, DT, KNN, SVM, NB to train the job posting ads datasets and build the classification models to detect the recruitment fraud postings. Compared with the performance among algorithms, we found that LR and DT techniques outperform other algorithms and achieve accuracy at 87 % respectively. The performance of these two algorithms is slightly lower than previous research conducted by Vidros .

For objective two, we mainly take company profile and company logo in EMSCAD dataset as our main attributes to conduct experimental analysis. For a better understanding of how the company-related information influences the performance, we trained the model using two groups of features (1)features with company information in the datasets (2) features without company information in the datasets. The results show that the company information highly impacts the classification model and the model with company information has higher performance on the detecting model. Then, to better understand how the company information influences the performance, we analyze the company logo using bar charts and company profile using word cloud. We found that the genuine job ads tend to be posted with company logs and attract candidates using work experience in company profile, while the fake job ads are more likely to be posted without company logo and promise money or goods while recruiting the candidates.

For objective three, we utilize several different ensemble learning algorithms bases on bagging and boosting techniques to train the datasets. The idea of bagging methods used for combing all the results of models and vote for the best performance model as the final result, and boosting is a sequential model to minimize errors. We found that Random Forest(bagging) techniques achieve better performance at 93% accuracy and it is slightly better than the fraud detection model trained by machine learning algorithms such as DT (87% accuracy)and LR(87% accuracy).

5.2 Limitations

The highly unbalanced datasets limited the performance of models. To analyze the unbalanced datasets, we use under-sampling techniques to resize the datasets, from which the number of datasets is limited from around 170000 instances to 1600 instances. This technique results in limited data sets also reduces model performance. Besides, the datasets mainly come from English language speaking countries and it is hard to find the pattern or model from other countries. Hence, if we pursue a larger area, and we have to collect more data from other non-English countries.

5.3 Future Research

The scarcity of datasets may not only limit the performance of models, but also negatively affect the various techniques we can apply. For example, deep learning is an emerging technique and has been widely used in the image classification realm, while it needs more data to improve its performance. As the limited datasets collected in this experiment, the neural network would behave poorly than any other algorithms. Consequently, in the future, we would collect more data and train the fraud detection model using deep learning to facilitate better performance.

Furthermore, in this research, we mainly focus on company-related information attributes in the job recruitment postings and find that machine learning classifiers with company information features could behave much better than those without company information. It gives us an insight that we could create or collect more attributes about job ads. For instance, we would collect information about the recruiters and use metrics such as scores to measure the reliability of the recruiters. In future work, we could analyze the new features as our main attributes and validate if or not it could improve the model performance for recruitment detection purposes.

REFERENCES:

Abbas, M., Ali Memon, K., & Aleem Jamali, A. (2019). Multinomial Naive Bayes Classification Model for Sentiment Analysis. *IJCSNS International Journal of Computer Science and Network Security*, *19*(3), 62.

Ajao, O., Bhowmik, D., & Zargari, S. (2018). Fake news identification on Twitter with hybrid CNN and RNN models. *ACM International Conference Proceeding Series*, 226–230. https://doi.org/10.1145/3217804.3217917

Alghamdi, B., & Alharby, F. (2019). An Intelligent Model for Online Recruitment Fraud Detection. *Journal of Information Security*, *10*(03), 155–176. https://doi.org/10.4236/jis.2019.103009

Allcott, H., & Gentzkow, M. (2017). Social media and fake news in the 2016 election. In *Journal of Economic Perspectives* (Vol. 31, Issue 2). https://doi.org/10.1257/jep.31.2.211

Anaissi, A., Goyal, M., Catchpoole, D. R., Braytee, A., & Kennedy, P. J. (2016). Ensemble feature learning of genomic data using support vector machine. *PLoS ONE*, *11*(6). https://doi.org/10.1371/journal.pone.0157330

Banerjee, P., & Gupta, R. (2019). Talent Attraction through Online Recruitment Websites: Application of Web 2.0 Technologies. In *Australasian Journal of Information Systems Banerjee & Gupta* (Vol. 23).

Bhardwaj, U., & Sharma, P. (2019). Email spam detection using ensemble methods. *International Journal of Recent Technology and Engineering*, *8*(3). https://doi.org/10.35940/ijrte.C5485.098319

Brandão, C., Silva, R., & dos Santos, J. V. (2019). Online recruitment in Portugal: Theories and candidate profiles. *Journal of Business Research*, *94*. https://doi.org/10.1016/j.jbusres.2018.04.011

Breiman, L. (1996). Bagging predictors. *Machine Learning*, *24*(2). https://doi.org/10.1007/bf00058655

Breiman, L. (2001). Random forests. *Machine Learning*, *45*(1). https://doi.org/10.1023/A:1010933404324

Chen, T., & Guestrin, C. (2016). *XGBoost*. https://doi.org/10.1145/2939672.2939785

Dong, X., Yu, Z., Cao, W., Shi, Y., & Ma, Q. (2020). A survey on ensemble learning. In *Frontiers of Computer Science* (Vol. 14, Issue 2). https://doi.org/10.1007/s11704-019-8208-z

