THE DETECTION OF FAKE JOB POSTS BY USING K-NEAREST NEIGHBOR (KNN)

KHIN MAR HTAY

M.C.Sc. SEPTEMBER 2022

THE DETECTION OF FAKE JOB POSTS BY USING K-NEAREST NEIGHBOR (KNN)

By

KHIN MAR HTAY

B.C.Sc.

A Dissertation Submitted in Partial Fulfillment of the Requirements for the Degree of Master of Computer Science (M.C.Sc.)

University of Computer Studies, Yangon

SEPTEMBER 2022

ACKNOWLEDGEMENTS

First of all, I would like to express my inmost gratitude to **Dr. Mie Mie Khin**, Rector, the University of Computer Studies, Yangon, for allowing me to develop this research and giving me excellent guidance during the period of my dissertation.

I would like to express my deepest gratitude to my supervisor **Dr. Yu Yu Than**, Professor, Faculty of Computer Science, the University of Computer Studies, Yangon, for her caring and encouragement, and providing me with excellent ideas and guidance during the time of writing this dissertation.

I would like to express very special thanks to **Dr. Si Si Mar Win**, and **Dr. Tin Zar Taw**, Professors, Faculty of Computer Science, the University of Computer Studies, Yangon, as a Dean of Master's Course, for giving me the valuable guidance and suggestions during the development of this thesis.

In addition, I wish to thank **Daw Khin Hnin Wai,** Assistant Lecturer, English Department, the University of Computer Studies, Yangon. I would like to thank her for valuable supports and revising my dissertation from the language point of view.

Finally, I would like to extend my thanks to all my teachers who taught me throughout the master's degree course and my friends for their cooperation. The completion of my dissertation would not have been possible without the support and nurturing of my family.

STATEMENT OF ORIGINALITY

I here by certify that the work embodied in this thesis is the result of original research and has not been submitted for a higher degree to any other University or Institution.

------------------------- -------------------------

Date Khin Mar Htay

ABSTRACT

Every day, there are issues on many job-calling websites on the internet. Some of the jobs advertised online are actually fake jobs that lead to the theft of sensitive information. So it needs to be identified. Using the term frequency inverse document frequency (TF-IDF) method and the K-Nearest Neighbor (KNN) algorithm, the system detects fake job postings. TF-IDF is one of the statistical techniques for calculating the importance or score of each word in a document. In order to extract features, TF-IDF is employed. The purpose of this system is to classify real or fake jobs by using the KNN classifier. This system is implemented using 17866 jobs as the Fake Job Posts detection dataset. The accuracy of the proposed system is measured by using a confusion matrix (precision, recall, and F-Measure). The experiment results have been many times used with the existing actual data by using K-Nearest Neighbor algorithm with K value changes (K=1, 3, 5, 7, 9). According to the comparison results, the proposed system has achieved high accuracy many times.

CONTENTS

LIST OF FIGURES

LIST OF TABLES

LIST OF EQUATIONS

CHAPTER 1

INTRODUCTION

The development of the internet has made the recruitment process a much quicker process. Additionally, the present pandemic has had a big impact on the way that people are hiring these days. Online recruitment has made it possible to find more applicants and streamline the recruiting process, and it has been very helpful in connecting the gap between employers and potential candidates. On the internet, job seekers can now quickly and easily apply to a wide range of jobs according to their specialization with simply a click of a button. With the aid of E-recruitment, businesses utilize a variety of internet-based solutions. Users can expand their job searches and find the best applicants by using online recruitment. This makes it easier for them to connect with qualified applicants from all over the world. When the client relies on online recruitment, it ends up hiring the best applicant. Companies can choose the most qualified candidates and improve efficiency by using tools like pre-employment screening, personality assessments, and tests for candidate screening. Therefore, there is very little human influence in this process. In terms of communication, online recruitment has a cost-effective advantage.

However, some of these advertised positions are only fake jobs used as traps to take potential data instead of actual jobs. When candidates apply for these jobs, it's possible that their personal information will be stolen, or in some situations, their computers may be hacked to steal important information. Cybercriminals combine the victim's data and either resells it on the dark web for use by another, or continue to use for years. This fake job post detection attracts considerable attention for developing an automated solution for classifying fake jobs and reporting them to people to avoid applying for such jobs [4].

In the proposed system, the fake jobs can be detected and data theft can be avoided using Term Frequency-Inverse Document Frequency (TF-IDF) and K-Nearest Neighbor algorithm. This system consists of three steps: preprocessing, feature extraction, and identifying fraudulent job by the KNN classifier. Lastly, this system evaluates the accuracy with precision, recall and f-measure with different K values.

1.1 Objectives of the Thesis

The main objectives of this thesis are:

- To apply K-Nearest-Neighbor algorithm for a fake job posting detection.
- To study TF-IDF method which is used for the feature extraction of words.
- To prevent stealing about personal information during looking for jobs.
- To evaluate the performance of the classification fake jobs by using confusion matrix.

1.2 Motivation of the Thesis

Nowadays, the process of online job recruitment is changing. On online job recruitment sites, employers advertise job vacancies, and applicants submit their applications online. It has been proved that this online hiring process is more time-efficient and less stressful for both employers and job seekers. But as was already said, these processes have their disadvantages as well. There have been reports today of a number of crimes involving online recruiting. Many people had to face the loss of personal information as a result of this process. Scammers frequently post advertisements with enticing promises that attract people's attention to those manufactured posts. People frequently fall victim to this trap because people don't know enough about the issue. Such occurrences included monetary losses and the disclosure of confidential information, like bank account numbers, etc. [27]. Additionally, scammers are getting increasingly good at producing these adverts, making it challenging for people to distinguish between real and false posts. In some extreme circumstances, people may receive calls for interviews as a result of these ads, and may later be subjected to detrimental consequences. For the abovementioned reasons, it is now essential to confirm if a job posting is real or a fake.

1.3 The Overview of the Proposed System

Today, people are living in a digital age, and the internet is the main tool that this digital world uses. The internet is now used for everything, including banking and recharging out mobile phone balances. In the same way, the internet is regarded as the best resource both for applying and seeking jobs. Now people don't need to physically go from one office to another to collect job circulars or to apply for a job. With the use of the internet by sitting in their own home people can quickly seek for new jobs and apply within a few seconds. But just as everything has a positive and a negative side, so does this internet approach for hiring people. The most dangerous and threatening type of crime is cybercrime.

In this system, TF-IDF and KNN is utilized to categorize and detect real jobs or fake jobs in a dataset of job postings. In addition, it can be noted that, data cleaning is an important stage in this proposed system. The accuracy of the system is measured by confusion matrix. This system will give the better result in detecting the fake jobs or real jobs due to the combination of TF-IDF (feature extraction) and KNN algorithm.

1.4 Related Works

Nasser et al. illustrated a number of machine learning classifiers, including Multinomial Naive Bayes, Support Vector Machine, Decision Tree, K Nearest Neighbors, and Random Forest in a text categorization problem,. The data contains both real and fake job postings. For feature extraction, the cleaned and preprocessed data were applied to TF-IDF. The evaluation metrics utilized are accuracy, precision, recall, and f-measure, with one attribute for description. Finally, the system made decision that Random forest classifier and K-Nearest Neighbor classifier achieved the highest recall [14].

Alghamdi et al. created a model for identifying fraudulent job postings in online job ad systems. The authors had utilized the Employment Scam Aegean Dataset (EMSCAD) dataset on several machine learning algorithms. The methodology consists of three steps: preprocessing, feature selection, and classifying fraudulent by the classifier. In this stage, tags and one unnecessary noise are removed from the data and added to the general text. Selective features are chosen with the use of a support vector machine and random forest classifier to reduce extraversion features that are underutilized. According to reports, the classification accuracy for detecting fraudulent job postings was 97.4% [7].

Van Huynh et al. used deep neural network models that have been retrained using text datasets. The classification of IT-related jobs was done. Text CNN, Bi-GRU CNN, and Bi-GRU-LSTM CNN were the models utilized. There are layers of convolution and pooling in the text CNN model, which is fully associated (Mujtaba et al., Mujtaba & Ryu). Layers were used during the training (convolution and pooling). This model uses the Softmax function for classification, and an ensemble classifier was used to increase accuracy. According to text CNN, the reported accuracy was 66%. BiGRU-LSTM CNN accuracy is 70% [26].

1.5 Organization of the Thesis

This structure of this thesis is five chapters, abstract, acknowledgement and references.

In chapter 1, fake job post detection system is introduced. This chapter further discusses objectives of the thesis, the motivation, the system overview, related works, and organization of the thesis are described. In chapter 2**,** the fundamental of the machine learning techniques are described. Among the various types of machine learning classifier, Support Vector Machine, Decision Tree, K Nearest Neighbors, Artificial Neural Network Algorithm and Random Forest are briefly explained in this chapter.

In chapter 3, depicts the system design , datasets, Preprocessing, Feature Extraction, the K Nearest Neighbors and term weighting schemes (TF-IDF) method calculations are discussed. In chapter 4, expresses implementations and experimental results with figures are described. In chapter 5, the conclusion of the research work is presented. In addition, advantages and further extensions of the system are depicted.

CHAPTER 2

BACKGROUND THEORY

The internet is filled with numerous job advertisements, even on the reputed job advertising and marketing sites, which never seem fake. But after the selection, the so-called recruiters start requesting the cash and the financial institution information. A large number of applicants fall into their lure and every so often lose their current paintings as well as a variety of cash. Therefore, it is higher to discover whether or not a process posting on the site is genuine or fraudulent. It is very difficult and nearly impossible to identify it manually. This proposed system apply machine learning to teach a version for fake task classification. It could be trained the usage of previous real and fake job postings, and it is accurate at identifying a fraudulent job [25].

The various machine learning algorithms based on labeled or unlabeled datasets are presented in this chapter. The background theory is briefly explained for several machine learning classifiers, including Decision Tree, Support Vector Machine, K Nearest Neighbors, Artificial Neural Network (ANN), Random Forest and text classification.

2.1 Overview of Machine Learning

Machine learning is one of the fastest growing areas of computer science, with a wide variety of approaches. It implies to the automated recognition of meaningful styles in information. Machine learning tools are concerned with endowing programs with the capability to research and exchange. Machine Learning is becoming one of the foundations of Information Technology and with that, as an alternative relevant, albeit usually hidden, part of our life. With the ever increasing amount of information becoming to be had, there is a superb motive to count on that clever facts analysis becomes even greater universal as a necessary aspect for technological progress. There are many applications for Machine Learning (ML), the maximum significant of that is records mining. People are frequently at risk of making mistakes in the course of analyses or, may be, while attempting to set up relationships among numerous functions. Data Mining and Machine Learning are Siamese twins from which numerous insights can be derived through suitable studying algorithms. There has been fantastic development in data mining and machine learning due to the development of clever and Nano era, which delivered approximately curiosity in finding hidden styles in records to derive value [19].

The Machine Learning (ML) field manipulates computational programs which might be capable of imitating the getting to know system in humans. It has been used for predictive analytics in a number of industries, including the analysis of call patterns in telecommunications, diagnosis models in medical, and the analysis of substantial amounts of data for various reasons in theoretical research. Because of ML is considered as artificial intelligence, its fashions must be able to research, and reply to the changes in its surroundings, so it is able to offer solutions for all feasible circumstances. ML models learn by improving their performance based on past experience (training data). The majority of machine learning (ML) classification programs are learnt through supervision, supervised learning involves providing the model with already classified data, so learning makes progress by comparing the actual versus predicted class and adjusting their mathematical equations till it got proper margin between the real elegance and the anticipated one. In the other hand, there may be unsupervised gaining knowledge of, in which is not any labels (classes) provided, The model must find patterns within the supplied information, produce affordable labels, and then try to map training facts with those labels.

Supervised machine learning strategies applied on a text classification problem, which is have become an important studies subject matter because of the big amounts of textual content statistics and files that cope with online [8]. Machine learning is being applied to one-of-a-kind packages such as business intelligence, Facebook recommendation engines, human resource information systems (HRIS) and self-riding vehicles. Based totally on the information kind, i.e., labeled, or unlabeled records, there are numerous styles of ML algorithms; typically, these algorithms are accumulated into two classes: unsupervised and supervised learning. Figure 2.1 is shown the machine learning algorithms.

