# DETECTION OF FAKE JOB POSTING

# USING

# MACHINE LEARNING APPROACH

ASHUTOSH SHANKAR

(Student Id: 939803)

M.Sc. in Machine Learning & Artificial Intelligence,

Liverpool John Moore's University

Supervisor: Snehansu Sekhar

FINAL THESIS REPORT

JUNE 2021

# CONTENTS

## ACKNOWLEDGEMENTS

Without the competence of our dear Thesis Adviser, Dr. Ahmed kaky, this research paper could not be completed. I also appreciate my supervisor for the thesis Snehansu Sekhar for having time to study my thesis.

A thanksgiving is also to Sokratis Vidros, Constantinos Kolias, Georgios Kambourakis and Leman Akoglu for the journal "Automatic Detection of Online Recruitment Frauds: Characteristics, Methods, and a Public Dataset" which helps me to get inside idea about this domain.

Finally, also I would like to thank my parents Dr. Ashok Kumar Singh and Mr. Virendra Kumar without you I would have not got moral support to complete this.

## ABSTRACT

The crucial recruitment system was just recently transferred to the cloud. The computer programmes in control of completing new hire online recruitment are designed to make the hiring process faster, more reliable, and less expensive. However, the internet disclosure to certain conventional company practises has created new points of vulnerability, which may result in applicants' privacy being abused and organisations' reputations being affected. Companies like Naukri, Monster, Glassdoor, and Indeed constantly working on architecture to identify fake job detection but till this area is not explored much in machine learning approach. The most popular form of Online Recruiting Fraud (ORF) so far has been a job scam. Unlike other forms of online fraud, the issue of ORF has gained less coverage and has remained mostly unexplored until now. To better resolve this current data security issues, the project gathers and articulates the features of cutting-edge and timely knowledge on the topic of cyber security. A release of an extensive dataset of 17,880 annotated work advertisements across the internet that is built up from across the course of time and augmented at the same time serves to aid us in our knowledge. We introduce through this research a model based machine learning which can detect fake job in real-time. While evaluating the machine learning models with Accuracy , Precision, AUC and ROC for classification of fake job detection.

## LIST OF TABLES

## LIST OF FIGURES

## LIST OF ABBREVATIONS

| Abbreviation | Full form |
|---|---|
| SMOTE | Synthetic Minority Over Sampling Technique |
| EMSCAD | Employment Scam Aegean Dataset |
| LSTM | Long Short-Term Memory |
| RQ | Research Question |
| ORF | Online Recruitment Frauds |
| SVM | Support Vector Machine |
| HTML | Hypertext Markup Language |
| ip | Internet protocol |
| LR | Literature Review |
| ROC | Receiver operating characteristic curve |
| AUC | Area under the ROC Curve |
| ReLU | Rectified Linear Unit |
| CBF | Circular Binary feature |
| WFS | Wrapper Based feature Selection |
| IG | Information Gain |
| FSA | Feature Selection Algorithm |
| ANN | Artificial Neural Network |
| RNN | Recurrent Neural Network |
| BN | Bayesian Net |
| XSS | Cross-site scripting |
| IDS | Intrusion Detection System |
| TF-IDF | term frequency–inverse document frequency |
| BOW | Bag of words |
| LMT | Logistic Model Trees |
| LB | Logistic Boosting |
| KNN | K nearest neighbour |
| U2R | User to Root Attack |
| R2L | Remote to Local Attack |

| | |
|---|---|
| KS | Kappa statistics |
| ATS | Applicant Tracking System |
| MAE | Mean Absolute Error |
| RMSE | Root Mean Squared Error |
| RAE | Relative Absolute Error |
| RRSE | Root Relative Squared Error |
| PCFG | Probabilistic Context Free Grammars |
| ED | Euclidean distance |
| MLP | multilayer Perceptron |
| Tfidf/ tf-idf | Term Frequency Inverse Document Frequency. |
| LR | Logistic Regression |
| fp | False Positive |
| fn | False Negative |
| tp | True positive |
| tn | True negative |

# CHAPTER 1: INTRODUCTION

## 1.1 Background of the Study

Job fraud is one of the most serious problems recently tackled in online recruitment fraud. (ORF) In recent days, many organizations have opted to post their openings online so that job seekers can access them conveniently and quickly. However, since they give jobs in favour of getting money from them, job hunters, this purpose could be one form of Scam by fraudsters. It is possible to post fake work ads against a reputed business for breaching their reputation. To acquire an automated method to recognise fake jobs and report them to individuals to avoid applying for such jobs, this fraudulent work post identification attracts great interest.

During recent days we see ascend in fake employment posting where job posting appear quite sensitive, frequently these organisations will also have a designed site and they have enlistment procedure that is same very much in common like genuine organisation. In precise look on these posting it can be segregated from fake or genuine posting, more often it noted these fake posting doesn't have an organisation logo also the underlying reaction from the organization is from not having company domain email account or informal email account in many cases this job scammer has basically two approaches.

First method where lure the applicants to fill application form to create database with ill motivated and sold these data to 3rd party, these data can be Telephone, full name & zip code etc. More sophisticated scammer may sell educational or professional experience details to send bulk email links to increase total page hits containing links, Scammer also contacts website administrator for this.

The second method is to use full data on fraud that could be subsequently utilised as part of a criminal enterprise's economic crimes, example as racketeering of finance and re-shipment of fraud. In this instance, the fraudsters pretend that the job is real or imaginary employer and using the applicant tracking system as a mode for propagating resembles fake jobs. These reports direct users to additional contact techniques (i.e., site, email add. or tele. Num.). They will take part in a range of events from that point on, such as the distribution of fake skills tests, the organizing of fake interviews. For dissemination of encouraging emails in the event of a successful onboarding procedure, etc. One of the final goals is to induce the victim to unwittingly release extremely sensitive information or to act as a money mule by using their bank accounts, which include social security numbers, identity cards, and passports, to aid the suspect in laundering funds. In comparison to related online fraud concerns, ORF has not earned ample coverage yet, to date, it remains largely unknown. and it is clear to see that work scam detection shares similar features with related concerns, Email spamming, phishing, cyber bullying, vandalism on Wikipedia, opinion fraud and trolling for instance.

## 1.2 Problem Statement

A recent well explained research on Online Recruitment Fraud Detection Model was put forward by (Alghamdi and Alharby, 2019) To differentiate scams or fraud objects from the data collection, the model core principle is to use the ensemble based classifier and SVM Algorithm for the feature selection. Another recent research on fake job recruitment detection using machine learning approach was put forward by (Dutta and Bandyopadhyay, 2020) where few classifiers are used, such as the , the Multi-Layer Perceptron Classifier, the K-nearest Neighbour Classifier, the AdaBoost Classifier, the Gradient Boost Classifier are used. First base got proposed in the year 2017 by (Vidros et al., 2017) where several analysis using Bag of Words model, Empirical Analysis, Geography and dataset complete evaluation are shown.

As fake job posting fraud resembles very much to most the online fraud hence Behaviour based email analysis with Spam Detection Application proposed by (Hershkop, 2006), professional classification algorithm of serious unethical activities, for example, inappropriate body dates format of the message or clear contradictions in users' past email behaviour. Another proposed paper in email spam by (Blanzieri and Bryl, 2008) The solutions suggested vary from different protocols for sender authentication to qualified classifiers that differentiate between regular and junk emails.

Another resembles online fraud is phishing which shares most of the common character where author

(Jain and Gupta, 2018) where it does client-side identification of phishing websites using a machine learning method where Elevated accuracy of detection approach proposed by Misclassification of real websites as phishing (false positive) must be the minimum and accurate classification of phishing websites (true positive).

In above proposed paper machine learning approach is applied which employs several classification algorithms for recognizing fake posts. In this case idea to purpose Bidirectional LSTM (Ma and Hovy, 2016) deep learning model to predicate fake job posting.

## 1.3 Aim and Objectives

The main aim of this research is to propose a model to predict fake job posting using EMSCD (Laboratory of Information and Communication Systems, University of the Aegean, Samos, 2021) data also compare of the most popular models available for detecting accuracy, which helps to identify the jobs which are true or fake. The goal of this research to contribute in tackling employment scam detection which guide job seeker to get only legitimate offer from companies.

The research target is formulated on the basis of the purpose of this review, which is as follows:

- Using text, binary and meta tags of HTML analysis to examine the pattern and relationship between distinct features.
- To propose effective balancing approaches that can be applied to the dataset of imbalances.

- To purpose best suited model to predicate fake job posting.
- To compare between the different predictive model and identify the most accurate model among them.

## 1.4 Research Questions

- How to decide the relevant characteristics used in posting fake jobs?
- What is the right algorithm to use to assess the posting of fake jobs?
- Is the Bidirectional LSTM model suitable for the determine fake job posting?

## 1.5 Scope of the Study

### 1.5.1 In Scope

This study stressed that online recruitment requires comparisons between areas such as email spam phishing, cyber bullying, and so on that were historically possibly the best. Furthermore, the only functional EMSCAD free dataset within such a scope and used in this study. In comparison to related online fraud concerns, ORF has not earned ample coverage yet, to date, it remains largely unknown. and It is clear to see that work scam detection shares similar features with related concerns**, Email spamming, phishing, cyber bullying, vandalism on Wikipedia, opinion fraud** and **trolling** for instance. Also, scope of this research is bound to only English fake job predication corpus.

### 1.5.2 Out of Scope

Non-English fake job prediction corpus for is out of scope for this research. Moreover, ip tracking and bot attack (A bot attack is the use of automated online requests to manipulate, deceive, or disrupt a website, application, API, or end-users to get access to their information or to their accounts.) tracking of fakers are kept out of scope for this research

## 1.6 Significance of the Study

From this experiment that we come with several word vec. transformation into model which help model in better accuracy and other evaluation matrices. Our works contributes building better ATS through which community can segregate the fake job posting, during this pandemic where the maximum job loss happen, and many people needs to go through job hunting and then with the help of these models we can lead scammer to get identified.

## 1.7 Structure of the Study

The thesis assumes the following form. The first chapter introduces the research subject and addresses the problem statement. In section 1.3, the study's goal and objectives are addressed, as well as how they relate to the issue statement. The research questions that we are attempting to answer through this research report are presented in Section 1.4. In section

1.5, the research study's scope is explored and explained, as well as what is beyond the scope of this research and study. The importance of the thesis is discussed in section 1.6, which explains why it is relevant and how it can contribute and assist in the science and industry domains. It explains how this research will be beneficial and who will learn from it.

In Chapter 2, you'll learn about previous research on the topic. This chapter is divided into parts that cover some basic observations, scientific findings, and discussion from previous research on fake job detection and related topics. The relevance of a Literature Review and Risk Factor was discussed in Section 2.1. Section 2.2 describes a form of internet scam that resembles a phoney work posting. Section 2.3 included a thorough explanation of data retrieval and interpretation in the context of internet fraud. Section 2.4 further discusses numerous articles in the field of fake job detection and related topics, as well as various technological aspects of fake job detection and related topics, as well as their evolution. The data format structure used in the identification of false work postings is presented in Section 2.5.

The analysis methodology is presented in Chapter 3 which includes details on the dataset used in this thesis as well as the methodology used. Section 3.1 gives an overview of the testing methods. The study method is defined in depth and with some specificity in Section 3.2. The comprehensive testing methodology and architecture are discussed in Section 3.3. This segment addresses numerous machine learning data pre-processing strategies that are relevant to this research design and analysis. This segment also addresses numerous EDAs, including immersive digital analytics and transitions, as well as the Class imbalance strategy utilised in analysis methodology. Finally, in section 3.4, the suggested approach is clarified in detail.

# CHAPTER 2: LITERATURE REVIEW

## 2.1 Introduction

The literature review addresses detailed study queries, while the generic sample paper offers a general overview of the domain. This literature review is carried out according to the instructions in (Group, 2007) , and the key studies are chosen in accordance with (Binyousef et al., 2017). The main goal of this study is to find the best possible function extraction strategies, as well as to present various current models for spam review identification and the parameters that can be used to test these models.

The SLR method aids in determining the various studies accessible in the field of spam analysis identification and answering various study questions. Figure 2.1 depicts the various stages of a comprehensive literature study. The researchers explore how various measures are taken in each process of SLR prior to the analysis.


Figure 2. 1: Various stages of a comprehensive literature study

Lastly, it discusses the best approaches for locating and reviewing primary research, including computerised databases and other methods. It additionally and most particularly elaborately focuses on the various testing issues, methods of inquiry that each question would provide. This thesis looks at various study topics, in this case, approaches that play a significant roles in answering different questions.

within these rules, this study is being performed in order to meet (Group, 2007) , After undertaking a literature review on the word to ensure there are no related tools accessible products, the findings are retrieved.

Following are keyword on which literature review carried out.

('Online recruitment fraud' or 'fake job posting' or 'scammer in job posting') AND (‘website phishing 'or 'phishing’) AND (‘Email spam' or 'spam' or 'ham') AND ('cyber bullying') AND (‘Intrusion Detection’) AND ('vandalism on Wikipedia') AND ("opinion fraud) AND ('trolling for instance')

### 2.1.1. Inspirations for Research Questions

| S.No | Research Questions | Inspiration |
|---|---|---|
| RQ_1 | How to decide the relevant characteristics used in posting fake jobs? | Fake job related various keyword used to assess the common characteristics among phishing, email spam etc. |
| RQ_2 | What is the right algorithm to use to assess the posting of fake jobs? | You may study how various feature engineering techniques, or datasets can support the feature extraction process and then find out how to properly grasp on them. |
| RQ_3 | Which performance indicators are used to assess the effectiveness of methods for detecting fake jobs? | Examine the various metrics that are used to assess the effectiveness of various fake job identification methods. |

Table 2. 1: Research Questions

For the collection of current studies, this literature review used the following criteria:

Criteria - an experiment or study, in which is included or is included in the research. It relates to "the hunt for a fake job" and the specific area of "doctored job descriptions".

Inclusion & Exclusion Criteria - Not many paper published in these areas but paper published to similar area are considered , in the range of 2013 to 2020. Some journal articles, regardless of their subject, must be written in English. There must be no question as to the validity of the study chosen.

Research Question criteria specified - A performance test on bogus employer verification strategies has been conducted for the research article RQ_3 one of the ways this study addresses the issue of work fraud is by the use of techniques to determine "fake job detection" RQ_2.

### 2.1.2. Risk Factor

Various machine learning-based fake job identification strategies have been addressed. However, there is also no single method for detecting different forms of fake jobs. Fake job identification is an example of supervised learning, so the solution's accuracy is calculated by the features collection, machine learning algorithm, and training data. The first problem is a zero-hour assault, since most fake job identification strategies equate features of questionable faker behavior to a pre-defined feature collection. As a result, the system's precision is determined by the features chosen and how precisely they are chosen. Most work portals, such as Naukri, Monster, Glassdoor, and Indeed, have different text languages in different nations, but the interface is nearly identical. Heuristics, such as the commonly

occurring keyword on the work site, are used in machine learning strategies. While these techniques can detect keywords written in English, they may not be able to detect text written in other languages, such as Chinese or Hindi. Embedded items, which may circumvent the fake job identification solution, are often a challenge in the fake job. To get around the fake job identification scheme, the intruder uses photos, JavaScript, and other methods instead of text. Therefore, identifying a fraudulent job that uses an embedded entity remains a problem.

## 2.2. Types of Online Fraud.

Because of the relatively lesser amount of publicity linked to online recruitment fraud(ORF), for this ORF has so far has not been an open to more extensive scrutiny. For your reference, it is a readily evident reality that it shares characteristics with similar practises, such as email spam, website phishing, cyber-bullying, Intrusion Detection, vandalism on Wikipedia and Trolling.

### 2.2.1. Fake Job Posting

Job fraud is one of the most serious problems recently tackled in the area of online recruitment fraud. (ORF) In recent days, many organizations have opted to post their openings online so that job seekers can access them conveniently and quickly. However, since they give jobs in favour of getting money from them, job hunters, this purpose could be one form of Scam by fraudsters. It is possible to post fake work ads against a reputed business for breaching their reputation. In order to acquire an automated method to recognise fake jobs and report them to individuals to avoid applying for such jobs, this fraudulent work post identification attracts great interest.

### 2.2.2. Email Spam

Mass email sent to several users to no particular person or group without prior permission to do so as a feature of the Simple Mail Transfer Protocol (SMTP) makes sure that the identity of the message's source is never becomes questionable, spammers integrate a malicious code and disguise it by sending the email as though they can, and then take advantage of a lack of the dynamic URLs, which is hard to track.

Identity theft may also happens as a result of the use of this as well as a precursor to other forms of internet fraud, including social engineering, fraud on the wire, and relationship scams. E-mail spam filtering has been researched and is useful for detecting a wide variety of issues from routers to mailboxes, the latter from consumers to make them less cluttered Another theory is that people broaden their social network of contacts as they develop new skills and obtain new roles, such as getting a career. That is, in other words, consumers that have set their email contacts to "expand" or comp rather than "contacts only" have a lower risks of receiving unnecessary messages. While we recognise that impersonation and the sender (e.g., spoofing of email addresses) are protocols that can be abused, we additionally recognise that characteristics which give the ability to submit high volumes of email so much credit, such as large volumes of messages too may be compensated for.