Dutta, S., & Bandyopadhyay, S. K. (2020). Fake job recruitment detection using machine learning approach. *SSRG International Journal of Engineering Trends and Technology*, *68*(4). https://doi.org/10.14445/22315381/IJETT-V68I4P209S

Hani, J., Nashaat, M., Ahmed, M., Emad, Z., Amer, E., & Mohammed, A. (2019). Social Media Cyberbullying Detection using Machine Learning. In *IJACSA) International Journal of Advanced Computer Science and Applications* (Vol. 10, Issue 5). www.ijacsa.thesai.org

Huang, S., Nianguang, C. A. I., Penzuti Pacheco, P., Narandes, S., Wang, Y., & Wayne, X. U. (2018). Applications of support vector machine (SVM) learning in cancer genomics. In *Cancer Genomics and Proteomics* (Vol. 15, Issue 1). https://doi.org/10.21873/cgp.20063

Jones, K. S. (n.d.). *A STATISTICAL INTERPRETATION OF TERM SPECIFICITY AND ITS APPLICATION IN RETRIEVAL*.

Kowsari, K., Brown, D. E., Heidarysafa, M., Meimandi, K. J., Gerber, M. S., & Barnes, L. E. (2017). *HDLTex: Hierarchical Deep Learning for Text Classification*. https://doi.org/10.1109/ICMLA.2017.0-134

Kowsari, K., Meimandi, K. J., Heidarysafa, M., Mendu, S., Barnes, L., & Brown, D. (2019). Text classification algorithms: A survey. In *Information (Switzerland)* (Vol. 10, Issue 4). MDPI AG. https://doi.org/10.3390/info10040150

Lakshmanarao, A., Swathi, Y., & Srinivasa Ravi Kiran, T. (2019). An effecient fake news detection system using machine learning. *International Journal of Innovative Technology and Exploring Engineering*, *8*(10), 3125–3129. https://doi.org/10.35940/ijitee.J9453.0881019

Lal, S., Jiaswal, R., Sardana, N., Verma, A., Kaur, A., & Mourya, R. (2019). ORFDetector: Ensemble Learning Based Online Recruitment Fraud Detection; ORFDetector: Ensemble Learning Based Online Recruitment Fraud Detection. In *2019 Twelfth International Conference on Contemporary Computing (IC3)*.

Muller, M. E., & Muller, M. E. (2012). Learning and ensemble learning. In *Relational Knowledge Discovery* (pp. 224–250). Cambridge University Press. https://doi.org/10.1017/cbo9781139047869.009

Radhakrishnan, A., & V, V. (2017). Email Classification Using Machine Learning Algorithms. *International Journal of Engineering and Technology*, *9*(2), 335–340. https://doi.org/10.21817/ijet/2017/v9i1/170902310

Reynolds, K., Kontostathis, A., & Edwards, L. (2011). Using machine learning to detect cyberbullying. *Proceedings - 10th International Conference on Machine Learning and Applications, ICMLA 2011*, *2*, 241–244. https://doi.org/10.1109/ICMLA.2011.152

Rosa, H., Pereira, N., Ribeiro, R., Ferreira, P. C., Carvalho, J. P., Oliveira, S., Coheur, L., Paulino, P., Veiga Simão, A. M., & Trancoso, I. (2019). Automatic cyberbullying detection: A systematic review. *Computers in Human Behavior*, *93*. https://doi.org/10.1016/j.chb.2018.12.021

Rusland, N. F., Wahid, N., Kasim, S., & Hafit, H. (2017). Analysis of Naïve Bayes Algorithm for Email Spam Filtering across Multiple Datasets. *IOP Conference Series: Materials Science and Engineering*, *226*(1). https://doi.org/10.1088/1757-899X/226/1/012091

Song, Y. Y., & Lu, Y. (2015). Decision tree methods: applications for classification and prediction. *Shanghai Archives of Psychiatry*, *27*(2). https://doi.org/10.11919/j.issn.1002-0829.215044

Vidros, S., Kolias, C., & Kambourakis, G. (2016). Online recruitment services: Another playground for fraudsters. *Computer Fraud and Security*, *2016*(3), 8–13. https://doi.org/10.1016/S1361-3723(16)30025-2

Vidros, S., Kolias, C., Kambourakis, G., & Akoglu, L. (2017). Automatic detection of online recruitment frauds: Characteristics, methods, and a public datase. *Future Internet*, *9*(1). https://doi.org/10.3390/fi9010006

Yaman, E., & Subasi, A. (2019). Comparison of Bagging and Boosting Ensemble Machine Learning Methods for Automated EMG Signal Classification. *BioMed Research International*, *2019*. https://doi.org/10.1155/2019/9152506

Zhang, Z. (2016). Introduction to machine learning: K-nearest neighbors. *Annals of Translational Medicine*, *4*(11). https://doi.org/10.21037/atm.2016.03.37

Žižka, J., Dařena, F., Svoboda, A., Žižka, J., Dařena, F., & Svoboda, A. (2019). Adaboost. In *Text Mining with Machine Learning*. https://doi.org/10.1201/9780429469275-9