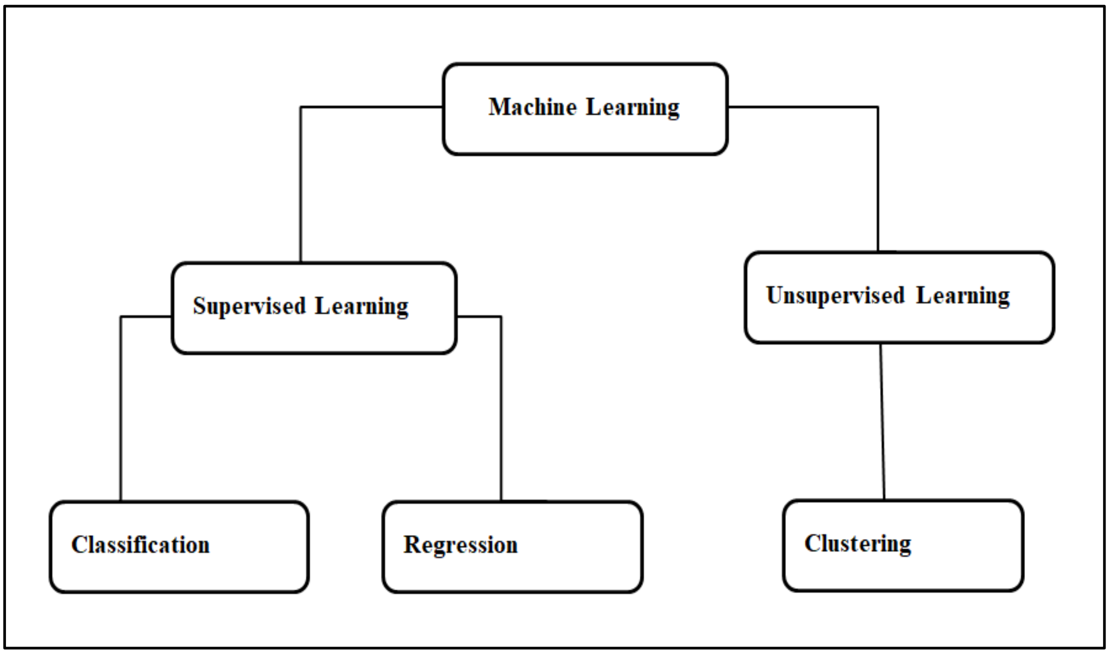


Figure 2.1 Machine learning Algorithms

(i) Supervised Learning

Supervised learning is the machine learning task of learning a characteristic that maps an input to an output depending on input-output pairs. It infers a feature from classified getting ready records along with a fixed of preparing instances. In supervised mastering, every example is a pair inclusive of an enter object and the desired output fee. This set of algorithms uses getting ready statistics to generate a function that maps the inputs to desired outputs (additionally referred to as labels). As an example, within the troubles of type, the gadget seems at instance statistics and uses it to arrive at a function mapping enter records into classes [8]. Supervised learning can be divided into two forms of problems: class and regression. Supervised learning algorithms are as shown in figure 2.2.

(a) Classification: Classification is an approach for predicting the class of given records factors. Classes are once in a while known as objectives/ labels or classes. Jobs detection can be identified as a classification problem. There are only two labels as fake and real jobs. A classifier utilizes some training statistics to apprehend how given input variables relate to the elegance. In this example, regarded fake and real jobs need to be used as the training data.

When the classifier is trained accurately, it could be used to discover fake jobs [19]. Support vector machines (SVM), decision trees, random forest and K-nearest neighbors are all common types of category algorithms.

(b) Regression: Regression is another kind of supervised learning method that uses a set of rules to apprehend the relationship between structured and impartial variables. Regression is a system that allows modeling the non-stop-valued capabilities. For that reason, the consequent regression model can be used to predict the unattainable of missing values for numerical facts more comfortably than magnificence labels of discrete statistics. A few famous regression algorithms are linear regression, logistic regression and polynomial regression.

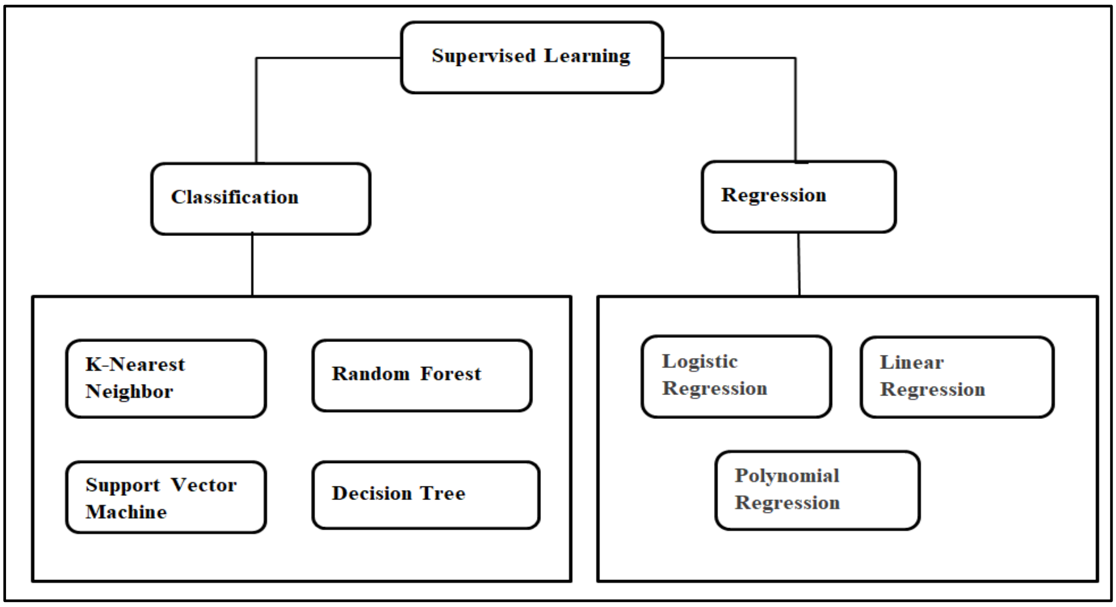


Figure 2.2 Supervised Learning Algorithms

(ii) Unsupervised Learning

Unsupervised learning is a kind of machine learning in which the calculation is not provided with any pre-assigned labels or ratings for the making ready facts. Unsupervised learning algorithms should first self-discover any evidently occurring patterns in those getting ready parameters. This set of calculation works without previously labeled statistics.

The number one purpose of those calculations is to locate common bureaucracy in previously unseen records. The most common type of unsupervised learning is clustering. Other types of unsupervised learning include self-organizing maps and hidden Markov models. Algorithms for unsupervised learning include clustering, anomaly detection, neural networks, and so on. These calculations discover hidden patterns or parameter groupings without the obtained for human intervention [8].

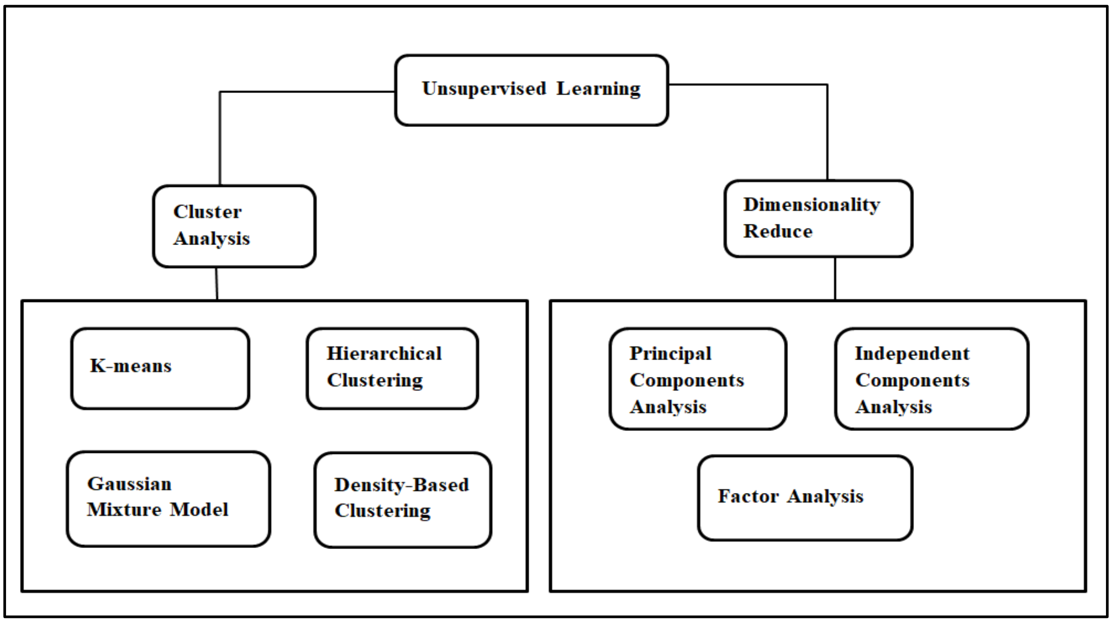


Figure 2.3 Unsupervised Learning Algorithms

(iii) Reinforcement learning

Reinforcement Learning (RL) is a subfield of artificial intelligent where learning is achieved when a specialist collaborates with its current environment to accomplish an objective. This is frequently used in situations where a framework must decide how to act optimally in a given situation in light of positive or negative feedback on a previous move. The framework is trained for the situation based on specific information boundaries alongside a measure of the score.

In the wake of understanding the principal thought of how the calculation works, it is worth attention on the positive and negative parts of RL. The main benefit is that there is a harmony between taking a stab at something previously demonstrated and trying new things to arrive at additional improvement. This makes the calculation more solid to attempt new patterns or arrangements in a steady configuration. That is the reason it can find new bits of knowledge and methods to expanding the predictive power. A potential inconvenience could be that is difficult to expect to consolidate express principles later on. Moreover, a great deal of information data sources may be necessary for the machine to get the legitimate input. Then again, RL is more challenging to implement and demands a lot of mastery compared with the other two artificial intelligent strategies [20].

2.2 Common Classification Algorithms

Classification is one of the most important functions of a supervised learning method. Informational approaches and machine learning algorithms are used in this field to solve real-world issues. On both structured and unstructured data, the classification can be performed. It determined the classification algorithm that would be used to determine the target that new data would observe on training data [17]. The classification method involves learning from the provided dataset or training corpus before classifying the new observation into several targets or categories. The classification task for each training set of data is performed under supervision and supported by real training set data results. In general, there are two main steps in the classification process:

- **Training Phase**: In this step, which is known as the construction of a classification model, several methods are applied to create a classifier that builds a model that learns from a variety of training samples. This model needs to be trained to predict results with the highest accuracy.
- **Testing Phase**: In this step, the model is used to predict class labels by evaluating the developed model on test samples and estimating the accuracy output of the classification rules.

The amount of data available on the internet is increasing at an unstoppable rate. The need for tools to go through or organize these enormous amounts of information has also grown. Many tasks involved in information management depend on classification. As a result of these facts, the classification process must not only perform better but also maintain accuracy rates.

As the most of the information input was in the form of text, automatic classification of text documents is beneficial when dealing with the enormous textual data, such as in spam filtering systems, library management systems, news classification systems, and organizational data analysis. In the classification approach, various theories and machine learning approaches have been applied [22].

Human experts are capable of classifying documents, but are unable to keep up with the rising datasets, and the task takes a lot of time. Additionally, human errors in the process also cause delays and add to the costs. Therefore, automatic text classification is a key technique for intelligent information systems that has gained a lot of attention recently. It is a powerful tool for organizing data. Text classification will become a more important approach in the digital age as the necessity for it grows along with its commercial value. This method has been used to categorize new electronic documents, discover interesting web content, and direct user search using hypertext. It's important that the text categorization model is accurate and effective. The quantity of training set data and the model's fitness determine how accurate it is. With many classification methods, the usefulness of an approach depends on comprehension, model development, computing cost, and outcome accuracy measures [24].