### 2.2.3. Website Phishing

When phishing is paired with specialised attack vectors, it guides visitors to unauthorised websites in order to extend to distribute malware; it can promote the spread of XSS(Cross-site scripting) and enables hackers to exploit backdoored scripts to be accessed, it collects confidential information as well. Many laymen users were unable to discern between genuine and non-looking websites, although user-experience testing showed that participants had little trouble discriminating between actual and bogus web sites, at least when it came to detecting phishing web sites

### 2.2.4. Cyber Bullying

It can be characterised as a systemic, deliberate communication by a group or a person, the aim of which is to push the envelope in electronic media. The most likely victims of online abuse are typically be affected by their age, for instance, such as teens who are unable to go to take legal action due to a lack of awareness. In many published paper author tries to solve the issue, the first attempts concentrated on scrutinising each statement for threats and intimidation, instead of looking for patterns of behaviour.

### 2.2.5. Vandalism on Wikipedia

The bulk of the information contained in a crowd-sourced encyclopaedia like Wikipedia is at risk of being altered or completely falsified. This greatly increases the administrative costs as well as the expense of undoing the hard work of reliable administration; as a result, it necessitates the continual material revisions. In most cases, a simple change in names, settings, date markers, or minor errors to one in historical stories aren't thought to be important or particularly noticeable, but some can include rearranging settings, renaming, or adding information that may not have possibly been known at the time the writers. A catalogue of revisions is provided on the web for finding edits depending on spite of how long they have been running, with quick access to their meaning, if it is desired. indicators of vandalism include metadata, such as if an article has been edited anonymously, as well as users of static IP addresses, although the contributions of the use of which an article has been made of can even be monitored to uncover a concerted assault on objectivity.

### 2.2.6. Trolling

Trolls are those who like to deliberately interrupt an on-topic, conversation of online users, specifically attempting to get them angry and upset so as a tactic to get a reaction from them. Bolting down an ill-conceived plugin or trojan to a site's executable code into a website may compromise the quality of the website itself or the mental health of the visitors. Both statistical and syntactic classifiers are educated on various messages previously found to be in app posts that have identified identifiers and then deployed to the public message boards to identify more unusual or inappropriate messages.

### 2.2.7. Intrusion Detection

These modern-day hackers will penetrate by following different trends or styles of attacks; these are harder to spot than they have in the past One good example of this is that some trends for packet scans and scans have a lot of data to be processed; for example, some packet probing attacks can take some time. Owing to the low sensitivity of conventional approaches, it can be impossible to discover that anyone has conducted unauthorised entry. A trademark violation occurs when anyone intentionally misuses the trademark to deceive others about the source of their goods or services. As part of signature-based identification, the unknown pattern will be paired with a known pattern, and then checked for either normal or irregular activity.

Generally speaking, anomaly detection is designed to uncover anomalous behaviour, but as such behaviours deviate from usual The benefits of both are unique, while the disadvantages of each are distinct. Since signature-based detection can work well, it is effective for detecting patterns of known behaviours, it is especially effective for identifying novel attacks. False positives are common in statistical analysis systems because they use statistical analysis techniques and novel attacks are detected by statistical methods, which results in poor precision and significant increases in waste of resources.

## 2.3 Data Extraction and Analysis in online fraud

Primary data was collected with a data extraction form that had been defined in a separate worksheet, to serve as a reference. Data was extracted from primary studies using a data extraction form, mentioned under the Data Extraction section. A more detailed figure highlights the kinds of studies that come under the expansion headings "Data Extraction" will contribute to. Expanding on Figure 2: This plot reveals the distribution of studies with regards to publication year.

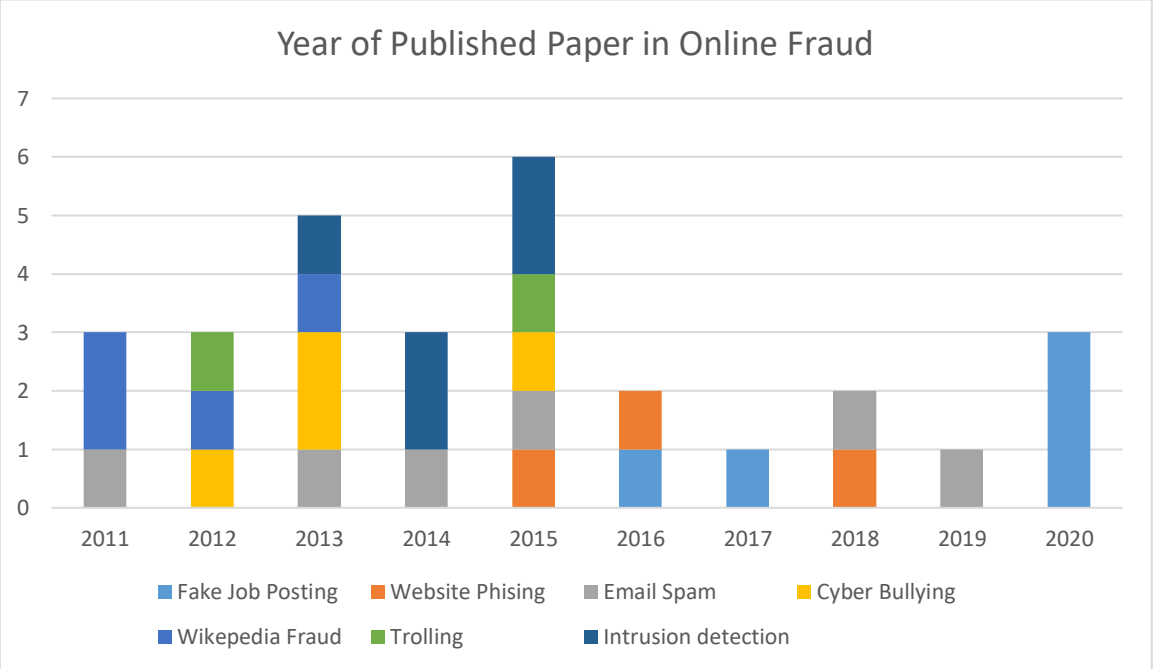


Figure 2. 2**:** Year of Published Paper in Online Fraud

During recent days we see ascend in fake employment posting where job posting appear quite sensitive, frequently these organisations will also have a designed site and they have enlistment procedure that is same very much in common like genuine organisation. In precise look on these posting it can be segregated from fake or genuine posting, more often its noted these fake posting doesn't have an organisation logo also the underlying reaction from the organization is from not having company domain email account or informal email account in many cases.

In various areas, there is a wealth of literature on cybercrime detection models. Data mining methods for various other identification purposes were often explored in related works. A few studies have looked at online recruiting scams. That being said, there are only two trials, one descriptive and the other experimental. One of them is an observational study that looked at online recruiting manipulation and scams.

## 2.4 Related Research Publications in Fake job Posting and Related Areas

### 2.4.1 Fake Job Posting Publications

(Dutta and Bandyopadhyay, 2020) A very recent paper of classifying job posts as fake, a variety of classifiers are used, including DecisionTree Classifier, AdaBoost Classifier, Naive Bayes Classifier, Multi-Layer Perceptron Classifier, Knearest Neighbour Classifier, Gradient Boost Classifier, and Random Tree Classifier. It should be noted that the dataset's attribute fraudulent' is used as the target class for classification. The classifiers are trained using 80% of the entire dataset at first, and then 20% of the entire dataset is used for prediction at the end. Here MPL classifier also been incorporated with 5 different set of hidden classifier, size sis as follows: 128, 64, 32, 16 and 8.

Another Comparative Study (Nasser, 2020) is present for fake job Posting where after cleaning and pre-processing data set, TF-IDF used to extract features. We learned and tested the classifiers after they were implemented. Precision, recall, f-measure, and accuracy are the evaluation measures used. The results of each classifier were summarised and compared to those of other machine learning classifiers.

Another recent well explained research on Online Recruitment Fraud Detection Model was put forward by (Alghamdi and Alharby, 2019) To differentiate scams or fraud objects from the data collection, the model core principle is to use the ensemble based classifier and SVM Algorithm for the feature selection. Another recent research on fake job recruitment detection using machine learning approach was put forward by (Dutta and Bandyopadhyay, 2020) where few classifiers are used, such as the , the Multi-Layer Perceptron Classifier, the K-nearest Neighbour Classifier, the AdaBoost Classifier, the Gradient Boost Classifier are used.

First base got proposed in the year 2017 by (Vidros et al., 2017) where several analysis using Bag of Words model, Empirical Analysis, Geography and dataset complete evaluation are shown. In regard to this significant and current extreme and time-critical subject, the project at hand describes and classifies the various aspects of this new and unusual research that has already been done. For the same amount of effort, it contributes and serves as a testable dataset of 17,880 jobs is being made available to the public for exploration, one, created from the collected usage data. Various model been used here like Random Forest, Naïve Bayes, ZeroR, OneR, Linear Regression and J48 and compared their accuracy.
(Vidros et al., 2016) The author discusses the seriousness of job abuse and the issues that have recently been addressed in the field of online recruiting fraud. (ORF) Several firms have recently agreed to advertise job openings online so that job hunters can locate them effectively and quickly. However, since they have work in exchange for money from job seekers, this may be a kind of bribery by fraudsters. It is possible to post false job advertisements against a well-known company in order to damage their image. The author requests that an automatic system be developed to recognise false positions and disclose them to people so that they can not apply for them. This bogus work after discovery generates a lot of attention.

### 2.4.2. Website Phising Publications.

(Jain and Gupta, 2018) where it does client-side identification of phishing websites using a machine learning method where Elevated accuracy of detection approach proposed by Misclassification of real websites as phishing (false positive) must be the minimum and accurate classification of phishing websites (true positive).
In above proposed paper machine learning approach is applied which employs several classification algorithms for recognizing fake posts. In this case idea to purpose Bidirectional LSTM (Ma and Hovy, 2016) deep learning model to predicate fake job posting.
(Jain and Gupta, 2016) This paper provides a thorough examination of phishing attack, their manipulation, and a comparison of some of the more recent machine learning-based methods for phishing identification. It offers a deeper understanding of the phishing challenge, the existing solution space in the machine learning domain, and the potential of future research to effectively deal with phishing attacks using machine learning-based approaches.

(Liang et al., 2015) Through analysing the impact of the feature selection method on classification results, researchers looked into the influence of feature selection on phish website identification. To remove a significant number of hybrid functions, an observational experiment was performed on a complex test-bed package. CBF, WFS, and IG were three feature selection algorithms (FSAs) function filtering. To qualify the way small subsets change detection precision, specificity, and sensitivity of the classification model to the best rates, a comparison of classification models was conducted. Due to their superior robustness, efficiency, and prediction, both feature selection methods outperformed their competitors significantly. The experiment's findings indicated a large increase in detection precision and observation accuracy in the susceptibility to stiffness, as well as predictability with high

throughput. The study's findings helped in the identification of the most successful subset of functionality for data collection and reliable phishing detection.

### 2.4.3. Intrusion Detection Publications.

(Gaikwad and Thool, 2015) In order to include an innovative intrusion mitigation technique, the bagging ensemble framework of machine learning was used. Selection of feature, building of the key classifier, packet sniffer, REPTree algorithm and detector were among the several modules involved to create duo instruments. They also suggested the bagging ensemble process algorithm as an intrusion detection scheme. To boost classification accuracy, a weak classifier was used. With a classification precision of 99.67% and the ensemble bagging machine-learning technique was the most accurate. They also found that the approach took less time to construct models and had less false positives(FP) than the AdaBoost algorithm with Decision stump. The study's findings show that the bagging category using REP Tree has the best classification accuracy. The bagging approach has the advantage of taking less time to build the model. In contrast to other machine learning methods, the proposed group approach produces fewer false positives.

(Sornsuwit and Jaiyen, 2015) using Ensemble Learning, they built an Intrusion Detection Model for U2N and R2R Attacks. Using a comparison determined algorithm, the ensemble learning was focused on detecting network intrusion data and reducing redundant functionality. By fixing the agreed problems in the KDD Cup'99 dataset for U2N and R2R attacks, it will increase the accuracy of the classifier. They used the Ada boost algorithm to create a strong classifier by combining weak classifiers in a linear fashion. To evaluate the required class of unseen data and calculate the posterior probability for each class, Nave Bayes was used. The mapping function from input space to undetectable level and from undetectable level to output space was performed using an MLP network. The classification problems were solved using an SVM method based on an ideal hyper plane in a high-dimensional domain. The findings of this study indicate that eliminating features enhances attack detection in works with a broad variety of low scales.

(Balakrishnan et al., 2014) Feature collection and classification techniques were used to research the intrusion detection system. The aim of the analysis was to develop and implement an intrusion detection method to combat potential attacks. The writers used a variety of strategies to get the data they requested from the dataset. Additionally, a rule-based classifier was used for evaluating the potential intrusions, as well as for binary and multiple classification for usage in a variety of complicated and simpler applications. Calculating the knowledge benefit ratio on attribute collection was used to apply a suggested algorithm for optimised function choice. The feature selection discovery algorithm selected only the most appropriate attributes, reducing the amount of time it took to find and identify the object. The use of a SVM and regex rule based classifier assisted in improving accuracy. The intrusion detection system that was provided decreased false positive rates and computing time. By determining the percentage of knowledge gain in characterization, this analysis led to the selection of the optimum benefit. It also sped up the process of locating

and identifying the item. Assist the dependent workbook's precision and the support conversion machine's reliability. The intrusion detection system minimised the number of false positives and cut down on estimation time.

(KumarShrivas and Kumar Dewangan, 2014) The use of BN - Bayesian Net - and ANN - Artificial Neural Network - was suggested as a solution. The approach uses a binary decision tree using Classification and Regression Tree, setting a record at each node with the Benefit Ratio (GR) preference selection process. The ANN was used to define data using a single attribute's property. Classification was developed in an ensemble fashion, with many frameworks working together to achieve classification accuracy. In order to even out the biases, a feature selection process was used. On the NSL-KDD and KDD99 datasets, different classification strategies were used to eliminate irrelevant features and increase classification accuracy. The suggested model had a 99.42 percent accuracy with the KDD99 dataset and a 98.07 percent accuracy with the NSL KDD dataset.

(A.M and S Irfan Ahmed, 2013) The SVM method was used to expand features in order to discover additional data while the RF model subspaces and the bootstrap artificial neural network (ANN) processed information. Because of their dependence on a common model, these blended models appear to offer an improved degree of precision over standard models. With a starting weight of zero and a maximum precision of one, each stage of the base algorithms is given a ranking from one to ten. According to the results, the ensemble approach is one of the most important advances in the area of machine learning.

### 2.4.4. Email Spam Publications.

(Sharaff et al., 2015) effort feature categorization analysis participants were able to analyse the effects of feature selection methodology on email classifications by considering their approach, focused on various groups of features The mathematical algorithm named the tree-based J48, which was proven to work well, was also, and the classifier known as the Bayes, and the SVM was evaluated as well. The features were chosen using a Chi-Square and IG. The SVM classification approach yielded the best overall performance without the use of feature selection techniques, based on the best overall results. Nave Bayes has little impact on feature selection techniques. In addition, J48 improved slightly with feature discovery, while info-Gain outperformed the Chi-square feature selection technique.

(Nizamani and Memon, 2014) introduced a methodology for detecting malicious emails development of appropriate feature selection. Because of its flexibility and inductive design, the J48 classification algorithms methodology was used. In addition, it was appropriate to use a cluster-based classification model to render the classifications as the data points were classified by obvious characteristics. Since the beginning of the dataset development, there have been 8000 emails. It's fine as long as long as you take the functions' frequency into account, regardless of the classification process, in order to ensure higher precision for email

identification. Also with a classification scheme in place, the classification techniques may be effective with the purpose of classifying and isolating emails whether they are frequency-based. As can be seen from the new studies, the degree of precision was observed to be influenced by classifiers of different forms rather than features of specificity. The email model evaluated in this study gained 96% of the information from the source material and features using the language to simulate it.

(Rathi and Pareek, 2013) an SVM was used to evaluate the results, with the classification as a primary objective in mind. With the help of the use of data mining to extract questionable emails, we went through the different data mining strategies to learn which classifier is most suitable for identifying spam. an NB classifier was used to determine if the role attribute relevance was unconnected to another role variable because the classes are different, which means that they can have similar distributions of attribute values, in models where attributes are not proportional to the roles. Further, the data was analysed using a functionality exclusion procedure, which first filtered out unnecessary and then included appropriate features. The analysis showed that classifier Random Tree has an accuracy of 99.15% when using the better feature selection algorithm, and an accuracy of 90.93% when using the strongest algorithm.

As fake job posting fraud resembles very much to most the online fraud hence Behaviour based email analysis with Spam Detection Application proposed by (Hershkop, 2006), professional classification algorithm of serious unethical activities, for example, inappropriate body dates format of the message or clear contradictions in users' past email behaviour.

Another proposed paper in email spam by (Blanzieri and Bryl, 2008) The solutions suggested vary from different protocols for sender authentication to qualified classifiers that differentiate between regular and junk emails.