Many information retrieval problems, including filtering, routing, and looking for vital information, benefit from classification research, which is an exciting field of research [1]. In general, the various types of classification approaches for discovering knowledge are Support Vector Machine (SVM), Bayesian Networks (BN), Rule Based classifiers, Decision Tree (DT), K- Nearest Neighbor (KNN), Artificial Neural Network (ANN) and Random Forest, etc. This section shows some of them, including the basic classification methods.

2.2.1 Support Vector Machine Algorithm

The Support Vector Machine (SVM) algorithm is based on the structure-based risk minimization principal and statistical learning theory. The purpose of SVM is to determine the position of decision boundaries, which are represented by a hyper plane or decision plane and generate the best class separation. A decision plane is a concept used to define decision boundaries that allow learning regression, classification, or ranking functions. A hyper plane creates divisions between a set of data with various class memberships [6]. It is an illustration of the training data as points in space that clearly divide each category as widely as possible. The original training data is converted into a higher dimension via a nonlinear mapping. The prediction of a category based on which side of the gap the new instances fall is then performed on the same space in which are mapped.

Initially, the SVM was created as a learning system for binary classification. However, it is difficult to apply to for multi-classification problem. It specifically divides a large optimization problem into a number of smaller problems, where each problem only consists of two carefully chosen variables to allow for efficient optimization. The iterative nature of this technique allows for successful resolution of all decomposed optimization problems. The SVM approach, even with a very high input space dimension, is regarded as a good classifier due to its high generalization performance without requiring a priori knowledge. The SVM classifiers have gained a lot of attention in recent years due to their ability to classify data in the most reliable and precise manner possible and their broad range of useful applications.

2.2.2 Decision Tree Algorithm

A decision tree is a classification that is defined as a recurrent split of the instance space. The decision tree is made up of nodes that together form a rooted tree, which is a directed tree with a node called "root" that has no outgoing edges. Every other node has precisely one incoming edge. An internal or test node is referred to as having outgoing edges. The leaves are all other nodes (also known as decision nodes or terminal). Each internal node in a decision tree splits the instance space into two or more sub-spaces in accordance with a certain discrete function of the value of the input attributes. Each test takes a single attribute in the simplest and most frequent case, dividing the instance space according to the attribute's value. The condition applies to a range in the case of numeric attributes.

Each leaf is provided a class that represents the most appropriate target value. As an alternative, the probability vector for the target attribute having a particular value may be stored in the leaf. According to the results of the tests along the way, instances are categorized by moving them from the tree's root all the way down to a leaf. A decision tree that evaluates whether a potential client will reply to a direct mail message. Internal nodes are shown as circles, while leaves are shown as triangles. Both nominal and numerical qualities are present in the decision tree in this instance. Given this classifier, the analyst can understand the behavioral characteristics of the entire population of possible customers for direct mailing and predict the response of a potential client (by descending it down the tree). The characteristic that each node tests serves as its unique identifier, and the values of that attribute are used to label the branches of each node.

Decision trees can be graphically visualized as a collection of orthogonal hyper planes when dealing with numerical attributes. Generally, decision-makers choose simpler decision trees since it may be regarded of as more comprehensible. Additionally, its accuracy is greatly affected by the complexity of the tree. The tree complexity is explicitly controlled by the stopping criteria used and the pruning technique employed. The total number of nodes, the total number of leaves, the depth of the tree, and the number of characteristics employed are typically used to determine how complicated a tree is. Rule induction and decision tree induction are closely linked. [16].

Additionally, this classifier can use a bottom-up approach to resolve the over fitting problem by applying various weights to the training data. The training dataset is analyzed by this method, which subsequently develops a classifier that may be able to appropriately arrange both the training and test datasets. The classifier model may use an input object as a test case, and it will forecast the outcome of an output value. This algorithm chooses one feature for each node of the decision tree that can most effectively divide the collection of samples into subsets such that the outcomes can fall into one of two categories. In this instance, the splitting requirement is the normalized information gain, which is the distinction of a non-symmetric measure between two probability distributions. The attribute with the greatest information gain is chosen for decision-making.

2.2.3 Artificial Neural Network Algorithm

An Artificial neural network (ANN) is also called a “Neural Network” (NN) which is a computational or mathematical model primarily based on biological neural networks. Especially, it is an emulation of biological neural machine containing an interconnected organization of synthetic neurons to procedure statistics using a connectionist technique for computation. First off, a neuron gets the input indicators from its enter links, after which computes an output sign and transmits this sign over its output links. An enter sign can be both raw information and outputs of other neurons. The output sign can be an enter to other neurons otherwise a very last answer of the problem. In most instances, an ANN is an adaptive machines that modification its shape primarily based on inner or external information at some point of the network for the duration of the gaining knowledge of segment.

Realization of an ANN taking place a single computer, it's far typically slower than extra conventional algorithms. Being the parallel nature of ANN, it could paintings with multiple processors that offer a top notch velocity advantage at very little development price. This parallel architecture permits ANN to deal with the massive amounts of records very successfully in less time [30]. Whilst managing incredible non-stop streams of data (i.e. machine sensor or speech recognition data), ANN can operate significantly faster than some other algorithms. As properly, ANN is useful in several real-global packages inclusive of textual content-to-speech, handwriting analysis, speech popularity and visible sample popularity handling complex and incomplete facts.

2.2.4 Random Forest Algorithm

Random forest is a supervised Machine Learning Algorithm that is used extensively in classification and regression problems. It builds choice trees on distinct samples and takes their majority vote for class and common in case of regression. One of the most vital capabilities of the random wooded area algorithm is that it can manage the data set containing non-stop variables as within the case of regression and specific variables as in the case of category. It plays higher consequences for class issues. That’s as it consists of a couple of selection trees just as a forest has many trees. On top of that, it makes use of randomness to beautify its accuracy and fight over becoming, which may be a big issue for such a complicated algorithm. These algorithms make selection bushes primarily based on a random choice of statistics samples and get predictions from every tree. After that, pick the nice feasible solution thru votes. It has several packages in daily lives which includes feature selectors, recommender systems, and picture classifiers. A number of its real-lifestyles programs consist of fraud detection, type of loan packages, and disease prediction [14].

2.3.5 K-Nearest Neighbor Algorithm

The KNN classifiers are built totally on gaining knowledge of an analogy that the training samples are described by way of n dimensional numeric attributes. Here, those training samples are required to be in memory on the run time and accordingly it is also denoted as the memory-primarily based technique. When given an unknown tuple, the KNN algorithm finds the pattern area for the k training tuple that are closest to the unknown tuple. When k=1, the unknown tuple is assigned the class of the training tuple that is closest to it in sample area [18]. The KNN classifiers can also be used for prediction. The distance characteristic is common to use the Euclidean distance, however other distance measures, consisting of the Manhattan or Hamming distance can theoretically be used as an opportunity.

In such scenario, the KNN classifier returns the average value of the real-valued related to k nearest neighbors of that unknown data sample. This algorithm is one of the simplest algorithms for implementation, effective despite the fact that training data is massive and robust for noisy training data. By taking this advantages, this thesis is exploited the KNN set of rules as a classifier to detect real or fake jobs. The precise strategies of the way decided on KNN search algorithm work by means of focusing on the fake job posts detection system will be discussed in Chapter 3.

(i) Development of KNN

K-Nearest-Neighbor classification was invented to carry out characteristic analysis when it was uncertain or challenging to identify clear parametric approximations of probability densities. Fix and Hodges presented a non-parametric approach for pattern categorization in a 1951 paper for the US Air Force School of Aviation Medicine that was never published. This algorithm is now known as the K-closest neighbor rule.

TO DETERMINE K: The ideal value of K is chosen after carefully examining the available data. Although this isn't ensured, greater K upsides are more accurate because reduce net noise. Using cross approval, it is also possible to resolve a respectable value of K. If K=1, the information is essentially assigned to the class of its nearest neighbor. For the preparation information, the error rate is absolutely zero at K=1. This occurs because the nearest highlight to any preparation-related information item is the preparation itself. Consequently K should equal 1 to achieve the optimal results. The limits, however, are over fitted when K=1.

The calculation is too delicate to even take noise into account in the case of minor upsides of "k." The preparation and approval set needs to be separated from the underlying dataset in order to obtain a reliable value for K. Assuming the two closet neighbors (K=2) have a place with two unique classes, the outcome is obscure. The number of closest neighbors is increased to a higher value in this method (say, five closest neighbors). This will distinguish an earliest neighbor region and provide clarity.

Larger upsides of "K" smooth out class boundaries, which is probably not appealing because different classes' grades might then be recalled for the neighborhood. It can be difficult to determine the value of K even while the preparatory information focuses are diffusely available.

(ii) Advantages of K-Nearest Neighbor Algorithm

K-Nearest Neighbor is known for its effortlessness, intelligibility and adaptability. It is not difficult to decipher. The estimation time is much less. Likewise, the prescient strength is exceedingly excessive which makes it compelling and powerful. Ok-nearest neighbor is relatively compelling for large training set. The manner continued in the grouping achieved with the aid of this calculation are fairly less confusing than that accompanied via exceptional calculations.

The numerical calculations aren't difficult to fathom and understand. The system does not consist of computations that look like difficult. Fundamental ideas like that of Euclidean distance estimation are utilized which upgrade the effortlessness of the calculation instead of deciding on other composite strategies like that of incorporation or separation. Its miles valuable for non-direct statistics. K-nearest neighbor is compelling for characterization as well as relapse.

(iii) Disadvantages of K-Nearest Neighbor Algorithm

K-nearest neighbor can be high priced in warranty of the facts is big. It needs a more prominent stockpiling than a powerful classifier. In K-nearest neighbor the expectation degree is delayed for a larger dataset. Likewise, calculation of precise distances assumes a chief component within the warranty of the calculation's precision. One of the significant stages in KNN is decided the boundary K. On occasion it isn't clear which sort of distance to utilize and which aspect will supply the first-rate outcome. The calculation cost is very high as the distance of every getting ready version is to be determined. K-nearest neighbor is a gradual gaining calculation as it does not gain from the education facts; its concept keeps it and afterward utilizes that statistics to symbolize the new facts.

(iv) K-Nearest Neighbor and its Variants

As examined before, the effectiveness of the calculation may be advanced with the aid of causing adjustments inside the variables that to supervise it. There are various variations of K-nearest neighbor which have been focused before to make this calculation extra a success, some of them are:

1. Locally Adaptive K-nearest neighbor:
   Regionally versatile k-nearest neighbor calculations proposed via [28]. It selections the well worth of ok that should be utilized to reserve a contribution with the aid of searching at the aftereffects of go-approval calculations in the close by community of the unlabeled facts.
2. Weight Adjusted K-nearest neighbor:
   The calculation by way of [13] recommends that the distances, on which the search for the closest associates is situated inside the initial step, must

be modified into comparable measures, which may be utilized as hundreds. The relegated loads conclude how lots a first-class influences the characterization interest. This classifier is particularly treasured for the state of affairs in which a dataset has many factors, some of which can be viewed as un-crucial; however it has excessive computational cost.

3. Improved K-nearest neighbor for Text Categorization:

   It proposes a refined K-nearest neighbor calculation for text order, which builds the characterization model by combining K-nearest neighbor text classification and confined one pass grouping calculation. On the off chance that a steady worth of K is utilized for every one of the classes, the class with bigger number of properties will enjoy a benefit. In better KNN, a reasonable number of closest neighbors are utilized by the conveyance of information in preparing sets, to foresee the class of unlabeled information [23].

4. Adaptive K-nearest neighbor:

   K-nearest neighbor recognizes same number of closest neighbors for each new information. Versatile K-nearest neighbor by [28] figures out a fit worth of K for each test. Initial an ideal worth of K is found. Then, at that point, to anticipate the arrangement of the unlabeled information, the worth of K is set equivalent to the ideal worth of K of its closest neighbor in the preparation data. The implementation of the proposed calculation is then tried on various datasets.

5. K-nearest neighbor with Shared Nearest Neighbors:

   A better K-closest neighbor calculation is introduced by utilizing divided closest neighbor comparability which can figure likeness among test tests with closest neighbor tests. It utilizes Similarity judgment calculation and works out the closest neighbor similitude an incentive for each preparing test. Then it ascertains the most extreme between these qualities.

6. KNN with K-Means:

   One more ad limbed way to deal with the calculation is portrayed by [29]. This calculation attempts to isolate a bunch of focuses into K sets or groups to the focuses in each bunch are near one another. The focuses of these newly made groups are taken as the new preparation tests. To foresee the characterization of an unlabeled information, its separation from the recently found preparing focus is determined, and the middle what shares the base separation from the information is allotted to that group. Dissimilar to standard K-nearest neighbor, there is the information boundary K isn't passed. This record is one of its benefits.

7. SVM K-nearest neighbor:

   Support Vector Machine (SVM) is an order strategy that can be applied on straight as well as non-direct information. It is a composite variant of K-nearest neighbor blended in with SVM for visual classification acknowledgment, and is expanded. In this calculation, the preparation is finished with the assistance of K closet neighbors to the un-named data of interest. To begin with, the K-closest information not set in stone. Then, at that point, pairwise distance between these K information focuses is processed. Subsequently get a distance framework from the determined distances. A Kernel network is then planned from the got distance framework. This bit framework is taken care of as contribution to SVM classifier. The outcome acquired is the group of the obscure piece of information. Then again, one could utilize SVMs yet time utilization is one of its downsides. Additionally, it includes estimation of pairwise distances.

8. K-nearest neighbor with Mahalanobis Metric:

   The measurement distance is critical in grouping of another data of interest. Mahalanobis is another distance metric, approach of which is canvassed .The metric guarantees that the K-closest neighbors are include in similar class and the examples having a place with various classes are isolated by an enormous level of contrast.

9. Generalized K-nearest neighbor:

   KNN can likewise be utilized for constant – esteemed class credits. For this arrangement, the typical quality determined among neighbors is allotted to the group property of the unlabeled information. It implements this calculation to foresee the constant – esteemed group characteristic.

10. Informative K-nearest neighbor:

    Typically the worth of okay depends on the records, making it difficult to pick the boundary as in keeping with diverse programs. It is offered any other metric that movements the enlightening ness of objects to be grouped. Schooling ness estimates the significance of focuses. On this approach, there are two facts barriers K and I. The extra element magnificence of most academic coming down fashions will be the elegance of the brand new test.

11. Bayes K-nearest neighbor:

    The information values encompassing the objective are created by using a comparable likelihood conveyance, extending outwards over the reasonable variety of buddies. It is recursively figured the probability of the closing exchange-factor and moved toward the goals, and registered the back likelihood dispersion over okay.

2.3 Text Classification

Text classification, also known as text categorization, is one of the main tasks in contemporary herbal language processing, which aims to assign labels or tags to textual gadgets together with sentences, queries, paragraphs, and files. Each text category problem follows similar steps and is being solved with extraordinary algorithms. Usually, text classification responsibilities may be done with rule-based strategies or gaining knowledge of-based totally (records-pushed) techniques. With the rule of thumb-based methods, the devices classify documents primarily based on sure styles such as key phrases or phrases, additionally referred to as normal expressions.

This kind of techniques is typically utilized in natural language processing responsibilities earlier than gadget studying techniques were extensively adopted. Rule-based method has its very own demanding situations: tailoring phrases or rules for particular pattern matching need a lot of guide rule writing and correction with several feedback loops to ensure trap the right ordinary expressions for the type responsibilities at positive self-belief level. Because simplest have a look at the pattern (keywords or phrases) itself without thinking about the context, rule-based method normally necessarily get many fake positives category. Additionally without close appearance on the context, its miles tough to seize the sentiment in the back of phrases, for you to make rule based totally method a chunk assignment for type responsibilities worried with sentiment factors. With learning-based (statistics driven) methods, the machine typically adopt gadget getting to know algorithms on the way to obtain a more holistic and meaningful information mapping and textual content category. Most classical device learning based methods follow the famous two step procedure, in which in the first step a few homemade features are extracted from the files (or every other textual unit), and inside the second step those features are fed to a classifier to make a prediction.

A mastering-based totally text type application usually follows those steps: 1. Text preprocessing & cleaning 2. Characteristic engineering to create handmade features from textual content 3. Characteristic vectorization (tf idf, count number vectorizer, etc.) or embedding (word2vec, doc2vec, bert, and many others.) four. Model education with device gaining knowledge of or deep studying algorithms. Besides the classical and famous machine studying classifiers like okay-nearest associates or random forest, there are greater than one hundred fifty deep learning frameworks proposed for diverse textual content category tasks. On this mission, it explores each method for detection of seasoned-social messages in online process postings and seeking to examine the performance between two strategies, which also offer method selections for comparable obligations in applicable area. In particular, beneath technique start with an unmanaged phrase embedding version to generate key phrases and terms that suggests pro-social messages, which name a seasoned-social dictionary and then use a rule-based approach to catch the ones keywords to further classify the task postings. [31].

2.3.1 Term Weighting Schemes (TF-IDF method)

TF-IDF is the well-known algorithm used in text mining research to calculate the weight of each term [12]. TF stands for term frequencies and IDF stands for inverse document frequency. It brings attention to a specific problem that might not be mentioned frequently in our corpus but is yet quite important. The TF-IFD score increases proportionally to a word's frequency in a document and decreases as the number of documents in the corpus that use the word raises.

Term Frequency (TF): The term frequency indicates how frequently it appears throughout the entire document. The probability of discovering a word within the document might be considered. It calculates the wide variety of instances a word t occurs in a document d, with appreciate to the whole quantity of words within the document d. It's far formulated as:

$$tf(t,d) = \frac{number\ of\ term\ t\ appear\ in\ a\ document}{total\ number\ of\ term\ in\ the\ document} \tag{2.1}$$

Inverse Document Frequency (IDF): The inverse document frequency measures a term's frequency or frequency across all documents in the corpus. It draws attention to words that are rare or have a high IDF score, or terms that appear in a small number of documents overall. IDF is a log normalized value that is calculated by taking the logarithm of the overall term and dividing the number of documents in the corpus by the number of documents containing the term [4].

$$idf = log \frac{total\ number\ of\ document}{number\ of\ documents\ with\ term\ t\ in\ it} \tag{2.2}$$

The formulation of Term Frequency-Inverse Document Frequency (TF-IDF) is as following:

$$W(t,d) = tf(t,d) * log \frac{N}{nt} \tag{2.3}$$

Where, W (t, d) is term weight in document d, t f (t, d) is term frequency in document, N is the total number of document and n $_t$ is the number of documents that have term t.

TF-IDF is one of the simplest and most effective weighting schemes for the data. TF-IDF and its algorithm version are frequently used for text classification because of it is simple formulation and good performance on a variety of different data sets.

CHAPTER 3

ARCHITECTURE OF THE PROPOSED SYSTEM

The primary goal of this chapter is to introduce the main methodology, which forms the basis of this thesis book. This chapter outlines the methodologies for the K-Nearest Neighbors algorithm, along with an overview of the system design, the dataset, preprocessing, feature extraction, and classification, in order to accomplish this goal. This chapter also discusses how to apply the K-Nearest Neighbor algorithm to identify fake job postings and prevent expatriates from falling for the trap.

3.1 Methodologies of the System

There have been numerous studies on clustering or classification using machine learning. The KNN search algorithm is currently being utilized for the detection of fake job postings. According to this approach, the following sections will describe the orientation of KNN algorithms in details.

3.1.1 K-Nearest-Neighbors (KNN)

The K-Nearest-Neighbors (KNN) calculation is a non-parametric order calculation; for instance, it doesn't make any assumptions about the simple dataset. It is recognized for being simple and practical. It is a supervised learning calculation. The information focus is organized into several classes in a marked preparation dataset, allowing the class of the unlabeled information to be predicted.

In classification, numerous attributes determine the class to which the unlabeled information has an area. KNN is typically applied as a classifier. Grouping records in mild of nearest or adjacent preparing models in a given area is applied. This technique is utilized for its straightforwardness of execution and low calculation time. For ceaseless facts, it utilizes the Euclidean distance to compute its nearest neighbors. Another example is when the K closest neighbors are identified and the majority of the neighboring information decides how the new information will be arranged. Although this classifier is simple, the importance of "K" plays a significant role in the ranking of the unlabeled data. There are multiple methods for choosing the qualities for "K" can essentially run the classifier on different occasions with different qualities to see which value produces the best results. Since, each estimation is made when the preparatory information is being organized rather than when it is really experienced in the dataset; the calculation cost is fairly high.

It is a slow learning calculation because, other from storing the preparation data and keeping the dataset overall, little much is done when the dataset is being prepared. The preparation dataset doesn't perform speculation. Therefore, while testing is being done, the entire primary dataset needs to be prepared. KNN forecasts constant properties in relapse. This value represents the average of its K-nearest neighbor's benefits [15].

(i) Tasks of KNN

KNN is a supervised learning classifier. Especially there are two phases in type:

1. Learning Step: the use of the training records a classifier is constructed.
2. Evaluation of the classifier.

As indicated by using the closest neighbor approach, the brand new unlabeled record is organized by way of figuring out which categories its associates have a place with. KNN calculation uses this concept in its computation. In the event of ok-nearest neighbor calculation, a specific really worth of K is constant which allows in ordering the obscure tuple. When a brand new unlabeled tuple is experienced inside the records, KNN plays tasks:

First of all, it breaks down the k focuses nearest to the brand new facts of interest, i.e., the K closest neighbor.

Next, utilizing the neighboring categories, K closet neighbor comes to a decision regarding which class the brand new records ought to be arranged into.

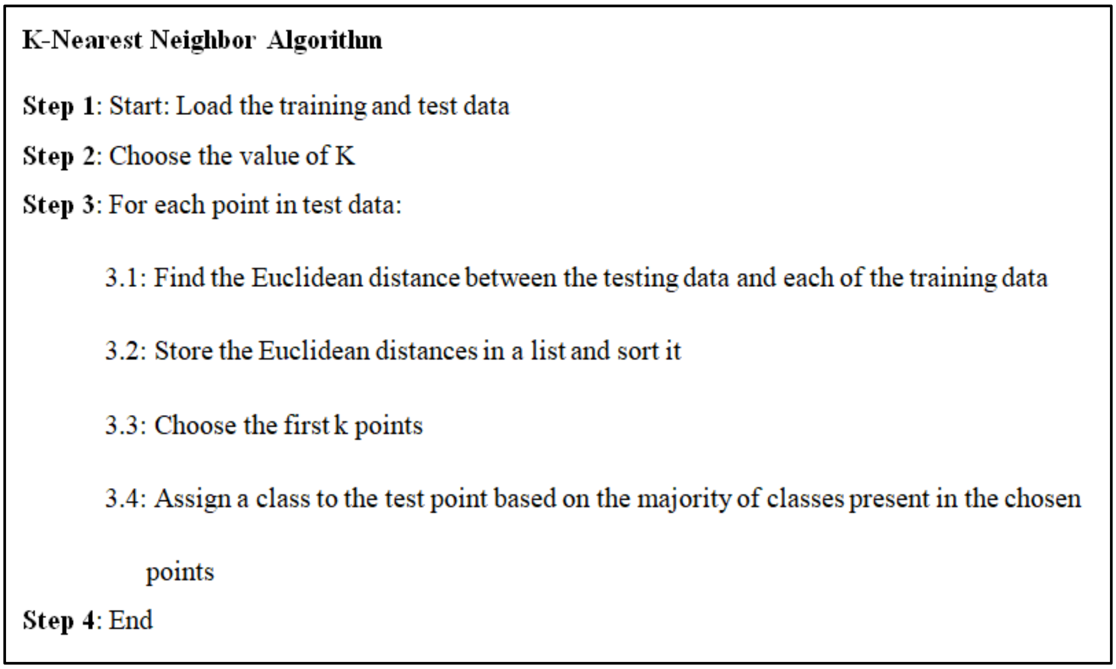
When a little new record is delivered, it characterizes the records in a like manner. It is more treasured in a dataset that is commonly partitioned into corporations and has an area with a selected district of the records plot. Consequently, this calculation receives extra precise, separating the records inputs into numerous classes in a clearer way. K closet neighbor types out the class having the greatest variety of focuses sharing minimal separation from the statistics manual that requirements closer to be ordered. Therefore, the Euclidean distance ought to be decided between the test and the predetermined preparation tests [15].