(Dada et al., 2019) Author provide a comprehensive analysis of some of the popular machine learning based email spam filtering approaches. The examination of various ideas is carried out through a review of previous work, which tests their efficacy, and looks at the future of new developments. The study to be carried out includes an evaluation of existing products and future options. Although in the details of the machine learning topic are held to a mystery, the leading ISPs, namely Gmail, Yahoo, and Outlook spam filters present in the findings as evidence for its practical applications in filtering spam. As with other people working to develop email spam filtering, we explore ways to go about it, a variety of options, such as training spam-filter algorithms using machine learning. We examine the capabilities and challenges of current machine learning technologies, and compare them to the risks and liabilities of open machine learning approaches.

### 2.4.5. Cyber Bullying Publications.

(Al-Garadi et al., 2016) They applied machine learning methods to the collection of features, which were extracted from each author, the network, users, and tweet content, in order to generate a set of specific features which in turn were used to recognise cyberbullying on Twitter. Author here demonstrate that an established detection model which was derived from proposed features, performed well with an area under the receiver operating characteristic (ROC) of 0.943, as well as a successful with an area under the ROC curve of 0.36. It can be concluded from these findings that the proposed model is suitable for locating online cyberbullying from these features. In the end, the author compared the proposed features to results using previous results to see how they can apply to new features. Comparing these features with the expected outcomes shows the magnitude of the magnitude of the importance of the proposed features have.

(Dinakar et al., 2011) The overall identification issue is decomposed into sensitive subject detection, which lends itself to text classification sub-problems in the written article. The author used a corpus of 4500 YouTube comments to test a variety of binary and multiclass classifiers on a corpus of 4500 comments. Binary classifiers for independent brands often outperform multiclass classifiers, according to the report. Person topic-sensitive classifiers may be used to identify textual cyberbullying, according to the results.

(Chen et al., 2012) The Lexical Syntactic Function (LSF) architecture is introduced by the writers as a way to recognise offensive material and classify possible offensive users of social media. In addition, differentiate between pejoratives/profanities and obscenities when assessing inappropriate content, and use hand-authoring syntactic guidelines to classify name-calling harassments. They use a user's writing style, composition, and real cyberbullying content as features to predict whether or not the user would send out offensive information. Experiments revealed that the LSF platform worked considerably better than current approaches in detecting offensive material. In sentence offensive detection, it has a precision of 98.24 % and a recall of 94.34 %, and in user offensive detection, it has a precision of 77.9 % and a recall of 77.8 %. Meanwhile, LSF has a processing speed of around 10msec per sentence, indicating that it may be used effectively in social networking.

(Dadvar and De Jong, 2012) The authors suggest that incorporating the users' facts, attributes, and post-harassing behaviour, such as changing their status on another social network in response to their bullying experience, would increase the accuracy of cyberbullying identification. Cross-system analyses of users' attitudes, such as tracking their responses in multiple online contexts, may assist with this phase and include knowledge that may contribute to more effective cyberbullying identification.

### 2.4.6. Wikipedia Vandalism Publications.

(Wang and McKeown, 2010) In this article, the authors suggest a new Web-based shallow syntactic- semantic modelling approach for detecting vandalism that uses Web search results as a resource and trains topic-specific n-tag and syntactic n-gram language models. We used logistic boosting and logistic model trees classifiers to achieve strong F-measures,

surpassing the findings published by major Wikipedia vandalism detection systems, by integrating simple task-specific and lexical functionality.

(Chin et al., 2010) This paper creates mathematical language models by building term distributions from Wikipedia article revision histories. The fitness of a new edit as compared to language models developed from previous versions can well mean that an edit is a vandalism case, since vandalism often requires the use of unintended words to draw attention. In comparison, the paper uses an adaptive learning approach to address the issue of Wikipedia vandalism marking that is noisy and incomplete. The Wikipedia domain, with its revision histories, provides a novel framework for exploring the use of language models to define author intent. Models carry promise for vandalism, as the experimental findings presented in the paper reveal.

(Harpalani et al., 2011) The authors of this paper look at more linguistically motivated approaches to identifying graffiti. They claim that literary vandalism is a separate category in which a group of individuals acts in a similar linguistic manner. Experiments show that (1) mathematical simulations reveal special language types in vandalism, and (2) deep syntactic patterns based on probabilistic context free grammars (PCFG) distinguish vandalism better than shallow lexico- syntactic patterns based on n-grams.

(West et al., 2010) Authors use the spatio-temporal properties of revision metadata to detect vandalism on Wikipedia in this article. Rollback, an administrative method of reversion, allows for the marking of malicious edits, which can then be linked to non-offending edits in a variety of ways. None of these functions, crucially, necessitate a study of the article or correction text. Finally, a classifier is created that detects vandalism at a level equivalent to the natural-language efforts we plan to supplement (90.33 percent accuracy). The classifier is robust (it can handle 100+ edits every second) and has been used to find over 5,000 manually confirmed vandalism cases outside of our branded collection.

### 2.4.6. Trolling Publications.

(Cheng et al., 2015) Authors evaluate users that were banned from three broad online discussion forums to identify antisocial conduct. According to the writers, such users focus their attention on a limited number of threads, are more likely to post irrelevantly, and are more effective at eliciting feedback from other users. Even Author discovered that as these users progress from joining a group to being banned, they not only write worse than most users, but they also become less accepted by the community. Furthermore, they discover that when peer criticism is too negative, antisocial behaviour is amplified. Our research further shows distinct consumer classes with varying degrees of antisocial behaviour that can evolve over time. The authors use these insights to detect antisocial participants early on, which is a crucial challenge for group administrators.

(Vidros et al., 2016) A research on the detection of cybercrime in internet activities, especially cyberbullying on Twitter, was presented. The key goal was to create a range of

exclusive Twitter-derived features. They contained material from the network, operation, users, and tweets. Technology features were used to propose a model for detecting cyberbullying on Twitter. All the details about the outreach (acquisition) activities, number of those who were targeted, and amount of people being tracked, and the degree of security they had to enjoy while doing so doing so were through surveys. Users' interaction attributes were also used to determine a user's online networking activity. Personality, ethnicity, and age were among the features introduced. The following algorithms were used: Random forest, Nave Baye, KNN and Support vector machine. The f-measure was 93% in the random tree. According to the findings of this report, the proposed model contributes to the creation of a viable approach for identifying online bullying in online networking contexts.

As fake job posting fraud resembles very much to most the online fraud hence Behaviour based email analysis with Spam Detection Application proposed by (Hershkop, 2006), professional classification algorithm of serious unethical activities, for example, inappropriate body dates format of the message or clear contradictions in users' past email behaviour.
(Hussain et al., 2019) Author has presented People opinions about the goods they buy on internet communities. It could be useful to other buyers while they're deciding what to buy. In this sense, spammers may exploit feedback for financial benefit, necessitating the development of techniques to identify spam reviews. This can be achieved by removing attributes from the feedback and utilising Natural Language Analysis to do so (NLP). These features are then subjected to machine learning techniques. Machine learning algorithms that use a dictionary or corpus to eradicate spam feedback can be replaced by lexicon-based approaches.

## 2.5 Data Format

The data for this study comes from the University of the Aegean's Laboratory of Knowledge and Communication Systems Protection(Laboratory of Information and Communication Systems, University of the Aegean, Samos, 2021) , which assembled a dataset of 18K work descriptions. This dataset includes documents that have been manually annotated and categorised into two groups. The dataset includes 17,014 real work descriptions and 866 false job descriptions.

| Data Attribute | Data Type |
|---|---|
| Job ID | Integer |
| Title | Text |
| Location | Text |
| Department | Text |
| Salary Range | Integer |

| | |
|---|---|
| Company profile | Text |
| Description | Text |
| Requirement | Text |
| Benefits | Text |
| Telecommuting | Binary |
| Has Logo? | Binary |
| Has Questions? | Binary |
| Employment Type | Text (Categorical) |
| Required Experience | Text (Categorical) |
| Required Education | Text (Categorical) |
| Industry | Text |
| Function | Text |
| Fraudulent | Binary |

Table 2. 2**:** Data attribute with data type

Manually annotated EMSCAD records is divided into two groups. More precisely, 17,014 valid and 866 fake work ads released from 2012 to 2014 are included in the dataset.

### 2.6 Evaluation Matrix

After handling the data and scaling then we will utilize Bidirectional LSTM , Radom Forest, Decision Tree , KNN classifier model and compare these with Evolution matrix (M and M.N, 2015). Confusion matrix to show the result of different model, confusion matrix visualizes the performance of models.

| | **Expected Class** | | |
|---|---|---|---|
| **Actual Class** | | **Positive** | **Negative** |
| | **Positive** | True Positive | False Positive |
| | **Negative** | False Negative | True Negative |

Table 2. 3**:** Confusion matrix

TP (True Positive) was defined as the actual work posts that were properly classified as legitimate, FP(False Positive) was described as the actual job posts that were incorrectly categorised as false, FN(False Negative) is described as fake job posts that have been incorrectly identified as true, and TN(True Negative) was the bogus work posts that were correctly labelled as fake. In addition, the parameters used to assess the efficiency of the ML methods are: precision (Acc), recall (True Positive Rate or Sensitivity (TPR)), accuracy (Positive Predicted Value (PPV)) and F-measure (F1), Below are the equations

$$\textbf{Accuracy} = \frac{Number\ of\ correctly\ classified\ samples}{Number\ of\ total\ samples} = \frac{TP + TN}{TP + FP + FN + TN}$$

$$\textbf{TPR} = \frac{Number\ of\ samples\ that\ correctly\ classified\ as\ real\ posts}{Number\ of\ samples\ that\ classified\ as\ real\ posts} = \frac{TP}{TP + FN}$$

$$\textbf{PPV} = \frac{Number\ of\ real\ posts\ that\ correctly\ classified\ as\ real}{Number\ of\ samples\ that\ actually\ real\ posts} = \frac{TP}{TP + FP}$$

$$\textbf{F1} = 2 * \left(\frac{Precision * Recall}{Precision + Recall}\right)$$

Figure 2. 3**:** Accuracy, TPR, PPV and F1

Accuracy, this is a parameter which describes the proportion of real estimates about the total number of cases taken into account. The accuracy, however, might not be sufficient to calculate the efficiency of the model as it does not take incorrect predicted cases into account. If a fake post is treated as a true one, it creates a significant problem. Hence, it is necessary to consider false positive and false negative cases that compensate to misclassification. For measuring this compensation, precision and recall is quite necessary to be considered (M and M.N, 2015)

Precision, The ratio of accurate positive results to the expected number of positive results. Recall, denotes the number of positive findings that are right divided by the number of all related samples. F1-Score or F- measure is the parameter dealing with both recall and accuracy is determined as the harmonic average of accuracy and recall.

## 2.7 ROC & AUC

On the Y axis of ROC curves, there is a true positive trend and on the X axis, there is a false positive rate. The top left corner of the map is the "optimal" position, with a false positive rate of zero and a real positive rate of one.

This isn't really realistic, but a greater region under the curve (AUC) is usually safer. The "curvature" of ROC curves is also significant since the true positive rate should be maximised while the false positive rate should be minimised. ROC curves are widely used to investigate the efficiency of a classifier in classification problems. In order to extend the ROC curve and ROC area to a multi-label category, it is necessary to binarize the results. Each label may have its own ROC curve, but each variable of the label indicator matrix can be viewed as a binary predictor to establish a ROC curve (micro- averaging).

## 2.8 Precision and Recall

Precision - The proportion of successfully detected positive instances among all of the anticipated positive instances is calculated. As a result, it is advantageous when the costs of False Positives are substantial.

$$\text{Precision} = \frac{\text{True Positive}}{(\text{True Positive} + \text{False Positive})}$$

Figure 2. 4: Precision

Recall - The proportion of accurately detected positive cases among all of the actual positive instances is calculated. When the cost of False Negatives is large, it is critical to consider this option.

$$\text{Recall} = \frac{\text{True Positive}}{(\text{True Positive} + \text{False Negative})}$$

Figure 2. 5: Recall

### 2.9 F1 Score

With combined Precision and Recall, the Harmonic Mean is obtained, which provides a more accurate assessment of the instances that were erroneously categorised than the Accuracy Metric. The accuracy measure may be utilised when the class distribution is comparable; however, the F1-score metric should be used if the class distribution is uneven.

$$\text{F1-score} = \left(\frac{\text{Recall}^{-1} + \text{Precision}^{-1}}{2}\right)^{-1} = 2 * \frac{(\text{Precision} * \text{Recall})}{(\text{Precision} + \text{Recall})}$$

Figure 2. 6: F1 Score

### 2.10 Discussion

Following are some main findings from previous research on fake job postings and related topics: The functionality extraction for fake job posting will assist in the discovery of appropriate and meaningful features as well as the development of new features to boost performance.

For fake job forecasts, various algorithms performed best, and no single algorithm stuck out as the best performer. The best function collection was achieved by combining the best-performing algorithms using a hierarchical consensus vote or other fusion methods, which consistently outperformed all individual algorithms. It was deduced that, in addition to forcing individual algorithms to their limits, the best results can be obtained by integrating

many top-performing algorithms. The model's robustness and accuracy were found to be enhanced by assembling it.

It was found that the most effective models were developed based on a discriminative learning approach. Knowing how to find similarities between fake job related area like website phishing, Email Spam, wikipedia vandalism, trolling, Bullying and Intrusion detection helps in improving model in terms of performance and speed.

## 2.11 Summary

In this chapter, we looked at how the area of machine learning has advanced and developed over time in terms of fake job identification. We've gone through different algorithms and approaches that have enhanced online fraud detection over time. We've all seen the many forms of internet manipulation that have emerged in recent years. We've also briefly addressed articles on website phishing, email spam, wikipedia vandalism, trolling, abuse, and intrusion detection that have been reported in various internet fraud areas. Furthermore, we have gone through the various but significant difficulties that make it automatically detecting fake jobs to be a research and development field. Let's take a look at some of the various testing methodologies for identifying and classifying fake job in the next chapter and see if they can support job applicants and make a more efficient job recruiting process.

# CHAPTER 3: RESEARCH METHODOLOGY

## 3.1 Introduction

This chapter begins by describing the methodology used in this study. It explains the different phases of the research process. The procedure used in the analysis is then presented. The analysis technique talks into the particulars of the dataset and how it was selected. The data preprocessing methods used for features that are in a suitable format for the fake job prediction. The input data processing methods used for pre-processing was grouped into various groups.
SMOTE - class imbalance to refine the data because it's extremely imbalanced. We'll use a number of machine learning algorithms to see which one is better for identifying fake jobs. The metrics used to evaluate the model's efficiency. Finally, the proposed design is defined.

## 3.2 Research Process

The quest for current literature on the chosen subject of fake work prediction is the first step in the study process. It is then supplemented by an analysis of the current literature in order to achieve a better understanding of existing science, examine existing methodologies and architectures, learn about recent developments, and identify research holes on fake jobs. Predictions that illustrates the analysis method is seen in the following figure. The quest for current literature on the chosen subject of fake work prediction is the first step in the study process. It is then supplemented by an analysis of the current literature in order to achieve a better understanding of existing science, examine existing methodologies and architectures, learn about recent developments, and identify research holes on fake jobs predictions.

## 3.3 Research Approach

The research process and model used for the Fake Job Prediction are illustrated in the diagram below. The subsections and parts that follow offer a comprehensive overview of each phase in the research approach.

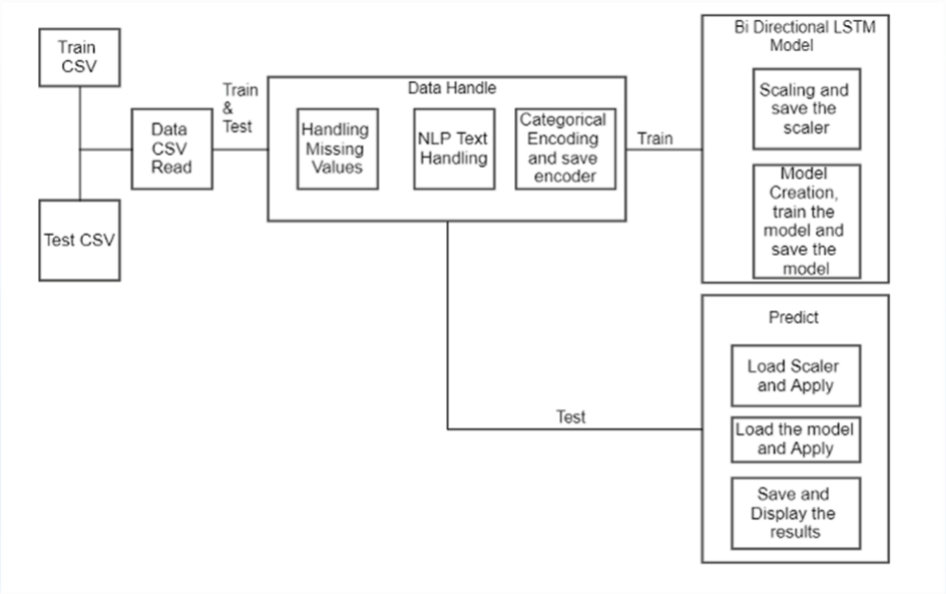



Figure 3. 1: Proposed research approach - Model: Bidirectional LSTM.