After K-Nearest Neighbors collect, essentially take most of them to predict the preparation model's class. Euclidean distance, a similarity metric used in KNN, is used to calculate the distance between two points. The distance between two points: X_{1i} and X_{2j} with i element is measured as in Equation 3.1:

$$dist(X1, X2) = \sqrt{\sum_{i,j=1}^{n}(X1i - X2j)} \tag{3.1}$$

The K nearest neighbors to the testing data are then chosen, and the majority class or label of the selected neighbors will become the predicted class or label for unknown testing data after determining the distance between the testing data and the training data using the equation 3.1.

With the help of the following steps, Figure 3.1 demonstrates how the KNN algorithm searches items using the closet training dataset:



```
K-Nearest Neighbor Algorithm
Step 1: Start: Load the training and test data
Step 2: Choose the value of K
Step 3: For each point in test data:
        3.1: Find the Euclidean distance between the testing data and each of the training data
        3.2: Store the Euclidean distances in a list and sort it
        3.3: Choose the first k points
        3.4: Assign a class to the test point based on the majority of classes present in the chosen
             points
Step 4: End
```

Figure 3.1 K-Nearest Neighbor Algorithm

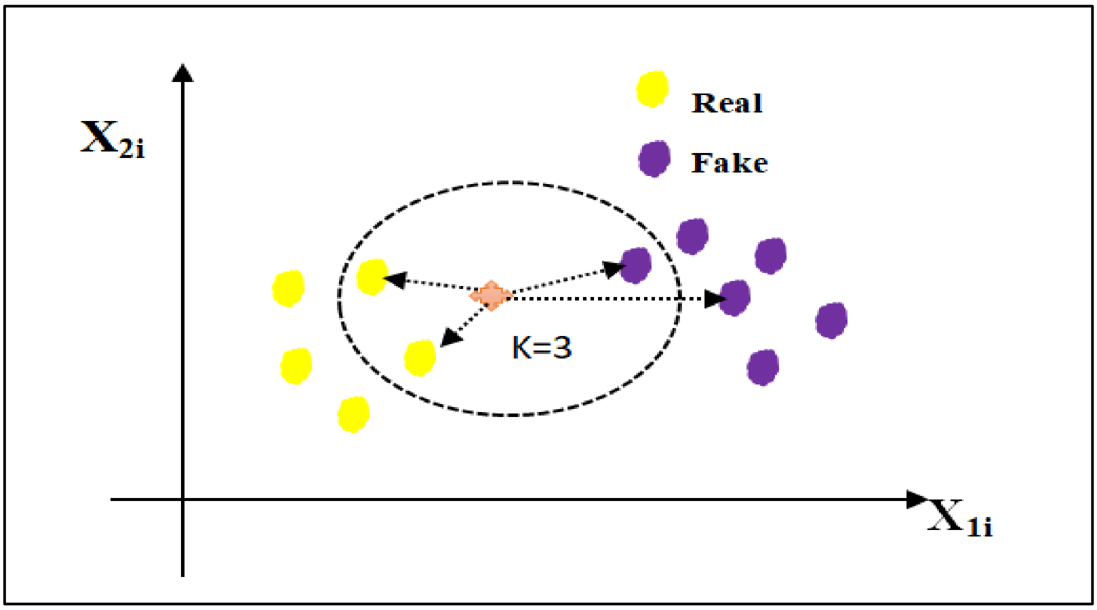


Figure 3.2 Sample Classification of KNN

In the wake of setting away the instruction, set all barriers ought to be standardized, with the intention that the estimations turn out to be extra straightforward. The outcome of the grouping is sensitive to the really worth of 'K'. The number of neighbors that should be taken into consideration is determined by the data variable "K." The value of "K" has an impact on the calculation because it allows us to construct the bounds of each class.

3.2 Proposed System Design

The overview design of the proposed system is shown in Figure 3.3. In this proposed system, there are two phases: training and testing. In both phases, two main stages are essential. In the first step, text preprocessing is performed for tokenizing, remove stop words and stemming. In the second step, the weights of words are calculated using TF-IDF features extraction. In the testing phase of the classification step, the system is used K-Nearest Neighbor classifier (KNN) to identify the class category. In the final step, the performance evaluation is measured by using confusion matrix: precision, recall and f- measure.

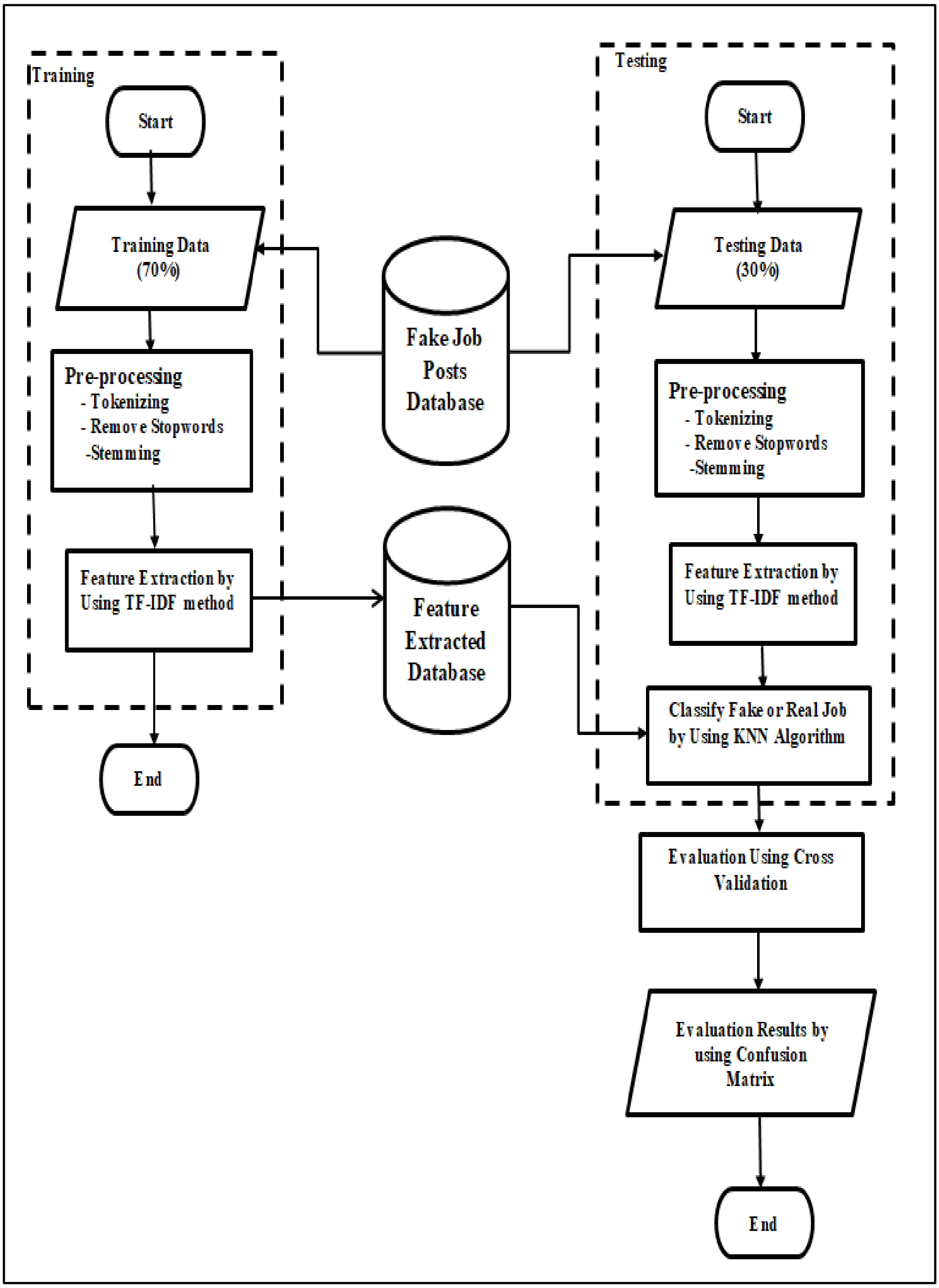

Training
Start
Training Data (70%)
Pre-processing
- Tokenizing
- Remove Stopwords
-Stemming
Feature Extraction by Using TF-IDF method
End
Fake Job Posts Database
Feature Extracted Database
Testing
Start
Testing Data (30%)
Pre-processing
- Tokenizing
- Remove Stopwords
-Stemming
Feature Extraction by Using TF-IDF method
Classify Fake or Real Job by Using KNN Algorithm
Evaluation Using Cross Validation
Evaluation Results by using Confusion Matrix
End

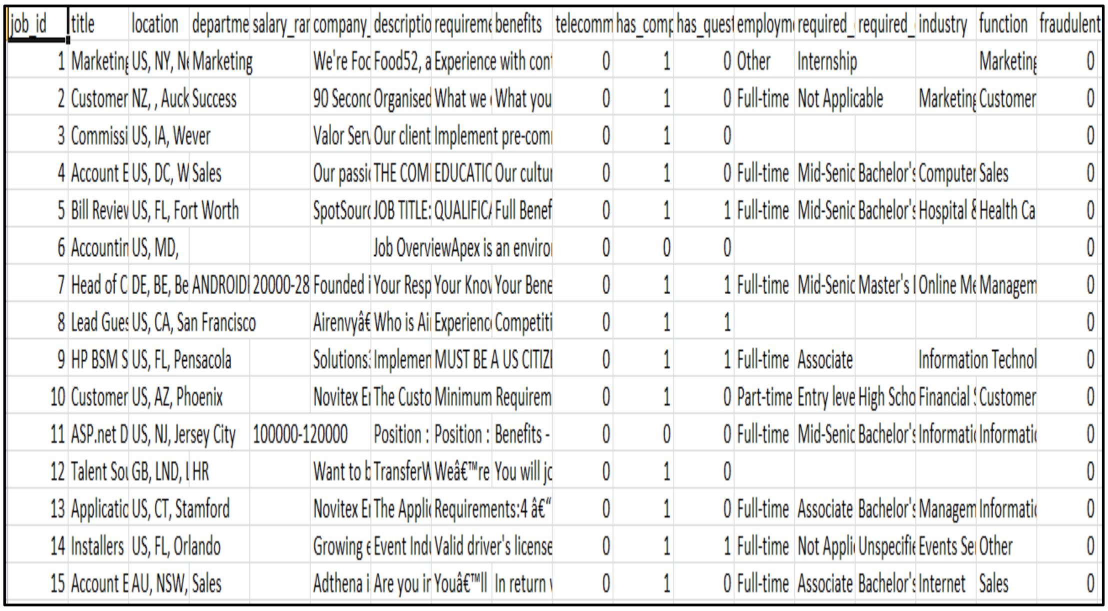
Figure 3.3 Proposed System Design

3.3 Sample Case Study of the Proposed System

The primary goal of this system is to construct a classifier for detecting fake job postings using the feature extraction (TF-IDF) method and the K-Nearest Neighbor (KNN) methodology. As can be seen below, the effectiveness of the KNN technique will be determined using fake or actual job posting prediction datasets.