### 3.3.1. Data Selection

The data for this study comes from the University of the Aegean's Laboratory of Knowledge and Communication Systems Protection(Laboratory of Information and Communication Systems, University of the Aegean, Samos, 2021) , which assembled a dataset of 18K work descriptions. This dataset includes documents that have been manually annotated and categorised into two groups. The dataset includes 17,014 real work descriptions and 866 false job descriptions.

| Attribute | Data Type |
|---|---|
| Job ID | Integer |
| Title | Text |
| Location | Text |
| Department | Text |
| Salary Range | Integer |
| Company profile | Text |
| Description | Text |
| Requirement | Text |
| Benefits | Text |
| Telecommuting | Binary |
| Has Logo? | Binary |
| Has Questions? | Binary |

| | |
|---|---|
| Employment Type | Text (Categorical) |
| Required Experience | Text (Categorical) |
| Required Education | Text (Categorical) |
| Industry | Text |
| Function | Text |
| Fraudulent | Binary |

Table 3. 1: Attributes with Data type

Manually annotated EMSCAD records is divided into two groups. More precisely, 17,014 valid and 866 fake work ads released from 2012 to 2014 are included in the dataset.

### 3.3.2. Data Pre-processing and Transformations

Since we are also interested into text classification also and hence doing the analysis using EDA also for the feature ex: company which doesn't have logo, company which has less job description and requirement, company tend to mention salary in the subject line.

Following are data pre-processing data engineering:

- Replace null to string "missing" - instead of dropping missing, use as valid observation. It could mean that fake posts often have missing data.
- Separate country, state and city from location column.
- Drop non-English text entries.
- Clean text columns - separate sentences, remove URLs, non-ascii characters, punctuation, extra spaces and white space.
- Redefine education bins - some rows have "some high school coursework" or "high school or equivalent" etc. which are replaced with "less than high school" for generalising it.
- Drop salary column: it is very often missing and unsure what units are used in foreign countries, inconsistent time frame. There is no way to standardise this column for such wide range of values.
- Using the natural language toolkit (NLTK) (Manning et al., 2014) , the job posts were tokenized with the word tokenize() function, which generates a tokenized text document., one of the commonly used natural language processing (NLP) libraries.
- Drop token with non-alphabetical characters.
- Commonly used stop terms like a, an, etc. that are overlooked by any search query when indexing the entries. There is no need for these terms because they reserve memory and require precious time to process.

- In reducing terms such as 'plays' and 'played' to their common form such as 'play,' Stemming plays a crucial role, so we used the porterStemmer class from stem library to our dataset.

### 3.3.3 Word vector transformation

Proposed word vector used in this paper:
Count Vectorizer -Text is tokenized(tokenization is defined as breaking down a phrase or paragraph or any text into words) and very minimal preprocessing is performed, such as removing punctuation marks and changing all of the words to lowercase, is performed by CountVectorizer.

Tf-idf Vectorizer - This is very common algorithm to transform text into a meaningful representation of numbers which is used to fit machine algorithm for prediction.

Tf-idf GloVe embedding - GloVe (Global Vectors for Word Representation) embedding is a method for creating word embeddings that is different from the traditional technique. It is based on the use of matrix factorization techniques on the word-context matrix to get its results. The data is compiled into a massive matrix of co-occurrence information, and you count the number of times each "word" (the rows) appears in some "context" (the columns) in a vast corpus. Here we are using pretrained vector to convert word vector transformation

### 3.3.4. Class Balancing

SMOTE sampling on training data such that each class also has a number of observations. The 80/20 train/test split feature also does this. SMOTE: Synthesize new examples instead of an oversample for the minority class, which does not add any new details. SMOTE first generates a random example of a marginalised group and discovers its immediate neighbours in the marginalised group. The synthetic example is then generated by simply choosing one of the k nearest neighbours y and integrating x and y in the space of the function to form a line segment. The synthetic instances are generated as a conic integration of the two selected x and y instances.

## 3.4 Proposed Method

Using Bidirectional LSTM, Random Forest Classifier, and Logistic Regression Classifier, proposed an architecture based on previous work on Fake Job Prediction. The model's performance will be measured using the parameters Accuracy, Precision, TPR, specificity, sensitivity, ROC, and AUC.

### 3.4.1. Bi-directional LSTM

After lemmatization corpus using pad sequence need to create embedded docs for this corpus. Bidirectional LSTM model with embedded feature vector with sigmoid activation

function (Han and Moraga, 1995) with adam optimizer (Kingma and Ba, 2015). The adam optimiser chosen as it works well with noisy and sparse gradients.

Forget Gate: This gate controls how much information needs to be discarded from the previous cell state ($C_{t-1}$) depending upon the new input. In event recognition from text problem the gate needs to decide how much information to retain from the previous frame, if the same action is happening repeatedly then very less information should be discarded and if the action changes the forget gate forgets lots of information.

Update Gate: It makes an update to the previous cell state by writing a new piece of information to it. In the event recognition problem in text, when the action changes the gate will update the cell state with information relevant to the new action. In case when the action is the same as the previous text there will be negligible information will be written to the cell state. If the scene or action changes drastically the update will be drastic too.

Output Gate: The gate controls how much information needs to be passed on to the next LSTM layer based on the current state. The output gate contains a filter that makes sure only relevant information is passed to the next layer.

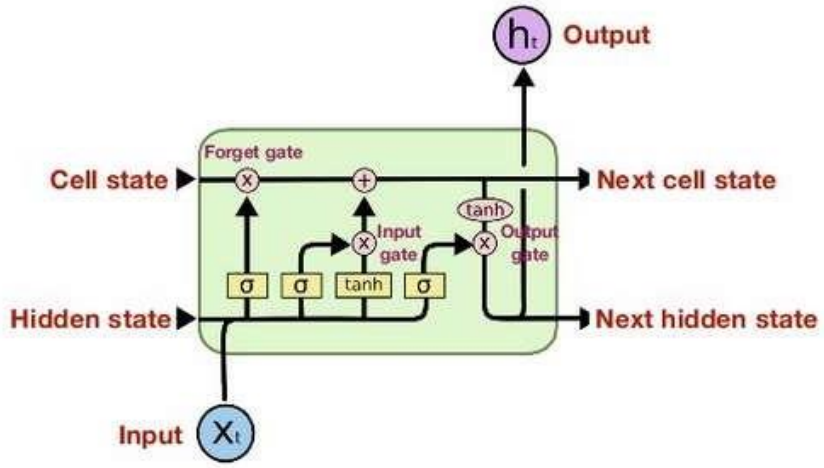



Figure 3. 2: Long short term memory network's cell structure.

$$f_t = \sigma(\boldsymbol{W}_f[h_{t-1}, x_t] + b_f) \quad (1)$$

$$i_t = \sigma(\boldsymbol{W}_i[h_{t-1}, x_t] + b_i) \quad (3)$$

$$C' = \tanh(Wc[h_{t-1}, x_t) + b_c) \quad (4)$$

$$C_t = f_t C_{t-1} + i_t C' \quad (5)$$

$$O_t = \sigma(\boldsymbol{W}_o[h_{t-1}, x_t] + b_o) \quad (6)$$

$$h_t = O_t \tanh(c_t) \quad (7)$$

Where: $\boldsymbol{W}_i, \boldsymbol{W}_f, \boldsymbol{W}_c, \boldsymbol{W}_o$ are the weight matrices. $h_{t-1}$ is the previous state output and $h_t$ is the current output state, $C_{t-1}$ is the previous cell state while $C_t$ is the current cell state.

$f_t, i_t, O_t$ are the forget gate, input gate, and output gate of the cell.

It is a variant of the vanilla LSTM model in which two LSTM layers are stacked face to face and in one layer the input flows from left to right and in another layer, the input flows from right to left in a backward sequence. As the event recognition from text is an offline task and we have the sequence of the frame beforehand we can use this for Bi-Directional LSTM. The advantage of using Bi-Directional LSTM in our model is that the model will be able to see the change of action beforehand and learn temporal features more efficiently which will increase the performance of the model.

### 3.4.2. Random Forest Classifier

After lemmatization corpus using pad sequence need to create embedded docs for this corpus It uses packing and feature randomness when constructing each individual tree to strive to build an uncorrelated forest of trees whose board prediction is more accurate than that of any individual tree.

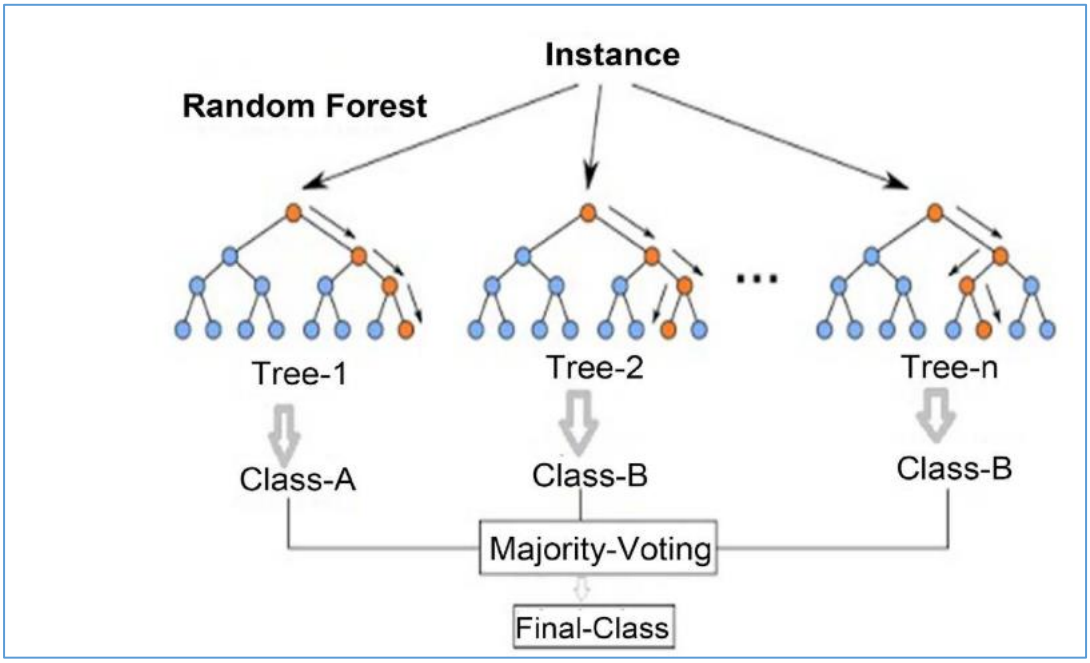



Figure 3. 3: Random Forest classifier.

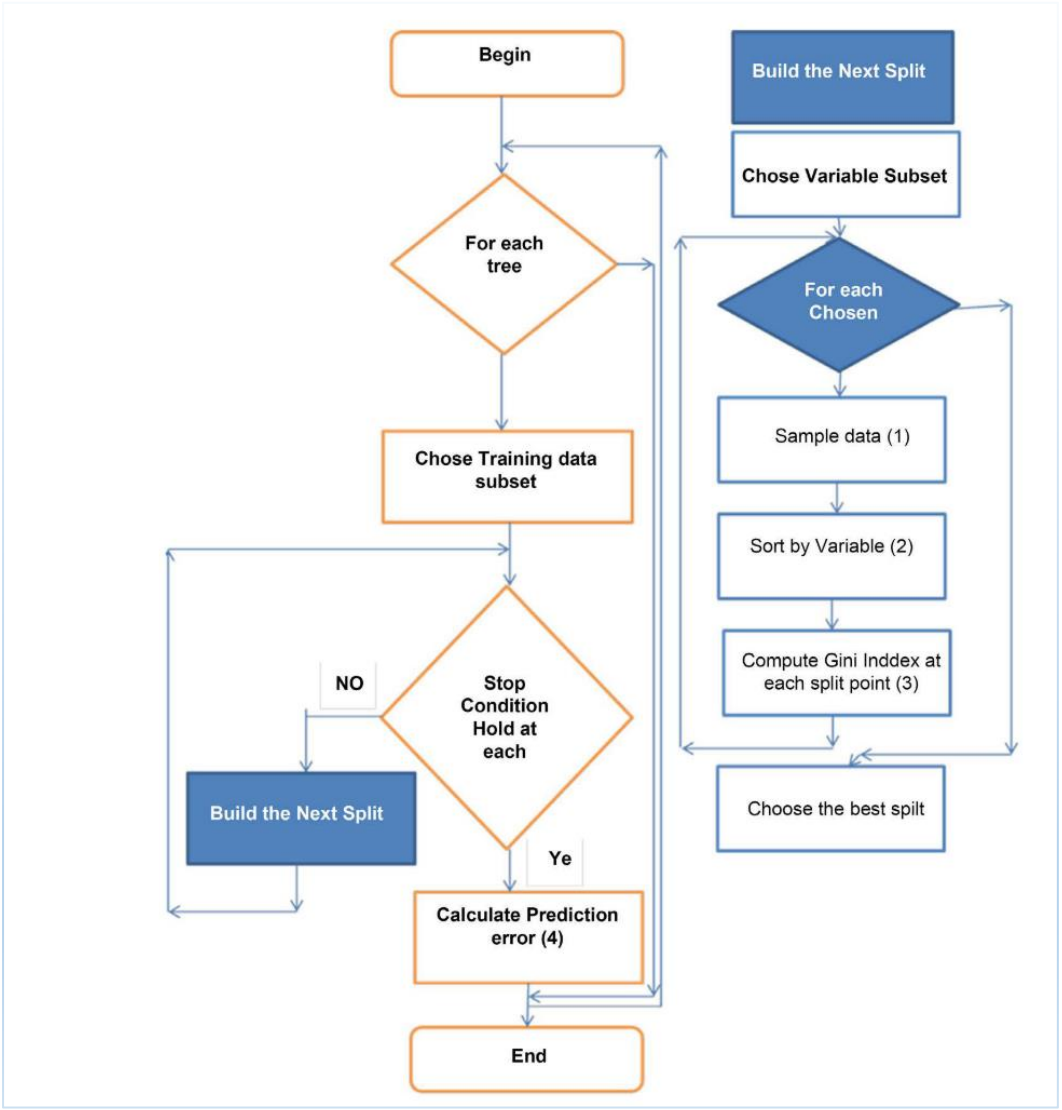



Figure 3. 4: Random forest classifier flow chart.

### 3.4.3. Logestic Regression

Supervised learning predict the probability of a target variable by using classification algorithm, After Lemma corpus on text and word vector transformation data fed into logistic regression model with balanced class weight data parameter. Basic Logistic Regression model explained in figure

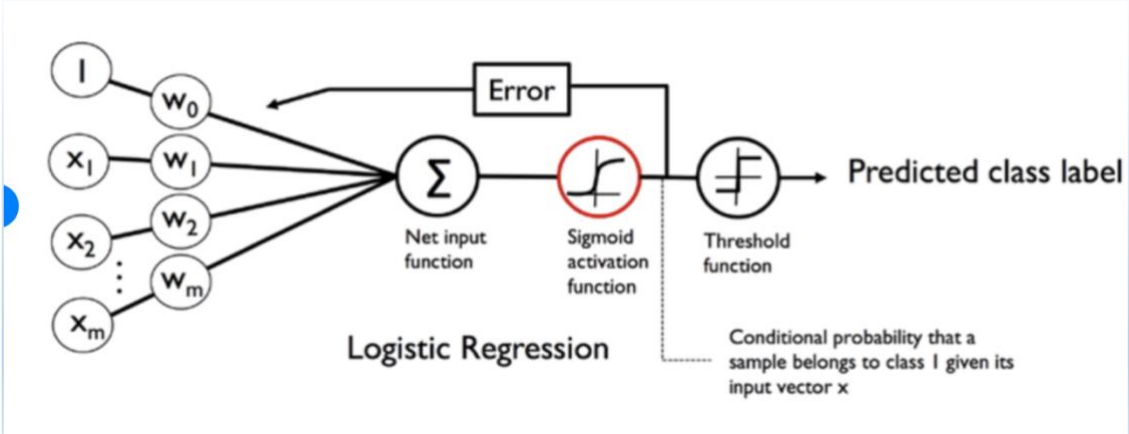



Figure 3. 5 Logistic Regression model Diagram

## 3.5 Summary

Various data preprocessing, data transformation, and other strategies for identifying fake jobs have been discussed in this portion. This chapter discusses the analysis methods and technique that can be utilised to establish a recruiting procedure that is free of significant fraud. We briefly discussed fraud forms linked to fake job postings in this segment, and we talked about classification methods that can be used to spot fake job postings, which can support job seekers. This chapter also contains technical information regarding the dataset that was included in the study. The numerous measurement measures have also been clarified to better clarify how the analysis methodology can be measured, as well as what these metrics mean and why they are relevant. We'll look through more actual implementation and review information in the following chapter.

# CHAPTER 4: IMPLEMENTATIONS

## 4.1 Introduction

In order to have a high-quality model which predicts the highest accuracy, it's important that proper cleaning of the text have been done. The data will be using for this analysis is a dataset of 18K job descriptions compiled by the University of the Aegean, Laboratory of Information & Communication Systems Security (http://emscad.samos.aegean.gr/). This dataset contains records which were manually annotated and classified into two categories. More specifically, the dataset contains 17,014 legitimate and 866 fake job descriptions.

The implementation of the different strategies and the values of the parameters we used in modeling to address the problem of fake job posting from the emscad dataset. First and foremost, we cleaned all entries and eliminated any unexpected non-English terms by recognizing non-ascii letter sequences(including html tags) in texts using regular expression pattern matching and filtering out any unexpected non-English terms. After that we removed English stop words like – the, is, are when, then etc. through stop words list {Formatting Citation}. Afterwards we checked class imbalance in the data also we analyzed whether Class imbalance really makes any difference in the performance or not, Following that, we employed the well-established TF-IDF modelling technique, and then we trained four distinct popular classifier models and evaluated their performance.

## 4.2 Exploratory Data Analysis

From the exploratory data analysis, we can evaluate important conclusion from different like Country, Department List, Employment types etc.

1. Countries with most applicant

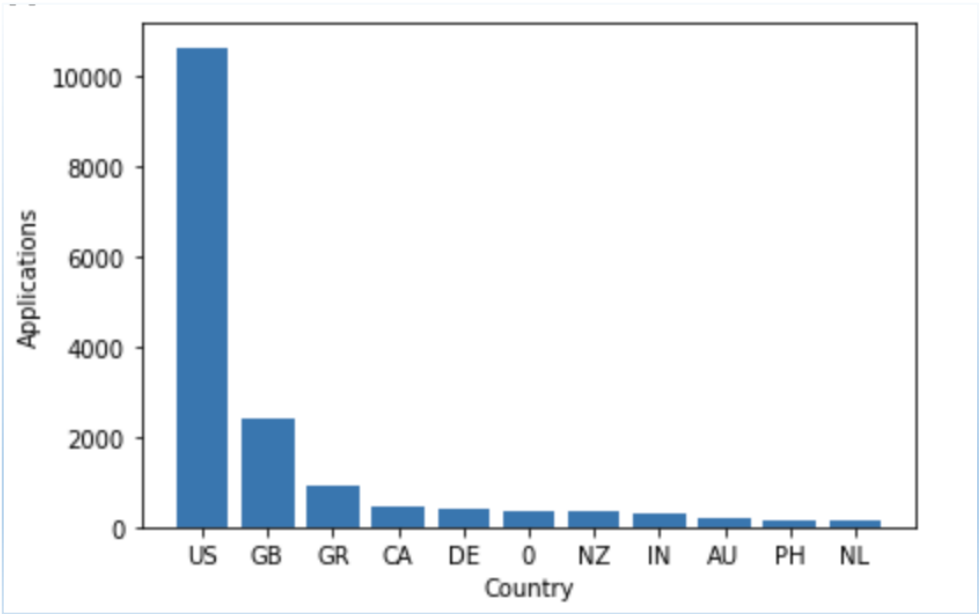



Figure 4. 1: Applicant with countries.

Most number of postings were from US it was acting as an outlier.

2. Top 20 Department.

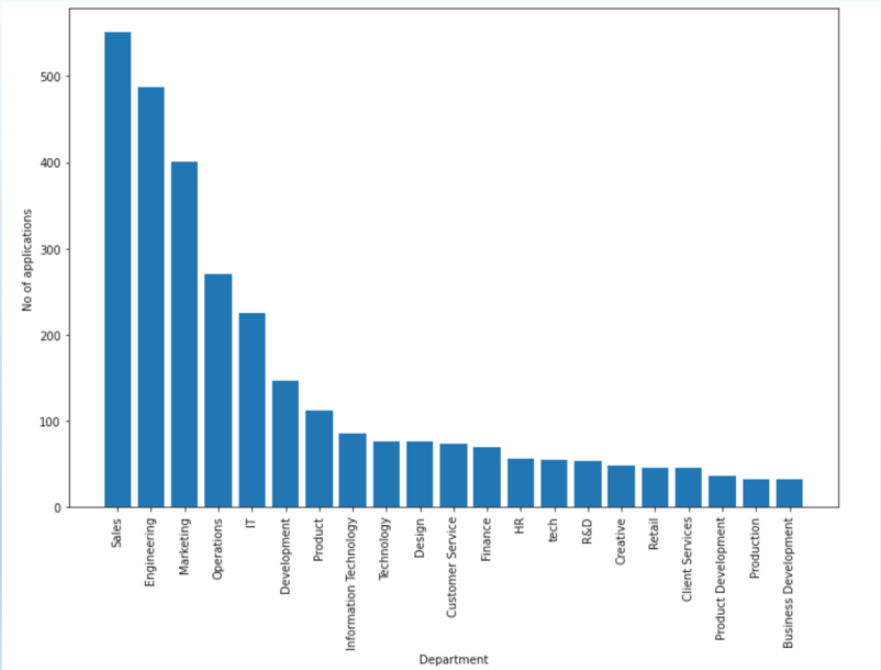



Figure 4. 2: Top 20 Departments.

Clearly sales Department is having top applicants.

3. Visualisations of categorical variables with real and fraudulent.

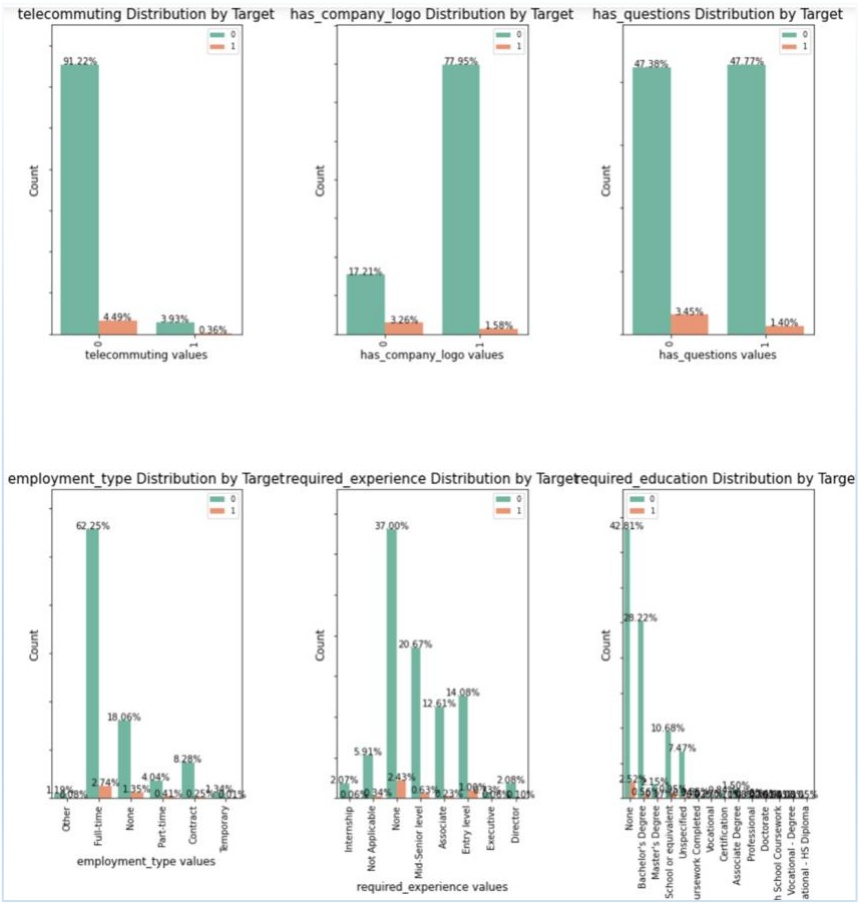



Figure 4. 3 : Categorical variables with real and fraudulent.

4. Employment types

a. Pie plot for employment types

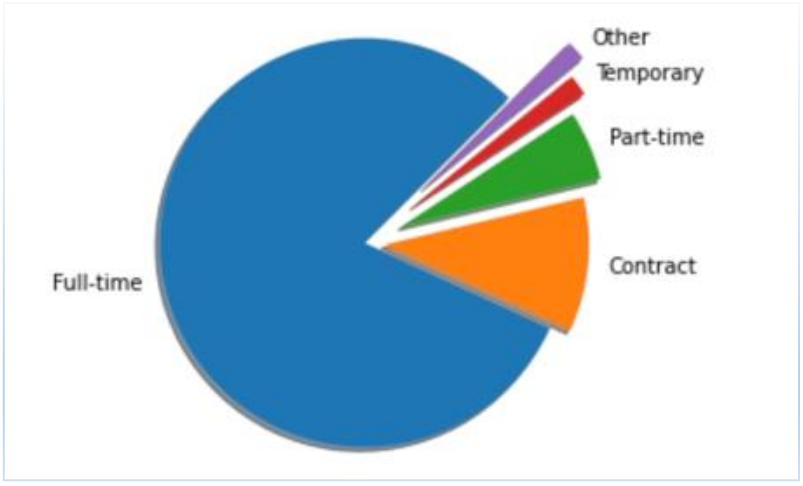



Figure 4. 4 Pie plot for employment type.

It is clear from pie plot full type has maximum applicants.

b. Counter plot for employment type

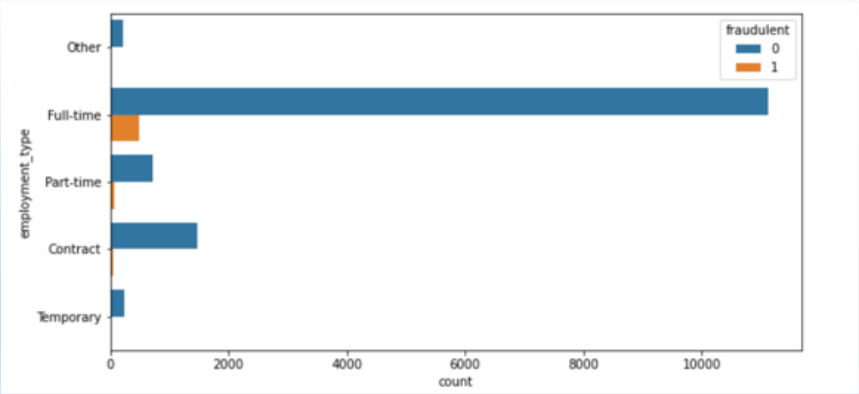



Figure 4. 5: Counter plot for employment.

c. Relationship between the target class and employment type.

| employment_type | Contract | Full-time | Other | Part-time | Temporary |
|---|---|---|---|---|---|
| fraudulent | | | | | |
| 0 | 1480 | 11130 | 212 | 723 | 239 |
| 1 | 44 | 490 | 15 | 74 | 2 |

Figure 4. 6: Relationship between the target class and employment type.

5. Required Education

a. Pie plot for Required Education

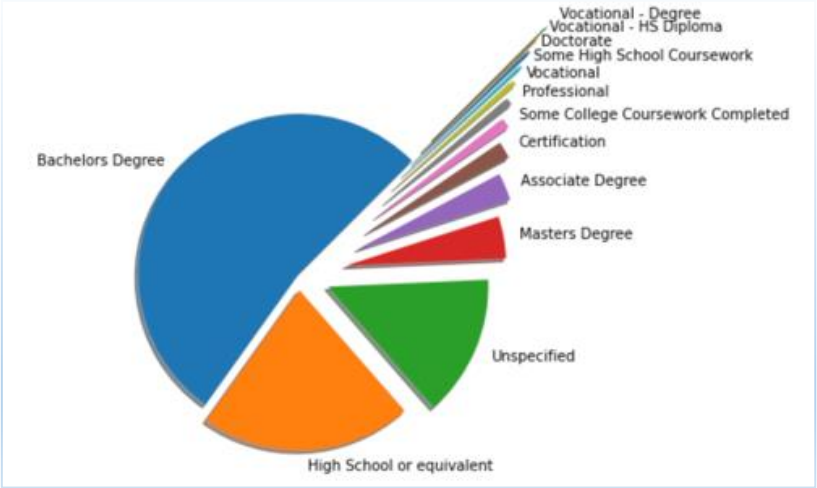



Figure 4. 7 : Pie plot for required education.

It is clear from pie plot bachelor has maximum applicants for required education.

b. Counter plot for required education

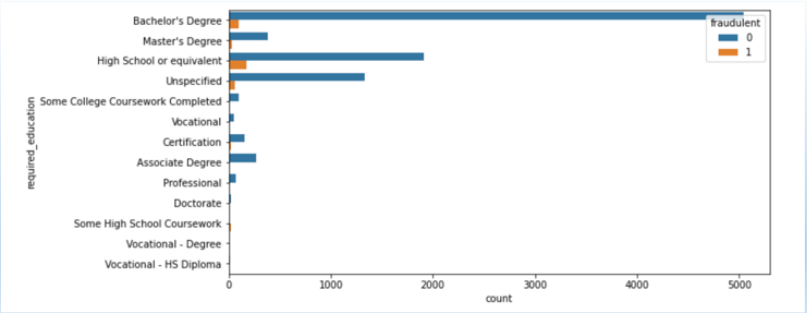



Figure 4. 8: Counter plot for required education.

c. Relationship between the target class and required education.

| required_education | Associate Degree | Bachelor's Degree | Certification | Doctorate | High School or equivalent | Master's Degree | Professional | Some College Coursework Completed | Some High School Coursework | Unspecified |
|---|---|---|---|---|---|---|---|---|---|---|
| fraudulent | | | | | | | | | | |
| 0 | 268 | 5045 | 151 | 25 | 1910 | 385 | 70 | 99 | 7 | 1336 |
| 1 | 6 | 100 | 19 | 1 | 170 | 31 | 4 | 3 | 20 | 61 |

Figure 4. 9: Relationship between the target class and required education.

6. Required Experience

a. Pie plot for Required Experience

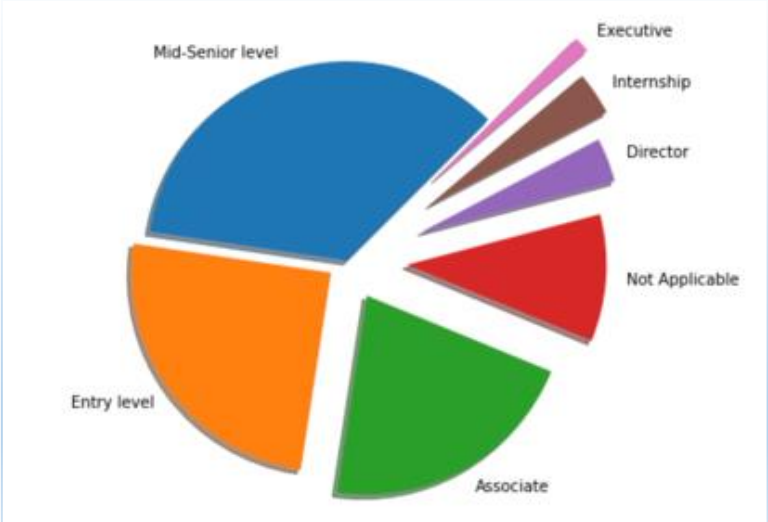



Figure 4. 10 : Pie plot for required experience.

It is clear from pie plot mid senior has maximum applicants for required experience.

b. Relationship between the target class and required experience.

| required_experience | Associate | Director | Entry level | Executive | Internship | Mid-Senior level | Not Applicable |
|---|---|---|---|---|---|---|---|
| fraudulent | | | | | | | |
| 0 | 2255 | 372 | 2518 | 131 | 371 | 3696 | 1056 |
| 1 | 42 | 17 | 179 | 10 | 10 | 113 | 60 |

Figure 4. 11: Relationship between the target class and required experience.

7. Company Logo.

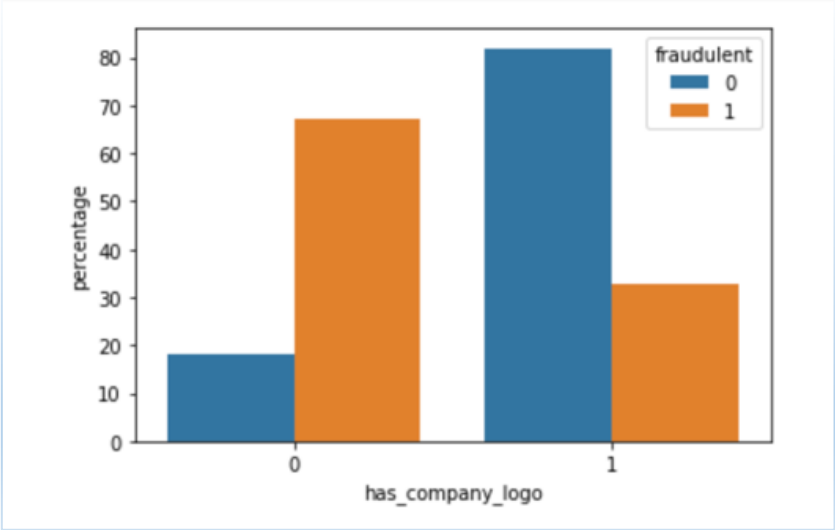



Figure 4. 12 : Company logo bar plots.

Most of the fraudulent job ads don't have a company logo.