3.3.1 Datasets

The dataset name is Real or Fake Job Posting Prediction from the website Kaggle [10]. This dataset has seventeen attributes. These are “job id”, “title”, “location”, “department”, “salary range”, “company profile”, “description”, “benefits”, and “telecommuting”, “has the company logo”, “has questions”, “employment type”, “required experience”, “required education”, “industry” and “function”. The label is binary for the particular domain of the problem; the real is "0" and the fake is "1". The attribute types are Boolean and Text. Boolean attributes are salary range, and telecommuting, has the company logo, has questions. Some attributes of the description are the same as the text (e.g., department, required experience, required education, industry, and function). Therefore, there are only seven attributes used. Figure 3.4 is shown the sample real or fake jobs prediction datasets.



job_id	title	location	departme	salary_rar	company_	descriptio	requireme	benefits	telecomm	has_comp	has_quest	employm	required_	required_	industry	function	fraudulent
1	Marketing	US, NY, N	Marketing		We're Foc	Food52, a	Experience with con		0	1	0	Other	Internship			Marketing	0
2	Customer	NZ, , Auck	Success		90 Second	Organised	What we	What you	0	1	0	Full-time	Not Applicable		Marketing	Customer	0
3	Commissi	US, IA, Wever			Valor Serv	Our client	Implement pre-com		0	1	0						0
4	Account E	US, DC, W	Sales		Our passi	THE COM	EDUCATIC	Our cultur	0	1	0	Full-time	Mid-Senic	Bachelor's	Computer	Sales	0
5	Bill Reviev	US, FL, Fort Worth			SpotSourc	JOB TITLE:	QUALIFICA	Full Benef	0	1	1	Full-time	Mid-Senic	Bachelor's	Hospital &	Health Ca	0
6	Accountin	US, MD,				Job Overview	Apex is an enviro		0	0	0						0
7	Head of C	DE, BE, Be	ANDROIDI	20000-28	Founded i	Your Resp	Your Knov	Your Bene	0	1	1	Full-time	Mid-Senic	Master's I	Online Me	Managem	0
8	Lead Gues	US, CA, San Francisco			Airenvyâ€	Who is Ai	Experienc	Competiti	0	1	1						0
9	HP BSM S	US, FL, Pensacola			Solutions:	Implemen	MUST BE A US CITIZI		0	1	1	Full-time	Associate		Information Technol		0
10	Customer	US, AZ, Phoenix			Novitex Er	The Custo	Minimum Requirem		0	1	0	Part-time	Entry leve	High Scho	Financial S	Customer	0
11	ASP.net D	US, NJ, Jersey City		100000-120000		Position :	Position :	Benefits -	0	0	0	Full-time	Mid-Senic	Bachelor's	Informatic	Informatic	0
12	Talent Sou	GB, LND, l	HR		Want to b	TransferW	Weâ€™re	You will jc	0	1	0						0
13	Applicatic	US, CT, Stamford			Novitex Er	The Applic	Requirements:4 â€“		0	1	0	Full-time	Associate	Bachelor's	Managem	Informatic	0
14	Installers	US, FL, Orlando			Growing e	Event Indu	Valid driver's license		0	1	1	Full-time	Not Appli	Unspecific	Events Se	Other	0
15	Account E	AU, NSW,	Sales		Adthena i	Are you ir	Youâ€™ll	In return	0	1	0	Full-time	Associate	Bachelor's	Internet	Sales	0

Figure 3.4 Sample real or fake jobs prediction datasets.

There are seventeen job posts attributes in this sample dataset. The descriptions of the each attributes are as shown in Table 3.1.

Table 3.1 Contents of the dataset's description

Number	Attribute	Datatype	Description
1	job_id	int	Identification number given to each job posting
2	title	text	A name that describes the position or job
3	location	text	Information about where the job is located
4	department	text	Information about the department this job is offered by
5	salary_range	text	Expected salary range
6	company_profile	text	Information about the company
7	description	text	A brief description about the position offered
8	requirements	text	Pre-requisites to qualify for the job
9	benefits	text	Benefits provided by the job
10	telecommuting	boolean	Is work from home or remote work allowed
11	has_company_logo	boolean	Does the job posting have a company logo
12	has_questions	boolean	Does the job posting have any questions
13	employment_type	text	5 categories- Full-time, part-time, contract, temporary and other
14	required_experience	text	"Can be-Internship, Entry level, Associate, Mid-senior level,
15	required_education	text	Director, Executive or Not Applicable"
16	industry	text	Can be-Bachelor's degree, high school degree, unspecified, associate degree, master's degree, certification, some college coursework, professional, some high school coursework, vocational
17	function	text	The industry the job posting is relevant to
18	fraudulent	boolean	The umbrella term to determining a job's functionality

3.2 Preprocessing

Preprocessing data is a phase that includes text cleaning and text conversion into a format that is suitable for the classification method. Text is unstructured data that is separated into useful and useless data. So, preprocessing is necessary to remove noise from text [2] .Text preprocessing are tokenization, removing stop words and stemming.

(i) Tokenization

Tokenization is the simplest preprocessing step that needs to perform on text. This approach, which can be broadly characterized as the process of dividing a stream of text into more smaller chunks, sets the level of granularity at which evaluate and produce textual data (historically called tokens). Most natural language processing models employed words as their preferred atomic unit until recently, but more recent methods have decomposed text into smaller units.

Tokenization is a complicated subject in and of itself, and being usually taken for granted. Conventional techniques are rule-based and can be as simple as splitting words with a space, using punctuation, and using contractions. Clearly, considerably more advanced knowledge-based techniques have been developed that are nevertheless grounded in language ideas. However, recent developments have greatly favored data-driven tokenizes, which produce better results but in which tokens no longer match the conventional concept of a typographic unit. In other words, the act of separating a sentence into components that, crucially, do not have a linguistic motivation or explanation is referred to as "tokenization" in modern usage. Conventional techniques are now frequently referred to as "pre-tokenizes" because some of these methods still require classic tokenizes as an initial step. It is more typical to associate conventional methods with pre-tokenizes because tokenization and classification approaches were created simultaneously [3]. It is the process of dividing a string of characters into smaller units, such as words, phrases, symbols, and other so-called tokens. Tokens can be single words, short phrases, or even complete sentences. In this step, digits are removed or filtered. Examples of digits include dates, numbers, times, and other regular expressions and others. All letters are changed to either lowercase letters or uppercase letters.

(ii) Remove Stop Words

The words that are commonly filtered out earlier than natural language processing are called stop words. Those are sincerely the maximum common phrases in any language (such as prepositions, pronouns, articles, conjunctions, etc.), and do not significantly add to the text's information. Some examples of stop words in English are "the," "a", "an", "so", "what". In any language used by people, stop words are common. By eliminating these words, The low-level information is removed from our text by this approach to make the key information more prominent. In other words, for our purposes, the removal of such statements has no negative effects on the model train.

The removal of stop words obviously shrinks the dataset, which in turn shortens training time because there are fewer tokens to process [10].

(ii) Stemming

Stemming is a preprocessing step in text mining applications as well as totally not unusual want in natural language processing duties. In fact, it is very crucial in maximum of the facts retrieval structures. The principle goal of stemming is to reduce distinct grammatical forms or word forms of a word, such as its adjective, verb, adverb, noun, etc., to its root form. The aim of stemming is to reduce a word's inflectional paperwork and from time to time derivationally related styles of a word to a common base form.

There are mainly two errors in stemming – over stemming and under stemming. When two words with different stems are stemmed to the same root, this is referred to as over-stemming. This is also known as a false positive. When two words that should have the same root are not, this is known as under-stemming. A false negative is another name for this. Heavy stemmers, on the other hand, increase the over-stemming errors while decreasing the under-stemming errors [5]. The two categories of stemming algorithms are Porter Stemmer and Snowball Stemmer.

(i) Porter stemmer: This stemming algorithm is more established. Its major goal is to resolve words to their common forms by deleting the common ends from words. It's not overly complicated, and work on it has been suspended. It's usually a fine basic stemmer to start with, but it's not really advised to use it for any production or complicated applications. Instead, it serves as a pleasant, straightforward stemming procedure in research that can guarantee reproducibility. It also is a fairly mild method when compared to other stemming algorithm.

(ii) Snowball stemmer: This algorithm is also called the Porter2 stemming algorithm. It is virtually commonly known that it is superior to the Porter stemmer, and even the person who invented the Porter stemmer concedes as much. In addition, it is also more forceful than the Porter stemmer. Many of the features that were introduced to the Snowball stemmer were a result of problems with the Porter stemmer. In this proposed system, this algorithm is applied.

This system tests eight job posts from company profile attribute. Firstly, this approach carries out the tokenization; remove stop words and stemming process. And then, words are extracted from each job post. Example jobs are presented in Table 3.2.

Table 3.2 Examples of Training Data

Name	**Company Profile**	**Class**
Instance 1	Growing event production company providing staging, scenic, and drapery primarily in the state of Florida. We have a secondary location and will be adding a third location in Southeast Florida. We are a small team passionate about creating high quality events and providing excellent customer service, both on show and in the office. Â	Real
Instance 2	Super Soccer Stars is the country's most popular soccer development program for kids. For over a decade, we have provided outstanding instruction for thousands of children in 400+ locations in NY, NJ, CT, MA, CA, FL, IL, Washington,	Real

Name	Company Profile	Class
	DC, and London, UK! Super Soccer Stars was founded in 2000, and since its inception, it has been providing outstanding soccer development instruction for children aged 2 and up.	
Instance 3	Fab is the place to discover the most exciting things for your life. Our modern, urban-inspired products allow everyone to design their lives and express their personal sense of style. Always unique, well-designed, and of the highest quality. Fab. Smiles. Guaranteed.	Fake
Instance 4	UAH is a multi-divisional, full-service private investment firm. With 41 locally managed offices around the nation, we offer private equity funds, residential and commercial clients individualized real estate development, management, and investment services. UAH is an employer that values diversity.	Fake

Step1: Data Cleaning: tokenizing and remove stop words

Name	Company Profile	Class
Instance 1	growing event production company providing staging scenic drapery primarily state florida secondary location adding third location southeast florida small team passionate creating high quality events providing excellent customer service show office	Real
Instance 2	super soccer stars country popular soccer development program kids decade provided outstanding instruction thousands children locations washington london super soccer stars founded since inception providing outstanding soccer development instruction children aged	Real
Instance 3	fab place discover exciting things life modern inspired	Fake

Name	Company Profile	Class
	products allow everyone design lives express personal sense style always unique well designed highest quality fab smiles guaranteed	
Instance 4	uah multi divisional full service private investment firm locally managed offices around nation offer private equity funds residential commercial clients individualized real estate development management investment services uah employer values diversity	Fake

Step 2 After Stemming

Name	Company Profile	Class
Instance 1	grow event product compani provid stage scenic draperi primarili state florida secondari locat ad third locat southeast florida small team passion creat high qualiti event provid excel custom servic show offic	Real
Instance 2	super soccer star countri popular soccer develop program kid decad provid outstand instruct thousand children locat washington london super soccer star found sinc incept provid outstand soccer develop instruct children age	Real
Instance 3	fab place discov excit thing life modern inspir product allow everyon design live express person sens style alway uniqu well design highest qualiti fab smile guarante	Fake
Instance 4	uah multi division full servic privat invest firm local manag offic around nation offer privat equiti fund residenti commerci client individu real estat develop manag invest servic uah employ valu divers	Fake

3.3.3 Feature Extraction

Feature extraction is the process of turning unprocessed raw data into numerical features that can be used for processing while preserving the original data sets content. It produces better outcomes than using machine learning directly to the raw data [11]. Feature extraction can be performed either manually or automatically:

- Identification and description of the characteristics that are relevant for a particular situation are necessary for manual feature extraction, as is the implementation of a method to extract such features. Knowing the background or domain well can often aid in making informed decisions about which features can be beneficial. Engineers and scientists have created feature extraction techniques for images, signals, and text through many years of research. The mean of a signal's window is a prime example of a simple feature.
- Automated feature extraction eliminates the need for human intervention by automatically extracting features from signals or images using specialized algorithms or deep networks. When attempting to transition quickly from collecting raw data to creating machine learning algorithms, this strategy can be quite helpful. An example of automated feature extraction is wavelet scattering.

Step 3: Calculate TF-IDF for Examples of Instance 1 and 3 of Training Data

According to the equation 2.3, Instance 1 and 3 of training data is calculated with TF-IDF for weight.