8. Has Questions

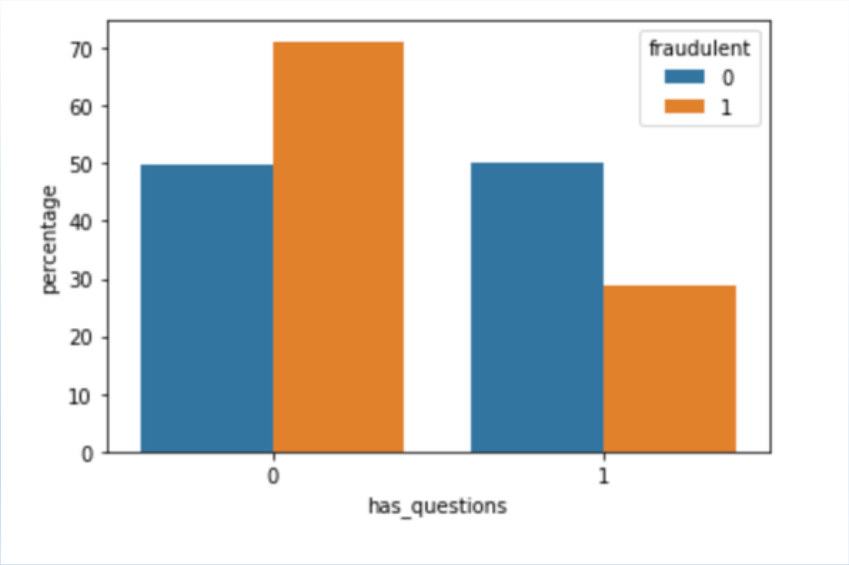


Figure 4. 13 : Has questions bar plots.

Most of the fraudulent job ads don't have a “has question” in form.

9. Text Length
    a. Histogram plot for text length

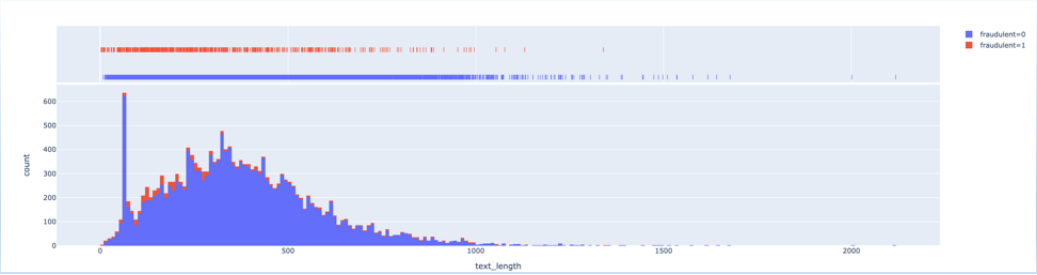



Figure 4. 14: Histogram plot for text length.

    b. Hist plot for text length

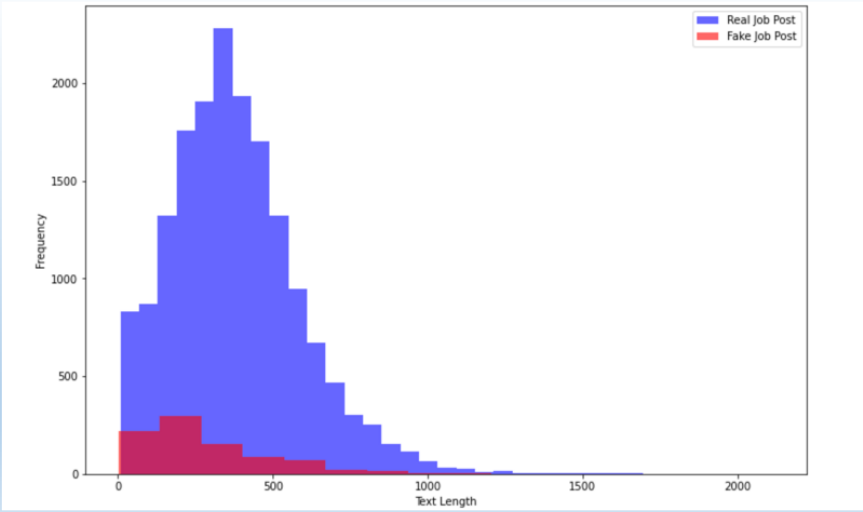



Figure 4. 15 Hist plot for text length.

It’s clear from histogram and hist plot that fraudulent job posting has less text length than real one.

10. Word cloud.

    a. WordCloud for Real job.

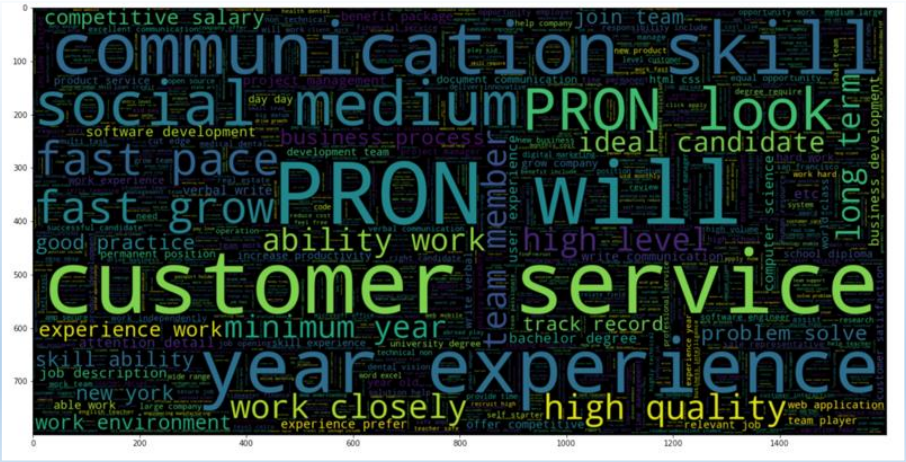



Figure 4. 16: WordCloud for real job.

b. WordCloud for fraudulent job.


Figure 4. 17: WordCloud for fraudulent job.

Customer Service, Communication skill & Year of skill are the top three most used phrases used in real job. Project, Customer Service, & Support are the top three most used phrases used in fraudulent job.

## 4.2 Data Preparation

Following are data pre-processing data engineering:

### 4.2.1 Data clean up

- Replace null to string "missing" - instead of dropping missing, use as valid observation. It could mean that fake posts often have missing data.
- Drop non-English text entries.

- For derived attribute ‘text’ perform regex operation to remove html tags, white space, non-ascii characters and URL.
- Drop salary column: it is very often missing and unsure what units are used in foreign countries, inconsistent time frame. There is no way to standardise this column for such wide range of values.
- Commonly used stop terms like a, an, etc. that are overlooked by any search query when indexing the entries. There is no need for these terms because they reserve memory and require precious time to process.
- In lemmatization algorithm it refer to dictionary to understand the meaning of the before reducing to its root words or lemma here we used wordnetlemmatizer library.

### 4.2.2 Dataset prepocessing

- Replaced nan with meaningful values in all categorical variables as these nan value may be having purpose ex : employment type, required experience and industry with nan data could be having meaning for fresher candidate job posting.
- Derived attribute ‘text’ formed by combining attributes like title, company profile, description, requirement, benefits.
- All the numerical data column scaled through standard scalar.
- All the categorical data are transformed through one hot encoder.

### 4.2.3 Word vector transformation

Different word vector transformer been used into this research and then evaluate the model based on their performance, word vector used in this paper and we discussed that in section 3.3.3:

- Count Vectorizer
- tf-idf Vectorizer
- tf-idf GloVe embedding

### 4.2.4 Dataset Split

This is the very last parameter. In order to get a similar distribution, we needed to stratify the dataset based on the class values. So we passed our class labels part of the data to this parameter and check what happens for EMSCAD dataset we split data into 80:20 train and test ratio.

## 4.3 Algorithm and Model Configuration

### 4.3.1 Word vector transformation

Conducted preliminary analysis word vector transformation modeling of the job description, benefits, requirements, and company profile HTML fields. Before feeding our data to three different logistic classifiers, named as Logistic regression with count vectorizer, Logistic regression with tfidf vectorizer and Logistic regression with tf-idf glove embedding.

Count Vectorizer:
As discussed on basic idea in section 3.3.3, corpus is being converted into count vectorizer with n gram range = (1,2).

```
bow=CountVectorizer(ngram_range=(1,2))
text_train=bow.fit_transform(x_train['text'])
text_test=bow.transform(x_test['text'])
```

tfidf Vectorizer:
As discussed on basic idea in section 3.3.3, corpus is being converted into tfidf vectorizer with n gram range = (1,2).

```
tfidf=TfidfVectorizer(ngram_range=(1,2))
text_train=tfidf.fit_transform(x_train['text'])
text_test=tfidf.transform(x_test['text'])
```

Glove Embedding:
As discussed on basic idea in section 3.3.3, here we use pertained glove vector, below are snippet for tf-idf glove embedding used in logistic regression model.

```
def glove_embedding(x):
    embed=[]
    tfidf_dict=tfidf.vocabulary_
    x=list(x.split())
    for word in x:
        try:
            embed.append(embeddings_index[word]*tfidf_dict[word])
        except:
            continue
    return np.mean(embed,axis=0)

text_train=x_train['text'].apply(glove_embedding)
text_test=x_test['text'].apply(glove_embedding)
```

### 4.3.2 Logistic Regression with word vec

A detailed explanation is given into figure 4.20

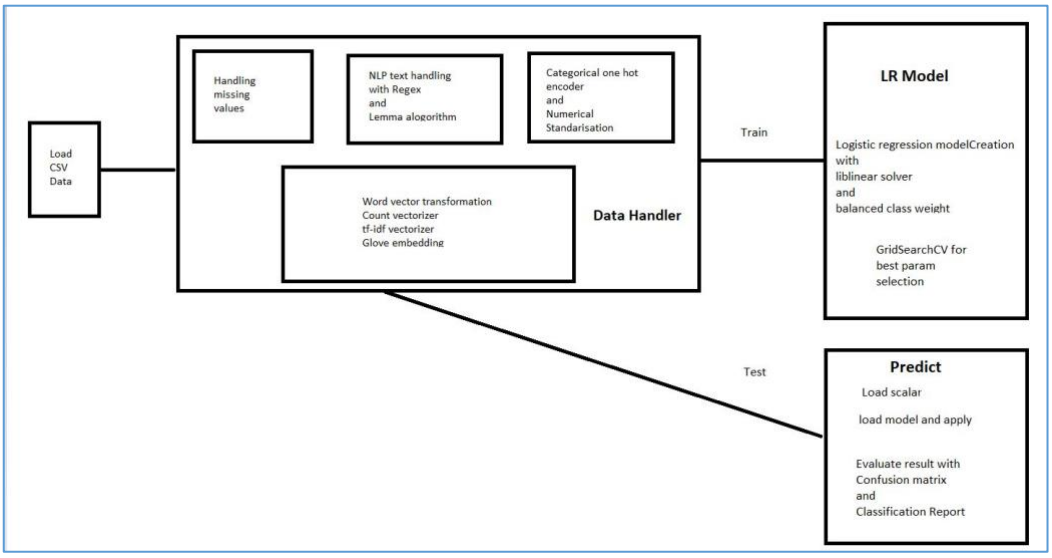



Figure 4. 18 : Logistic regression architecture

Logistic regression model experiment carried out with count vectorizer, tf-idf vectorizer and glove embedding, word vector transformation. Models build on balanced class weight and liblinear solver then with gridsearchcv and cross validation best param gets selected. For each run, the corpus was split into test and training through cross-validation subsets using gridsearchcv.

### 4.3.3 Random Forest

With tf-idf weighted Glove embedding test and train data fed into Random forest classifier model with number of estimator = 500. Detailed performance and evolution of matrix discussed in chapter 5.

```
clf=RandomForestClassifier(n_estimators=500,oob_score=True,n_jobs=-
1,random_state=0)
clf.fit(text_train,y_train)
```

### 4.3.2 Bidirectional LSTM layer

To process the sequence generated by the tokenizer, the number of neurons set in this LSTM layer is 32 neurons and then its processed into relu and sigmoid activation function with loss being selected ad binary entropy and adam optimizer, total number param generated from model is 6,426,417.

```
Model: "sequential"
_________________________________________________________________
Layer (type)                 Output Shape              Param #
=================================================================
embedding (Embedding)        (None, 250, 64)           6400000
_________________________________________________________________
spatial_dropout1d (SpatialDr (None, 250, 64)           0
_________________________________________________________________
bidirectional (Bidirectional (None, 64)                24832
_________________________________________________________________
dense (Dense)                (None, 24)                1560
_________________________________________________________________
dense_1 (Dense)              (None, 1)                 25
=================================================================
Total params: 6,426,417
Trainable params: 6,426,417
Non-trainable params: 0
_________________________________________________________________
```

Figure 4. 19 : Total number params in Bi-directional LSTM model

# CHAPTER 5 : RESULTS AND ANALYSIS

## 5.1 Results

Below are confusion matrices, evaluation metrics and accuracy results for test data.

### 5.1.1 Confusion Matrices

| Logistic Regression with count vectorises (word vector transformation) | | |
|---|---|---|
| | Fake (Predicated) | Real (Predicated) |
| Fake (Actual) | 3387 | 5 |
| Real (Actual) | 26 | 147 |
| Logistic Regression with tfidf (word vector transformation) | | |
| | Fake (Predicated) | Real (Predicated) |
| Fake (Actual) | 3398 | 5 |
| Real (Actual) | 17 | 156 |
| Logistic Regression with tf-idf glove embedding (word vector transformation) | | |
| | Fake (Predicated) | Real (Predicated) |
| Fake (Actual) | 2952 | 451 |
| Real (Actual) | 32 | 141 |
| Random Forest | | |
| | Fake (Predicated) | Real (Predicated) |
| Fake (Actual) | 3403 | 0 |
| Real (Actual) | 108 | 65 |
| Bi-Directional LSTM | | |
| | Fake (Predicated) | Real (Predicated) |
| Fake (Actual) | 3401 | 13 |
| Real (Actual) | 64 | 98 |

Table 5. 1: Confusion Matrices

### 5.1.2 Evaluation Matrices

**Logistic Regression with Count Vectorise (word vector transformation)**

| | precision | recall | f1-score | support |
|---|---|---|---|---|
| 0 | 0.99 | 1.00 | 0.99 | 3403 |
| 1 | 0.90 | 0.85 | 0.87 | 173 |
| accuracy | | | 0.99 | 3576 |
| macro avg | 0.95 | 0.92 | 0.93 | 3576 |
| weighted avg | 0.99 | 0.99 | 0.99 | 3576 |

Figure 5. 1: Evaluation Matrices for Logistic Regression(Count Vectorise)

**Logistic Regression with tfidf (word vector transformation)**

| | precision | recall | f1-score | support |
|---|---|---|---|---|
| 0 | 1.00 | 1.00 | 1.00 | 3403 |
| 1 | 0.97 | 0.90 | 0.93 | 173 |
| accuracy | | | 0.99 | 3576 |
| macro avg | 0.98 | 0.95 | 0.97 | 3576 |
| weighted avg | 0.99 | 0.99 | 0.99 | 3576 |

Figure 5. 2: Evaluation Matrices for Logistic Regression(tfidf)

**Logistic Regression with tf-idf glove embedding (word vector transformation)**

| | precision | recall | f1-score | support |
|---|---|---|---|---|
| 0 | 0.99 | 0.87 | 0.92 | 3403 |
| 1 | 0.24 | 0.82 | 0.37 | 173 |
| accuracy | | | 0.86 | 3576 |
| macro avg | 0.61 | 0.84 | 0.65 | 3576 |
| weighted avg | 0.95 | 0.86 | 0.90 | 3576 |

Figure 5. 3 Evaluation Matrices for Logistic Regression(tf-idfglove embedding)

**Random Forest**

| | precision | recall | f1-score | support |
|---|---|---|---|---|
| 0 | 0.97 | 1.00 | 0.98 | 3403 |
| 1 | 1.00 | 0.38 | 0.55 | 173 |
| accuracy | | | 0.97 | 3576 |
| macro avg | 0.98 | 0.69 | 0.77 | 3576 |
| weighted avg | 0.97 | 0.97 | 0.96 | 3576 |

Figure 5. 4Evaluation Matrices for Random Forest

**Bi-Directional LSTM**

| | precision | recall | f1-score | support |
|---|---|---|---|---|
| 0 | 0.98 | 1.00 | 0.99 | 3414 |
| 1 | 0.88 | 0.60 | 0.72 | 162 |
| accuracy | | | 0.98 | 3576 |
| macro avg | 0.93 | 0.80 | 0.85 | 3576 |
| weighted avg | 0.98 | 0.98 | 0.98 | 3576 |

Figure 5. 5: Evaluation Matrices for Bi-Directional LSTM

### 5.1.3 Accuracy

| | Accuracy |
|---|---|
| LR (Count Vectorise) | 98.88 |
| LR (tf-idf) | 98.78 |
| LR (tf-idf glove embedding) | 86.49 |
| Random Forest | 96.54 |
| Bi-Directional LSTM | 97.74 |

Table 5. 2: Accuracy for different models

## 5.2 Analysis

After comparing all the above model results discussed in 5.1 section, it turned out LR (tf-idf) followed by LR (Count Vectorise) and LR (tf-idf glove embedding) has highest recall. Its important to notice that all the algorithm from Logistic regression has better performance, this is because word vector been used in logistic regression. Also we have noticed that LR(tf-idf) followed by LR(Count Vectorise) best f-measure and accuracy hence we conclude that most effective machine learning dealing with classification problem is LR(tf-idf). Moreover, we can notice that it has least predicated incorrect sample (fp+fn = 22). Despite this, we attempted to reduce the quantity of the vocabulary, however this resulted in poor performance in other

machine learning approaches. Furthermore, we believe that the big dataset size made the Bi-directional LSTM the most efficient approach, which is consistent with prior study findings that demonstrated that Bi-directional LSTM achieves good accuracy when dealing with many occurrences.