TF-IDF for Instance 1

grow=1 /event=2/ product=1/ compani=1/ provid=2/ stage=1/ scenic=1/ draperi=1/ primarili =1/ state=1/ florida=2/ secondari=1/ locat=2 ad=1/ third=1/ southeast=1/ small=1/ team=1/ passion=1/ creat=1/ high=1/ quality=1/ excel=1/ custom=1/ service=1/ show=1/ office=1/

Total number of documents = 8

Total number of term in the document = 57

1.	TF-IDF(“growing”)=1/57*log(8/1) = 0.0158	14.	TF-IDF(“ad”)=1/57*log(8/1) = 0.0158
2.	TF-IDF(“event”)=2/57*log(8/1) = 0.0317	15.	TF-IDF(“third”)=1/57*log(8/1) =0.0158
3.	TF-IDF(“product”)=1/57*log(8/2) = 0.011	16.	TF-IDF(“southeast”)=1/57*log (8/1) =0.0158
4.	TF-IDF(“compani”)=1/57*log(8/3) = 0.0075	17.	TF-IDF(“small”)=1/57*log(8/1) =0.0158
5.	TF-IDF(“provid”)=2/57*log(8/5) = 0.0072	18.	TF-IDF(“team”)=1/57*log(8/1) =0.0158
6.	TF-IDF(“stage”)=1/57*log(8/1) = 0.0158	19.	TF-IDF(“passion”)=1/57*log(8/2) =0.0106
7.	TF-IDF(“scenic”)=1/57*log(8/1) = 0.0158	20.	TF-IDF(“creat”)=1/57*log(8/1) =0.0158
8.	TF-IDF(“draperi”)=1/57*log(8/1) = 0.0158	21.	TF-IDF(“high”)=1/57*log(8/3) =0.0075
9.	TF-IDF(“primarili”)=1/57*log(8/1) = 0.0158	22.	TF-IDF(“quality”)=1/57*log(8/3) =0.0075
10.	TF-IDF(“state”)=1/57*log(8/1) = 0.0158	23.	TF-IDF(“excel”)=1/57*log(8/1) =0.0158
11.	TF-IDF(“florida”)=2/57*log(8/1) = 0.0317	24.	TF-IDF(“custom”)=1/57*log(8/2) =0.0106
12.	TF-IDF(“secondari”)=1/57*log (8/1)= 0.0158	25.	TF-IDF(“service”)=1/57*log(8/4) =0.0053
13.	TF-IDF(“locat”)=2/57*log(8/2) = 0.0211	26.	TF-IDF(“show”)=1/57*log(8/1) =0.0158

TF-IDF for Instance 3

fab=2/ place=1/ discov=1/ excit=1/ life=1/ modern=1/ inspire=1/ product=1/ allow=1/ design=2/ live=1/ express=1/ person=1/style=1/ high=1/ quality=1/ guarantee=1/

Total number of documents = 8

Total number of term in the document = 41

1.	TF-IDF(“fab”) =2/41*log (8/1) =0.0441	9.	TF-IDF(“allow”)=1/41*log(8/1) = 0.02202
2.	TF-IDF(“place”) =1/41*log (8/3) =0.0104	10.	TF-IDF(“design”)=2/41*log(8/1) =0.0441
3.	TF-IDF(“discov”)=1/41*log(8/1) =0.02202	11.	TF-IDF(“live”)=1/41*log(8/1) = 0.02202
4.	TF-IDF(“excit”)=1/41*log(8/1) = 0.02202	12.	TF-IDF(“express”)=1/41*log(8/1) = 0.02202
5.	TF-IDF(“life”)=1/41*log(8/1) = 0.02202	13.	TF-IDF(“person”)=1/41*log(8/2) =0.0147
6.	TF-IDF(“modern”)=1/41*log(8/1) = 0.02202	14.	TF-IDF(“style”)=1/41*log(8/1) = 0.02202
7.	TF-IDF(“inspire”)=1/41*log(8/1) = 0.02202	15.	TF-IDF(“high”)=1/41*log(8/1) =0.02202
8.	TF-IDF(“product”)=1/41*log(8/3) = 0.0104	16.	TF-IDF(“quality”)=1/41*log(8/2) =0.0147

Step 4: To find the nearest neighbor, the distance calculate between the testing job and training jobs by using **Euclidean's Distance** of equation 3.1. Table 3.3 and Table 3.4 are shown as the calculation of K-Nearest Neighbor with Euclidean distance.

Table 3.3 Distance between training jobs and testing job

Words	**Training real job 1**	**Training real job 2**	**Training real job 3**	**Training real job 4**	**Training fake job 5**	**Training fake job 6**	**Training fake job 7**	**Training fake job 8**
provide	0.0017	0.0253	0.0061	0.0113	0.0055	0.0654	0.0088	0.0052
full	0.0012	0.01703	0.0024	0.0065	0.0017	0.0033	0.003	0.0018
time	0.0057	0.0116	0.0018	0.0032	0.0022	0.0024	0.0022	0.0029
permanent	0.0057	0.0116	0.0018	0.0032	0.0022	0.0024	0.0022	0.0029
position	0.0057	0.0116	0.0018	0.0032	0.0022	0.0024	0.0022	0.0029
medium	0.0057	0.0116	0.0018	0.0032	0.0022	0.0024	0.0022	0.0029
large	0.0057	0.0116	0.0018	0.0032	0.0022	0.0024	0.0022	0.0029
company	0.0012	0.01703	0.0024	0.0065	0.0017	0.0033	0.003	0.0018
interested	0.0057	0.0116	0.0018	0.0032	0.0022	0.0024	0.0022	0.0029
finding	0.0057	0.0116	0.0018	0.0032	0.0022	0.0024	0.0022	0.0029
recruiting	0.0057	0.0116	0.0018	0.0032	0.0022	0.0024	0.0022	0.0029
high	0.0076	0.0094	0.0042	0.0014	0.0073	0.0255	0.0053	0.0087
quality	0.0012	0.01703	0.0024	0.0065	0.0017	0.0033	0.003	0.0018
candidates	0.0057	0.0116	0.0018	0.0032	0.0022	0.0024	0.0022	0.0029
engineering	0.0057	0.0116	0.0018	0.0032	0.0022	0.0024	0.0022	0.0029
manufacturing	0.0057	0.0116	0.0018	0.0032	0.0022	0.0024	0.0022	0.0029
technical	0.0472	0.0139	0.0179	0.0019	0.0305	0.0199	0.12	0.0328
job	0.0057	0.0116	0.0018	0.0032	0.0022	0.0024	0.0022	0.0029

Let k = 3

Jobs	Distance	Rank	Neighbor	Class
Real job 1	0.3585	5	No	Real
Real job 2	0.5688	8	No	Real
Real job 3	0.2583	1	Yes	Real
Real job 4	0.2693	2	Yes	Real
Real job 5	0.2735	3	Yes	Fake
Real job 6	0.3867	6	No	Fake
Real job 7	0.4113	7	No	Fake
Real job 8	0.2946	4	No	Fake

Table 3.4 The category of the nearest neighbors

Use sample majority of the category of the nearest neighbors as the prediction values of the test job.

Two instances get two neighbors are "**Real**" class and one neighbor is "**Fake**" class. So, it concludes the test job is "**Real**" jobs.

3.4 Evaluation Performance

To evaluate the efficiency of the system, three evaluation performances of precision, recall, and f-measure—are used. These are Equation (3.1) (3.2) (3.3).

$$Precision = \frac{TP}{TP+FN} \tag{3.1}$$

$$Recall = \frac{TP}{TP+FP} \tag{3.2}$$

$$F - measure = \frac{2*Precision*Recall}{Precision+Recall} \tag{3.3}$$

Where:

- True Positive (TP): The true positive value occurs when the actual value and predicted value are same.
- False Positive (FP): the false positive value for a class will be the sum of values in appropriate column with expectation for TP.
- True Negative (TN): the true negative value for a class will be the total of all columns and rows excluding those for the class that is determining the values.
- False Negative (FN): the false negative value for a class will be the sum of values of corresponding rows except for TP.

Precision is the ability to correctly classify a group. The recall measures how well the classification performed in classifying the current class in the dataset. As a result, the recall is the completeness of the categorization model, and the precision is the exactitude. The accuracy and completeness of the system are being increased by the f-measure.

In this fake job posts detection system, four experiments are performed. Four separate training dataset and testing dataset combinations are used for each analysis (Test 1: 210 records of training data and 90 records of testing data; Test 2 uses 350 training records and 150 testing records; Test 3 uses 700 training records and 300 testing records; Test 4: 1306 records for training and 560 records for testing). This system bases its evaluation of the experiment's performance on the precision, recall, and f-measure of each analysis.

3.4.1 Cross Validation

A resampling technique called cross-validation is used to assess machine learning models on a limited data sample. The process contains a single parameter, k, that designates how many groups should be created from a given data sample. As a result, the process is frequently referred to as k-fold cross-validation.

The original data are randomly separated into k folds or subsets that are mutually exclusive and about equal in size. This process is known as k-fold cross-validation. A learning algorithm is created each time utilizing the remaining k-1 folds and tested on one of the k-folds. This procedure is carried out k times. The total number of correctly identified examples over all k iterations is divided by the total number of examples in the training set to arrive at the cross-validation estimate of accuracy.

This proposed system is implemented with 9-fold cross validation for 300 datasets are used as training data.

CHAPTER 4

IMPLEMENTATION AND EXPERIMENTAL RESULTS

This chapter serves as a demonstration of the system's implementation and an examination of its performance using various metrics. The first section of this chapter provides user interface illustrations for each stage of system development. The K-Nearest Neighbors technique evaluation findings are in the second section.

4.1 System Implementation

The proposed system employs the JAVA programming language to implement the fake job post detecting system. K-Nearest Neighbor and TF-IDF will be used to develop the classification analysis. The main graphical user interface (GUI) of the system includes four important buttons: Data, Preprocessing, Feature Extraction, and Classification and Evaluation as depicted in Figure 4.1.

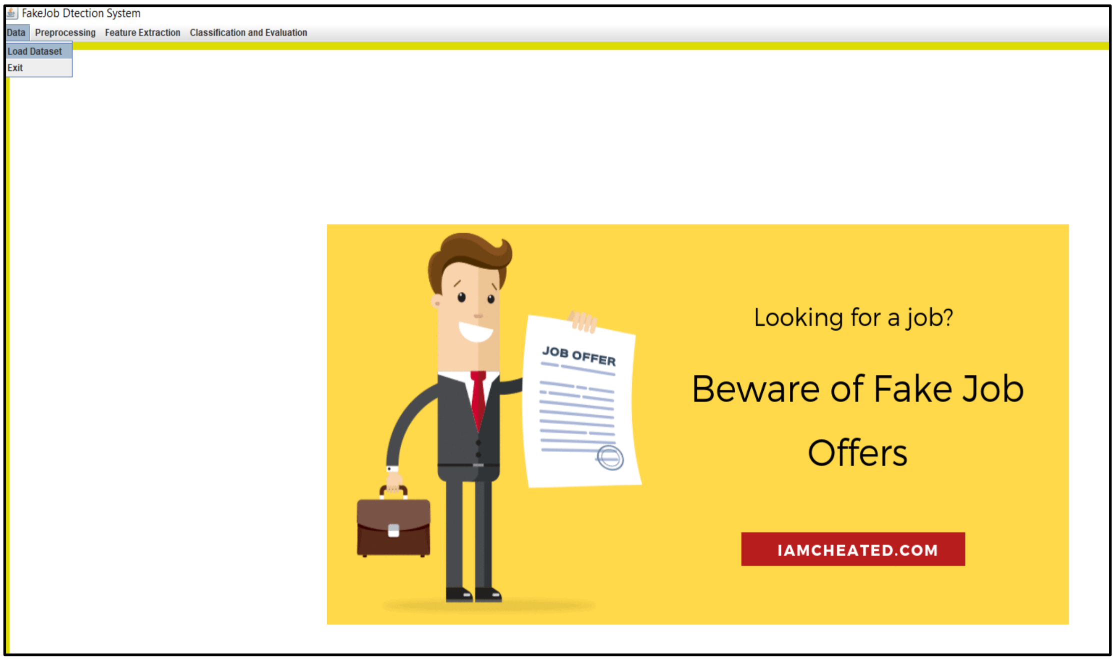
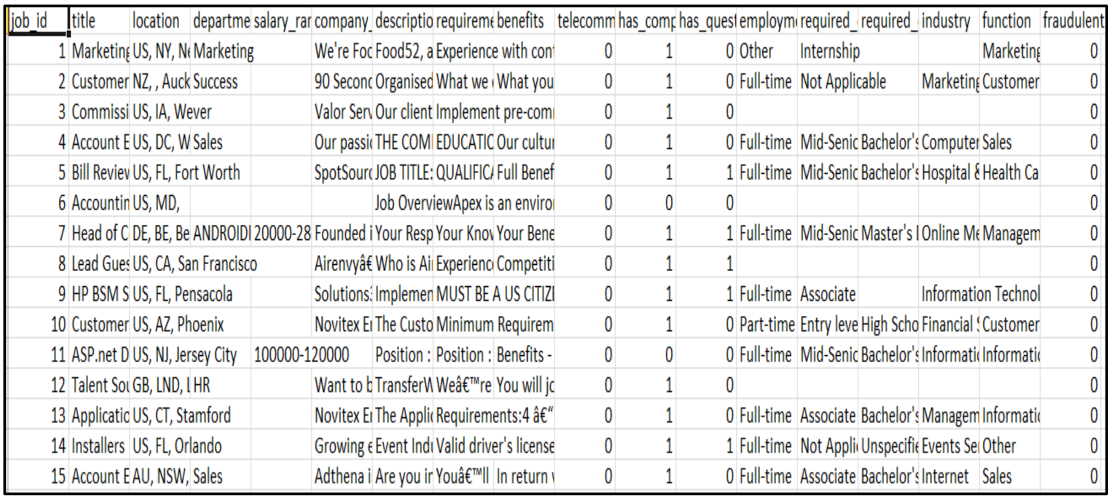