# CHAPTER 6: CONCLUSIONS AND FUTURE WORK

## 6.1 Conclusions

So, finally we can conclude, according to our text classification problem, and taking into consideration the text pre-processing we did, and feature extraction we have applied, the Logistic Regression with tf-idf word vector transformation is the best Machine Learning algorithm to solve the problem at hand.
EMSCAD is a publicly available dataset that contains both real valid and fraudulent job advertisements. In this paper, we investigated the possible aspects of employment scam, an unexplored research field that warrants further investigation. We also introduced EMSCAD, an unexplored research field that warrants further investigation. As seen, online recruitment fraud is a relatively young sector with varying degrees of severity that may swiftly evolve into a widespread fraud. According to our findings, employment scams are similar to well-studied problems such as email spamming, phishing, cyber bullying, vandalism on Wikipedia, opinion fraud and trolling. But they are distinguished by several peculiarities that make reliable scam detection through known methodologies difficult, necessitating the use of composite approaches when developing countermeasures. The EMSCAD dataset was also used as a basis for our experiments. Although the findings are early, they are thorough enough to demonstrate that text mining combined with metadata may serve as a rudimentary basis for job fraud detection systems. This dataset, we think, may be utilised as a component of an automated anti-scam solution by an application tracking system to train classifiers or get a more in-depth understanding of the features of the issue. It is also expected to stimulate and drive more research efforts in this very exciting, although still in its infancy, field.

## 6.2 Contribution to knowledge

From this experiment that we come with several word vec. transformation into model which help model in better accuracy and other evaluation matrices. Our works contributes building better ATS through which community can segregate the fake job posting, during this pandemic where the maximum job loss happen, and many people needs to go through job hunting and then with the help of these model we can lead scammer to get identified. Also, the key contribution of this study is that it is only the fourth experimental study in this area.

## 6.3 Conflict of interest

There is no conflict of interest while publishing this paper.

## 6.4 Future work

The ruleset will be further developed and enhanced in future works, with a particular emphasis on user behaviour, corporate and network data, as well as user-content-ip tracking patterns based on akamai routing logic, among other things. Furthermore, we would want to apply graph modelling to investigate the relationships that exist between fake job advertisements, firms,

and users. Ultimately, we want to present a useful employment fraud detection solution that may be used for commercial reasons and helpful for job seekers.

### 6.5 Limitations

With time and resource constraints, not able to try techniques like BERT (Bidirectional Encoder Representations from Transformers) which helps in text sequence either from left to right or combined left-to-right and right-to-left training which can results better classification. Also, these models only applied over English corpus.

## REFERENCES:


A.M, R. and S Irfan Ahmed, M., (2013) An Ensemble Classification Approach for Intrusion Detection. *International Journal of Computer Applications*, 802, pp.37–42.

Al-Garadi, M.A., Varathan, K.D. and Ravana, S.D., (2016) Cybercrime detection in online communications: The experimental case of cyberbullying detection in the Twitter network. *Computers in Human Behavior*, [online] 63, pp.433–443. Available at: http://dx.doi.org/10.1016/j.chb.2016.05.051.

Alghamdi, B. and Alharby, F., (2019) An Intelligent Model for Online Recruitment Fraud Detection. *Journal of Information Security*, 1003, pp.155–176.

Anon (2021) *IP Info, IP Geolocation Tools and API| IPInfoDB*. [online] Available at: https://ipinfodb.com/ [Accessed 21 Mar. 2021].

Anon (2021) *NSL-KDD | Datasets | Research | Canadian Institute for Cybersecurity | UNB*. [online] Available at: https://www.unb.ca/cic/datasets/nsl.html [Accessed 22 Mar. 2021].

Anon (2021) *PhishTank | Join the fight against phishing*. [online] Available at: https://www.phishtank.com/ [Accessed 20 Mar. 2021].

Balakrishnan, S., K, V. and A, K., (2014) Intrusion Detection System Using Feature Selection and Classification Technique. *International Journal of Computer Science and Application*, 34, p.145.

Binyousef, R.F., Al-Gahmi, A.M., Khan, Z.R. and Rawah, E., (2017) A rare case of Erdheim-Chester disease in the breast. *Annals of Saudi Medicine*, 371, pp.79–83.

Blanzieri, E. and Bryl, A., (2008) A survey of learning-based techniques of email spam filtering. *Artificial Intelligence Review*, 291, pp.63–92.

Chen, Y., Zhou, Y., Zhu, S. and Xu, H., (2012) Detecting offensive language in social media to protect adolescent online safety. *Proceedings - 2012 ASE/IEEE International Conference on Privacy, Security, Risk and Trust and 2012 ASE/IEEE International Conference on Social Computing, SocialCom/PASSAT 2012*, pp.71–80.

Cheng, J., Danescu-Niculescu-Mizil, C. and Leskovec, J., (2015) *Antisocial Behavior in Online Discussion Communities*. [online] *Proceedings of the International AAAI Conference on Web and Social Media*, Available at: www.aaai.org [Accessed 21 Mar. 2021].

Chin, S.C., Street, W.N., Srinivasan, P. and Eichmann, D., (2010) Detecting wikipedia vandalism with active learning and statistical language models. In: *Proceedings of the 4th Workshop on Information Credibility, WICOW '10*. [online] New York, New York, USA: ACM Press, pp.3–10. Available at:

http://portal.acm.org/citation.cfm?doid=1772938.1772942 [Accessed 20 Mar. 2021].
Cunningham, P. and Delany, S.J., (2020) k-Nearest neighbour classifiers 2nd edition (with python examples). *arXiv*, 1, pp.1–22.
Dada, E.G., Bassi, J.S., Chiroma, H., Abdulhamid, S.M., Adetunmbi, A.O. and Ajibuwa, O.E., (2019) Machine learning for email spam filtering: review, approaches and open research problems. *Heliyon*, 56.
Dadvar, M. and De Jong, F., (2012) Cyberbullying detection: A step toward a safer internet yard. In: *WWW'12 - Proceedings of the 21st Annual Conference on World Wide Web Companion*. [online] New York, New York, USA: ACM Press, pp.121–125. Available at: http://dl.acm.org/citation.cfm?doid=2187980.2187995 [Accessed 20 Mar. 2021].
Dinakar, K., Reichart, R. and Lieberman, H., (2011) *Modeling the Detection of Textual Cyberbullying*. [online] *Proceedings of the International AAAI Conference on Web and Social Media*, Available at: www.aaai.org [Accessed 20 Mar. 2021].
Dutta, S. and Bandyopadhyay, S.K., (2020) Fake job recruitment detection using machine learning approach. *SSRG International Journal of Engineering Trends and Technology*, 684, pp.48–53.
Gaikwad, D.P. and Thool, R.C., (2015) Intrusion detection system using bagging ensemble method of machine learning. *Proceedings - 1st International Conference on Computing, Communication, Control and Automation, ICCUBEA 2015*, pp.291–295.
Group, S.E., (2007) Методи за автоматично управление на подемни устройства при Jack-up системите.
Han, J. and Moraga, C., (1995) The influence of the sigmoid function parameters on the speed of backpropagation learning. *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*, 930, pp.195–201.
Harpalani, M., Hart, M., Singh, S., Johnson, R. and Choi, Y., (2011) Language of vandalism: Improving Wikipedia vandalism detection via stylometric analysis. *ACL-HLT 2011 - Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies*, 22009, pp.83–88.
Hershkop, S., (2006) Behavior-based email analysis with application to spam detection. *Architecture*.
Hussain, N., Mirza, H.T., Rasool, G., Hussain, I. and Kaleem, M., (2019) Spam review detection techniques: A systematic literature review. *Applied Sciences (Switzerland)*, 95.

Jain, A.K. and Gupta, B.B., (2016) Comparative analysis of features based machine learning approaches for phishing detection. *Proceedings of the 10th INDIACom; 2016 3rd International Conference on Computing for Sustainable Global Development, INDIACom 2016*, pp.2125–2130.

Jain, A.K. and Gupta, B.B., (2018) Towards detection of phishing websites on client-side using machine learning based approach. *Telecommunication Systems*, [online] 684, pp.687–700. Available at: https://doi.org/10.1007/s11235-017-0414-0.

Kingma, D.P. and Ba, J.L., (2015) Adam: A method for stochastic optimization. *3rd International Conference on Learning Representations, ICLR 2015 - Conference Track Proceedings*, pp.1–15.

KumarShrivas, A. and Kumar Dewangan, A., (2014) An Ensemble Model for Classification of Attacks with Feature Selection based on KDD99 and NSL-KDD Data Set. *International Journal of Computer Applications*, 9915, pp.8–13.

Laboratory of Information and Communication Systems, University of the Aegean, Samos, G., (2021) *Employment Scam Aegean Dataset*. [online] Available at: http://emscad.samos.aegean.gr/ [Accessed 16 Jan. 2021].

Liang, D., Tsai, C.F. and Wu, H.T., (2015) The effect of feature selection on financial distress prediction. *Knowledge-Based Systems*, 731, pp.289–297.

M, H. and M.N, S., (2015) A Review on Evaluation Metrics for Data Classification Evaluations. *International Journal of Data Mining & Knowledge Management Process*, 52, pp.01–11.

Ma, X. and Hovy, E., (2016) End-to-end sequence labeling via bi-directional LSTM-CNNs-CRF. *54th Annual Meeting of the Association for Computational Linguistics, ACL 2016 - Long Papers*, 2, pp.1064–1074.

Manning, C.D., Bauer, J., Finkel, J. and Bethard, S.J., (2014) The Stanford CoreNLP Natural Language Processing Toolkit. *Aclweb.Org*, [online] pp.55–60. Available at: http://macopolo.cn/mkpl/products.asp.

Nasser, I.M.., (2020) Machine Learning and Job Posting Classification: A Comparative Study. [online] 49, pp.6–14. Available at: https://www.academia.edu/44270115/Machine_Learning_and_Job_Posting_Classification_A_Comparative_Study.

Nizamani, S. and Memon, N., (2014) Detection of fraudulent emails by employing advanced feature abundance. *Egyptian Informatics Journal*, [online] 153, pp.169–174. Available at:

http://dx.doi.org/10.1016/j.eij.2014.07.002.
Rathi, M. and Pareek, V., (2013) Spam Mail Detection through Data Mining – A Comparative Performance Analysis. *International Journal of Modern Education and Computer Science*, 512, pp.31–39.
Sharaff, A., Nagwani, N.K. and Swami, K., (2015) Impact of Feature Selection Technique on Email Classification. *International Journal of Knowledge Engineering-IACSIT*, 11, pp.59–63.
Sornsuwit, P. and Jaiyen, S., (2015) Intrusion detection model based on ensemble learning for U2R and R2L attacks. *Proceedings - 2015 7th International Conference on Information Technology and Electrical Engineering: Envisioning the Trend of Computer, Information and Engineering, ICITEE 2015*, pp.354–359.
Tahir, N.M., Hussain, A., Samad, S.A., Ishak, K.A. and Halim, R.A., (2006) Feature selection for classification using decision tree. *SCOReD 2006 - Proceedings of 2006 4th Student Conference on Research and Development 'Towards Enhancing Research Excellence in the Region'*, SCOReD, pp.99–102.
Vidros, S., Kolias, C. and Kambourakis, G., (2016) Online recruitment services: Another playground for fraudsters. *Computer Fraud and Security*, [online] 20163, pp.8–13. Available at: http://dx.doi.org/10.1016/S1361-3723(16)30025-2.
Vidros, S., Kolias, C., Kambourakis, G. and Akoglu, L., (2017) Automatic detection of online recruitment frauds: Characteristics, methods, and a public datase. *Future Internet*, 91, pp.1–19.
Wang, W.Y. and McKeown, K.R., (2010) 'Got You!': Automatic vandalism detection in wikipedia with web-based shallow syntactic-semantic modeling. *Coling 2010 - 23rd International Conference on Computational Linguistics, Proceedings of the Conference*, 2August, pp.1146–1154.
West, A.G., Kannan, S. and Lee, I., (2010) Detecting Wikipedia vandalism via spatio-temporal analysis of revision metadata? In: *Proceedings of the 3rd European Workshop on System Security, EUROSEC'10*. [online] New York, New York, USA: ACM Press, pp.22–28. Available at: http://portal.acm.org/citation.cfm?doid=1752046.1752050 [Accessed 20 Mar. 2021].

# APPENDIX A : RESEARCH PROPOSAL

## 1. Abstract

We are living in such a situation where the unemployment rates is as high as it could be in anytime in near future. Thousands of employees have been removed on a daily basis creating a global effect on dependencies between the company and their employees.
In these frantic occasions, when thousands and a huge number of individuals are keeping watch for a vocation, it gives an ideal chance to online con artists to exploit their distress. The aim of this study is to compare of the most popular models available for detecting the accuracy, which helps to identify the jobs which are true or fake.
Choosing an analysis model is necessary and can be difficult given the surplus of choices for this study, as it is used more than one model at a time to take advantage hide disadvantage of some models.
The selection of a good Analysis model will provide effortless performance of models in the system to deliver the best result. In this paper, we will see the various aspects of these seven models for analysing their ROC AUC and accuracy of the models. The comparison between the seven models will be done based on various parameters that can help analyse and decide which model will be better suited for different aspects.

## 2. Background and Related work

Job fraud is one of the most serious problems recently tackled in the area of online recruitment fraud. (ORF) In recent days, many organizations have opted to post their openings online so that job seekers can access them conveniently and quickly. However, since they give jobs in favour of getting money from them, job hunters, this purpose could be one form of Scam by fraudsters. It is possible to post fake work ads against a reputed business for breaching their reputation. In order to acquire an automated method to recognise fake jobs and report them to individuals to avoid applying for such jobs, this fraudulent work post identification attracts great interest.
During recent days we see ascend in fake employment posting where job posting appear quite sensitive, frequently these organisations will also have a designed site and they have enlistment procedure that is same very much in common like genuine organisation. In precise look on these posting it can be segregated from fake or genuine posting, more often its noted these fake posting doesn't have an organisation logo also the underlying reaction from the organization is from not having company domain email account or informal email account in many cases these job scammer has basically two approach.
First method where lure the applicants to fill some kind of application form to create database with ill motivated and sold these data to 3rd party, these data can be Telephone, full name & zip code etc. More sophisticated scammer may sell educational or professional experience details to send bulk email links to increase total page hits containing links, Scammer also contacts website administrator for this.

The second method is to use full data on fraud that could be subsequently used as part of economic offenses, such as racketeering of money and re-shipment of fraud. In this case, scammers claim the role of a genuine or imaginary employer and using the ATS as a mode for propagating resembles fake jobs. These reports direct users to additional contact techniques(i.e., site, email add. or tele. Num.). They will take part in a range of events from that point on, such as the distribution of fake skills tests, the organizing of fake interviews . the distribution of laudatory emails for successful go ahead for onboarding, etc.. The ultimate goal is to inadvertently force the victim to distribute highly sensitive information or become a "money mule" and use their bank accounts, such as social security numbers, identification cards, and passports Which is assistance to the laundering cash of the suspect.

In comparison to related online fraud concerns, ORF has not earned ample coverage yet, to date, it remains largely unknown. and It is clear to see that work scam detection shares similar features with related concerns**, Email spamming, phishing, cyber bullying, vandalism on Wikipedia, opinion fraud** and **trolling** for instance.

A recent well explained research on Online Recruitment Fraud Detection Model was put forward by (Alghamdi and Alharby, 2019) To differentiate scams or fraud objects from the data collection, the model core principle is to use the ensemble based classifier and SVM Algorithm for the feature selection. Another recent research on fake job recruitment detection using machine learning approach was put forward by (Dutta and Bandyopadhyay, 2020) where few classifiers are used, such as the , the Multi-Layer Perceptron Classifier, the K-nearest Neighbour Classifier, the AdaBoost Classifier, the Gradient Boost Classifier are used. First base got proposed in the year 2017 by (Vidros et al., 2017) where several analysis using Bag of Words model, Empirical Analysis, Geography and dataset complete evaluation are shown.

As fake job posting fraud resembles very much to most the online fraud hence Behaviour based email analysis with Spam Detection Application proposed by (Hershkop, 2006), professional classification algorithm of serious unethical activities, for example, inappropriate body dates format of the message or clear contradictions in users' past email behaviour. Another proposed paper in email spam by (Blanzieri and Bryl, 2008) The solutions suggested vary from different protocols for sender authentication to qualified classifiers that differentiate between regular and junk emails.

Another resembles online fraud is phishing which shares most of the common character where author(Jain and Gupta, 2018) where it does client-side identification of phishing websites using a machine learning method where Elevated accuracy of detection approach proposed by Misclassification of real websites as phishing (false positive) must be the minimum and accurate classification of phishing websites (true positive).

In above proposed paper machine learning approach is applied which employs several classification algorithms for recognizing fake posts. In this case idea to purpose Bidirectional LSTM (Ma and Hovy, 2016) deep learning model to predicate fake job posting.

## 3. Question for Research Study

- How to decide the relevant characteristics used in posting fake jobs?
- What is the right algorithm to use to assess the posting of fake jobs?

- Is the Bidirectional LSTM model suitable for the determine fake job posting?

## 4. Research Significance

- The study aims to reduce online recruitment fraud, Bidirectional LSTM deep learning model used here to predicate fake job posting or reduce fraud. Earlier studies show the usage of different classification technique gives significant accuracy and precision. The study aims to use some of the same techniques and apply in deep learning model.
- In most of earlier studies class imbalance techniques has not been explored much during this study proper class imbalance technique will get applied to get better prediction result.

## 5. Scope of the Study

This study stressed that online recruitment requires comparisons between areas such as email spam phishing, cyber bullying, and so on that were historically possibly the best. Furthermore, the only functional EMSCAD free dataset within such a scope and used in this study. The key contribution of this study is that it is only the fourth experimental study in this area.
In comparison to related online fraud concerns, ORF has not earned ample coverage yet, to date, it remains largely unknown. and It is clear to see that work scam detection shares similar features with related concerns**, Email spamming, phishing, cyber bullying, vandalism on Wikipedia, opinion fraud** and **trolling** for instance.

## 6. Aims and Objective

The main aim of this research is to propose a model to predict fake job posting using EMSCD (Laboratory of Information and Communication Systems, University of the Aegean, Samos, 2021) data also compare of the most popular models available for detecting accuracy, which helps to identify the jobs which are true or fake. The goal of this research to contribute in tackling employment scam detection which guide job seeker to get only legitimate offer from companies.
The research target is formulated on the basis of the purpose of this review, which is as follows:

- Using text, binary and HTML analysis to examine the pattern and relationship between distinct features.
- To propose effective balancing approaches that can be applied to the dataset of imbalances.
- To purpose Bidirectional LSTM deep learning model to predicate fake job posting.
- To compare between the different predictive model and identify the most accurate model among them.

## 7. Research Methodology

**Research Introduction** - Using Bi-directional LSTM network, we will design model to predicate fake job posting. The reason Bi-directional LSTM has been chosen because they have tried to learn how and when to forget and when not to using gates in their architecture. Using Google Colb (Bisong, 2019), data will be uploaded and pre-processed, then feature extraction will get applicable using text, binary and HTML analysis. We split pre-processed data into 80:20 train and test data then we fed the data to bidirectional LSTM model.

**Dataset Description** - EMSCAD is a set of data which are easy to access to the public where 17,880 actual work ads aimed at presenting a clear image of the online recruitment fraud problem to the scientific establishment that can serve as a desirable test bed object for researchers employed in the field,

Manually annotated EMSCAD records is divided into two groups. More precisely, 17,014 valid

and 866 fake work ads released from 2012 to 2014 are included in the dataset.

The dataset contains:

- 17,880 rows
- 18 features
    - 5 features (title, company profile, description, requirements and benefits) are long texts
    - Rest 13 features are mainly numeric fields or categorical data

The dataset is provided with Fraudulent column where value of 1 denotes the job is a fraud and 0 for real jobs. Dataset contains lot of missing values which are used as a valid observation. It could mean that fake posts often have missing fields.

**Data Prepossessing and Transformation -** Since we are also interested into text classification also and hence doing the analysis using EDA also for the feature ex: company which doesn't have logo, company which has less job description and requirement, company tend to mention salary in the subject line.
Following are data pre-processing data engineering:

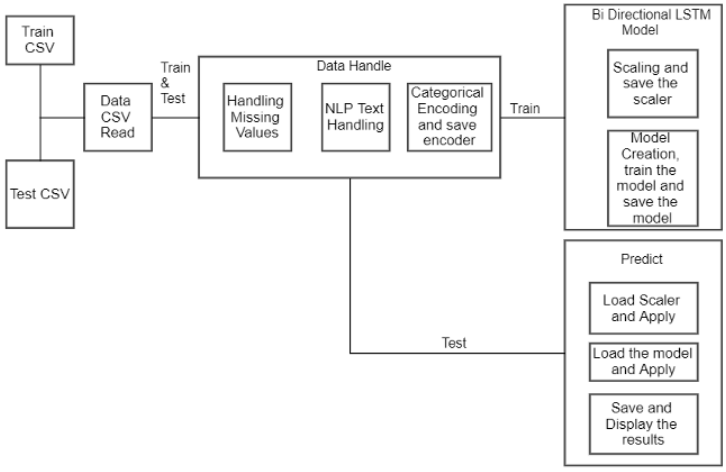
- Replace null to string "missing" - instead of dropping missing, use as valid observation. It could mean that fake posts often have missing data.
- Separate country, state and city from location column.
- Drop non-English text entries.
- Clean text columns - separate sentences, remove URLs, non-ascii characters, punctuation, extra spaces and white space.
- Redefine education bins - some rows have "some high school coursework" or "high school or equivalent" etc. which are replaced with "less than high school" for generalising it.

- Drop salary column: it is very often missing and unsure what units are used in foreign countries, inconsistent time frame. There is no way to standardise this column for such wide range of values.
- Using the natural language toolkit (NLTK) (Manning et al., 2014) , the job posts were tokenized with the word tokenize() function, which generates a tokenized text document., one of the commonly used natural language processing (NLP) libraries.
- Drop token with non-alphabetical characters.
- Commonly used stop terms like a, an, etc. that are overlooked by any search query when indexing the entries. There is no need for these terms because they reserve memory and require precious time to process.
- In reducing terms such as 'plays' and 'played' to their common form such as 'play,' Stemming plays a crucial role, so we used the porterStemmer class from stem library to our dataset.

**Smote Class imbalance -** SMOTE sampling on training data such that each class also has a number of observations. The 80/20 train/test split feature also does this. SMOTE: Synthesize new examples instead of an oversample for the minority class, which does not add any new details. SMOTE first generates a random example of a marginalised group and discovers its immediate neighbours in the marginalised group. The synthetic example is then generated by simply choosing one of the k nearest neighbours y and integrating x and y in the space of the function to form a line segment. The synthetic instances are generated as a conic integration of the two selected x and y instances.

**Models - Bidirectional LSTM**

After Stemming corpus using pad sequence need to create embedded docs for this corpus. Deep learning Bidirectional LSTM model with embedded feature vector with sigmoid activation function (Han and Moraga, 1995) with adam optimizer (Kingma and Ba, 2015). The adam optimiser chosen as it works well with noisy and sparse gradients.




**KNN Classifier -** K-Nearest Neighbours (KNN) algorithm mainly used to solve both

regression and classification machine learning problem. Based on similarity measurements, KNN algorithms use data and identify new data points (For Example: Dist. Func.) The naming of its neighbours is carried out by a plurality vote. (Cunningham and Delany, 2020)

**Decision Tree Classifier -** A test for an attribute is specified by every other node in the tree but every branch descending from that node corresponds to one of the distinct combinations of the attribute. (Tahir et al., 2006).

**Random Forest Classifier** - It uses packing and feature randomness when constructing each individual tree to strive to build an uncorrelated forest of trees whose board prediction is more accurate than that of any individual tree.

**Evaluation metrics -** After handling the data and scaling then we will utilize Bidirectional LSTM , Radom Forest, Decision Tree , KNN classifier model and compare these with Evolution matrix (M and M.N, 2015) Confusion matrix to show the result of different model, confusion matrix visualizes the performance of models.

| | Expected Class | | |
|---|---|---|---|
| Actual Class | | Positive | Negative |
| | Positive | True Positive | False Positive |
| | Negative | False Negative | True Negative |

TP (True Positive) was defined as the actual work posts that were properly classified as legitimate, FP(False Positive) was described as the actual job posts that were incorrectly categorised as false, FN(False Negative) is described as fake job posts that have been incorrectly identified as true, and TN(True Negative) was the bogus work posts that were correctly labelled as fake. In addition, the parameters used to assess the efficiency of the ML methods are: precision (Acc), recall (True Positive Rate or Sensitivity (TPR)), accuracy (Positive Predicted Value (PPV)) and F-measure (F1), Below are the equations.

$$\boldsymbol{Accuracy} = \frac{Number\ of\ correctly\ classified\ samples}{Number\ of\ total\ samples} = \frac{TP + TN}{TP + FP + FN + TN}$$

$$\boldsymbol{TPR} = \frac{Number\ of\ samples\ that\ correctly\ classified\ as\ real\ posts}{Number\ of\ samples\ that\ classified\ as\ real\ posts} = \frac{TP}{TP + FN}$$

$$\boldsymbol{PPV} = \frac{Number\ of\ real\ posts\ that\ correctly\ classified\ as\ real}{Number\ of\ samples\ that\ actually\ real\ posts} = \frac{TP}{TP + FP}$$

$$\boldsymbol{F1} = 2 * \left(\frac{Precision * Recall}{Precision + Recall}\right)$$

Accuracy, this is a parameter which describes the proportion of real estimates about the total number of cases taken into account. The accuracy, however, might not be sufficient to calculate the efficiency of the model as it does not take incorrect predicted cases into account. If a fake post is treated as a true one, it creates a significant problem. Hence, it is necessary to consider false positive and false negative cases that compensate to misclassification. For measuring this compensation, precision and recall is quite necessary to be considered (M and M.N, 2015)
Precision, The ratio of accurate positive results to the expected number of positive results. Recall, denotes the number of positive findings that are right divided by the number of all related samples. F1-Score or F- measure is the parameter dealing with both recall and accuracy is determined as the harmonic average of accuracy and recall.

**ROC & AUC:** Typically, ROC curves have a genuine positive rate on the Y axis and a fake positive rate on the X axis. This indicates the "optimal" point - a false positive rate of zero, and a true positive rate of one - is the top left corner of the map. This is not very practical, but it generally means that a wider area under the curve (AUC) is better. The "curvature" of ROC curves is also important since the true positive rate is desirable to maximise while decreasing the false positive rate.

In classification problem, ROC curves are usually often used study a classifier's performance. It is important to binarize the performance in order to expand the ROC curve and ROC region to a multi-label category. One ROC curve can be shown each label, but by treating each element of the label indicator matrix as a binary predictor, one can also create a ROC curve (micro-averaging).

## 8. Needed Resource

In the study, Experiment will be performed on PC and Cloud based GPU environment with following configuration of hardware, software and library requirement:

**Hardware Requirement –**

- **Local PC:**
    - Intel Core i7 10700K (8 Core, 16 Threads, Upto 5.1 Ghz)
    - 16GB DDR4 3000MHz.
    - Nvidia GeForce RTX 3070 8GB.
    - 256 GB S40G Nvme m.2 SSD.
- **GPU**
    - We can equip our machine with a high-end GPU like Nvidia's RTX 2080 TI or even its most strong Titan lineup for neural network training.

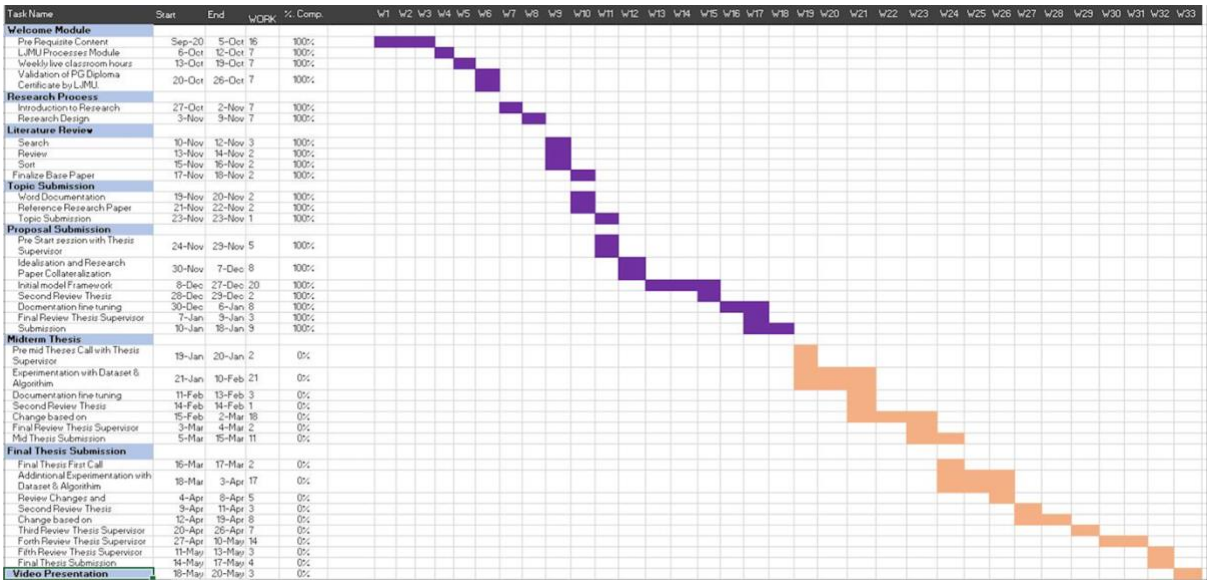
**Software Requirement -**

- Anaconda
- Python

**Libraries –**

- Pickle
- Pandas
- Nltk
- Sklearn
- Imblearn

# 9. Research Plan



| Task Name | Start | End | WORK | % Comp. |
|---|---|---|---|---|
| **Welcome Module** | | | | |
| Pre Requisite Content | Sep-20 | 5-Oct | 16 | 100% |
| LJMU Processes Module | 6-Oct | 12-Oct | 7 | 100% |
| Weekly live classroom hours | 13-Oct | 19-Oct | 7 | 100% |
| Validation of PG Diploma Certificate by LJMU | 20-Oct | 26-Oct | 7 | 100% |
| **Research Process** | | | | |
| Introduction to Research | 27-Oct | 2-Nov | 7 | 100% |
| Research Design | 3-Nov | 9-Nov | 7 | 100% |
| **Literature Review** | | | | |
| Search | 10-Nov | 12-Nov | 3 | 100% |
| Review | 13-Nov | 14-Nov | 2 | 100% |
| Sort | 15-Nov | 16-Nov | 2 | 100% |
| Finalize Base Paper | 17-Nov | 18-Nov | 2 | 100% |
| **Topic Submission** | | | | |
| Word Documentation | 19-Nov | 20-Nov | 2 | 100% |
| Reference Research Paper | 21-Nov | 22-Nov | 2 | 100% |
| Topic Submission | 23-Nov | 23-Nov | 1 | 100% |
| **Proposal Submission** | | | | |
| Pre Start session with Thesis Supervisor | 24-Nov | 29-Nov | 5 | 100% |
| Idealisation and Research Paper Collateralization | 30-Nov | 7-Dec | 8 | 100% |
| Initial model Framework | 8-Dec | 27-Dec | 20 | 100% |
| Second Review Thesis | 28-Dec | 29-Dec | 2 | 100% |
| Documentation fine tuning | 30-Dec | 6-Jan | 8 | 100% |
| Final Review Thesis Supervisor | 7-Jan | 9-Jan | 3 | 100% |
| Submission | 10-Jan | 18-Jan | 9 | 100% |
| **Midterm Thesis** | | | | |
| Pre mid Theses Call with Thesis Supervisor | 19-Jan | 20-Jan | 2 | 0% |
| Experimentation with Dataset & Algorithim | 21-Jan | 10-Feb | 21 | 0% |
| Documentation fine tuning | 11-Feb | 13-Feb | 3 | 0% |
| Second Review Thesis | 14-Feb | 14-Feb | 1 | 0% |
| Change based on | 15-Feb | 2-Mar | 18 | 0% |
| Final Review Thesis Supervisor | 3-Mar | 4-Mar | 2 | 0% |
| Mid Thesis Submission | 5-Mar | 15-Mar | 11 | 0% |
| **Final Thesis Submission** | | | | |
| Final Thesis First Call | 16-Mar | 17-Mar | 2 | 0% |
| Additional Experimentation with Dataset & Algorithim | 18-Mar | 3-Apr | 17 | 0% |
| Review Changes and | 4-Apr | 8-Apr | 5 | 0% |
| Second Review Thesis | 9-Apr | 11-Apr | 3 | 0% |
| Change based on | 12-Apr | 19-Apr | 8 | 0% |
| Third Review Thesis Supervisor | 20-Apr | 26-Apr | 7 | 0% |
| Forth Review Thesis Supervisor | 27-Apr | 10-May | 14 | 0% |
| Fifth Review Thesis Supervisor | 11-May | 13-May | 3 | 0% |
| Final Thesis Submission | 14-May | 17-May | 4 | 0% |
| **Video Presentation** | 18-May | 20-May | 3 | 0% |

## APPENDIX B: DATASET LINK

http://emscad.samos.aegean.gr/