Figure 4.1 Main Page of the System

Preprocessing is important before removing elements. Text is unstructured data that is separated into useful and useless data. So, preprocessing is necessary to remove noise from text. Text preprocessing are tokenization, removing stop word, and stemming. TF-IDF is utilized in text mining methods to calculate the weight of each term for Feature Extraction.

4.1.1 Load Training Data

The homepage of the load dataset menu can be summarized as shown in Figure 4.2 after selecting Load Dataset.

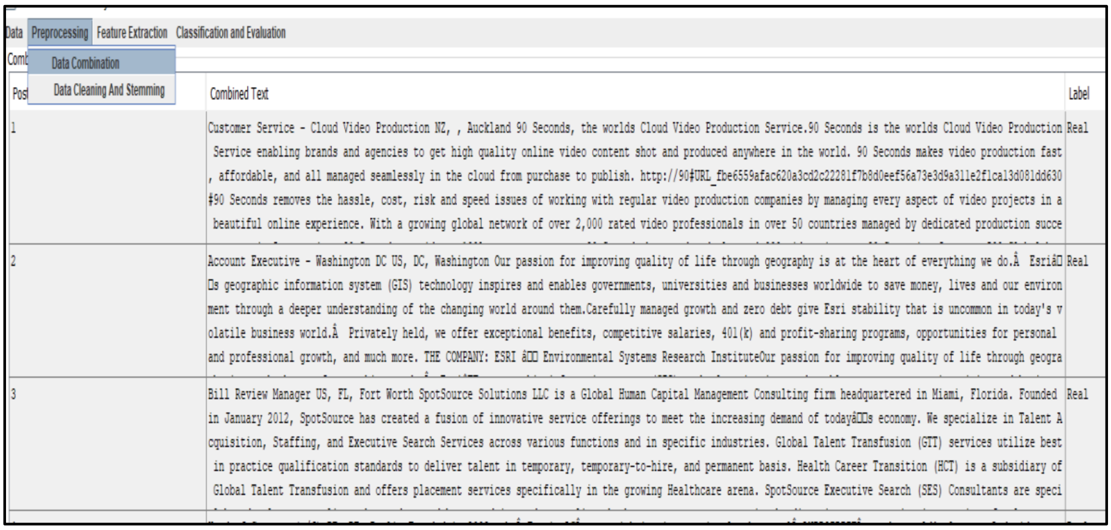


job_id	title	location	departme	salary_rar	company_	descriptio	requireme	benefits	telecomm	has_comp	has_quest	employme	required_	required_	industry	function	fraudulent
1	Marketing	US, NY, Ne	Marketing		We're Foc	Food52, a	Experience with con		0	1	0	Other	Internship			Marketing	0
2	Customer	NZ, , Auck	Success		90 Seconc	Organised	What we	What you	0	1	0	Full-time	Not Applicable		Marketing	Customer	0
3	Commissi	US, IA, Wever			Valor Serv	Our client	Implement pre-comi		0	1	0						0
4	Account E	US, DC, W	Sales		Our passi	THE COMI	EDUCATIC	Our cultui	0	1	0	Full-time	Mid-Senic	Bachelor's	Computer	Sales	0
5	Bill Reviev	US, FL, Fort Worth			SpotSourc	JOB TITLE:	QUALIFIC/	Full Benef	0	1	1	Full-time	Mid-Senic	Bachelor's	Hospital &	Health Ca	0
6	Accountin	US, MD,				Job OverviewApex is an enviroi			0	0	0						0
7	Head of C	DE, BE, Be	ANDROIDI	20000-28	Founded i	Your Resp	Your Knov	Your Bene	0	1	1	Full-time	Mid-Senic	Master's I	Online Me	Managem	0
8	Lead Gues	US, CA, San Francisco			Airenvyâ€	Who is Ai	Experienc	Competiti	0	1	1						0
9	HP BSM S	US, FL, Pensacola			Solutions:	Implemen	MUST BE A US CITIZI		0	1	1	Full-time	Associate		Information Technol		0
10	Customer	US, AZ, Phoenix			Novitex Ei	The Custo	Minimum Requirem		0	1	0	Part-time	Entry leve	High Scho	Financial S	Customer	0
11	ASP.net D	US, NJ, Jersey City		100000-120000		Position :	Position :	Benefits -	0	0	0	Full-time	Mid-Senic	Bachelor's	Informatic	Informatic	0
12	Talent Sou	GB, LND, L	HR		Want to b	TransferW	Weâ€™re	You will jc	0	1	0						0
13	Applicatic	US, CT, Stamford			Novitex Ei	The Applic	Requirements:4 â€“		0	1	0	Full-time	Associate	Bachelor's	Managem	Informatic	0
14	Installers	US, FL, Orlando			Growing e	Event Indu	Valid driver's license		0	1	1	Full-time	Not Appli	Unspecifie	Events Se	Other	0
15	Account E	AU, NSW,	Sales		Adthena i	Are you ir	Youâ€™ll	In return v	0	1	0	Full-time	Associate	Bachelor's	Internet	Sales	0

Figure 4.2 Load Training Data

4.1.2 Data Combination

The "Data Combination" step of the training process is depicted in Figure 4.3. In this step, there are title, location, company profile, description, requirements, benefits and employment type of texts are combined with one column.



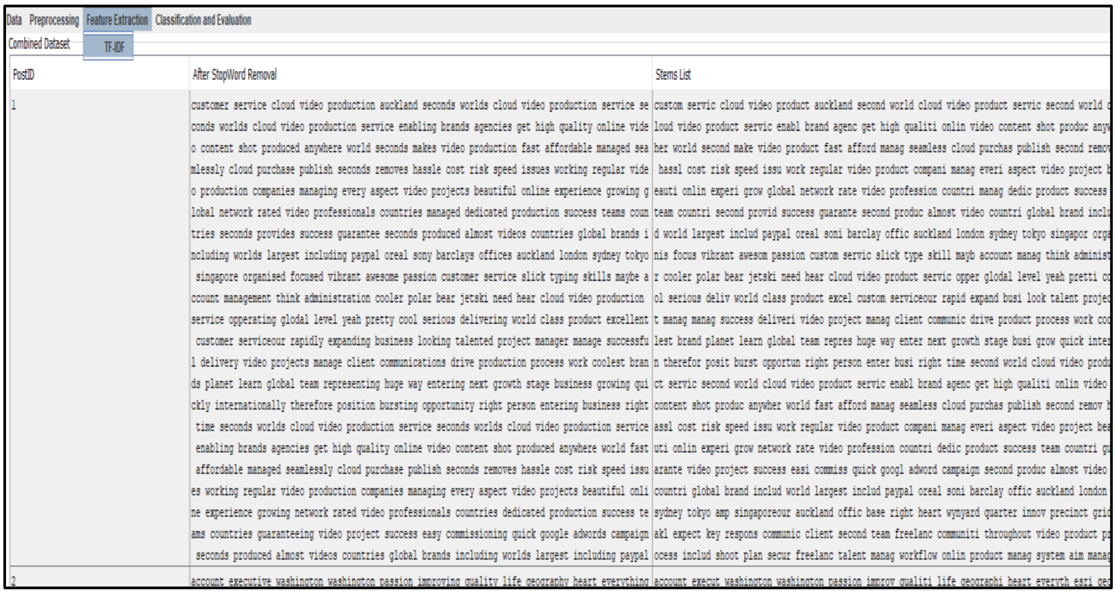
Post	Combined Text	Label
1	Customer Service - Cloud Video Production NZ, , Auckland 90 Seconds, the worlds Cloud Video Production Service.90 Seconds is the worlds Cloud Video Production Service enabling brands and agencies to get high quality online video content shot and produced anywhere in the world. 90 Seconds makes video production fast, affordable, and all managed seamlessly in the cloud from purchase to publish. http://90#URL_fbe6559afac620a3cd2c22281f7b8d0eef56a73e3d9a311e2f1ca13d081dd630#90 Seconds removes the hassle, cost, risk and speed issues of working with regular video production companies by managing every aspect of video projects in a beautiful online experience. With a growing global network of over 2,000 rated video professionals in over 50 countries managed by dedicated production succe	Real
2	Account Executive - Washington DC US, DC, Washington Our passion for improving quality of life through geography is at the heart of everything we do.Â Esriâ□□s geographic information system (GIS) technology inspires and enables governments, universities and businesses worldwide to save money, lives and our environment through a deeper understanding of the changing world around them.Carefully managed growth and zero debt give Esri stability that is uncommon in today's volatile business world.Â Privately held, we offer exceptional benefits, competitive salaries, 401(k) and profit-sharing programs, opportunities for personal and professional growth, and much more. THE COMPANY: ESRI â□□ Environmental Systems Research InstituteOur passion for improving quality of life through geogra	Real
3	Bill Review Manager US, FL, Fort Worth SpotSource Solutions LLC is a Global Human Capital Management Consulting firm headquartered in Miami, Florida. Founded in January 2012, SpotSource has created a fusion of innovative service offerings to meet the increasing demand of todayâ□□s economy. We specialize in Talent Acquisition, Staffing, and Executive Search Services across various functions and in specific industries. Global Talent Transfusion (GTT) services utilize best in practice qualification standards to deliver talent in temporary, temporary-to-hire, and permanent basis. Health Career Transition (HCT) is a subsidiary of Global Talent Transfusion and offers placement services specifically in the growing Healthcare arena. SpotSource Executive Search (SES) Consultants are speci	Real

Figure 4.3 Preprocessing step of Data Combination

4.1.3 Data Cleaning and Stemming

Figure 4.4 shows the "Data Cleaning and Stemming" step of the training process. This step is very important for preprocessing. In Data Cleaning, Tokenization and remove stop words process are contained. To begin with, a detailed explanation of the tokenization process is necessary. Tokenization is the process of dividing a string of characters into tokens, which can include words, phrases, keywords, symbols, and other objects. Tokens can be single words, short phrases, or even complete sentences. In this step, digits are removed. Examples of digits include dates, numbers, times, and other regular expressions and others. All letters are changed to either lowercase letters or uppercase letters. Secondly, remove stop words are the method of removing unimportant or unnecessary words that is not pointed any contents. The stop words that point to natural are prepositions, articles, some pronouns, and conjunctions. A few instances of stop words in English are "the", "a", "an", "so", "what", etc.

In Stemming, It reduces the word to its word stem means affixes to suffixes and prefixes, or the roots of words. In this proposed system, stemming process is stemmed by using snowball stemmer of rules.

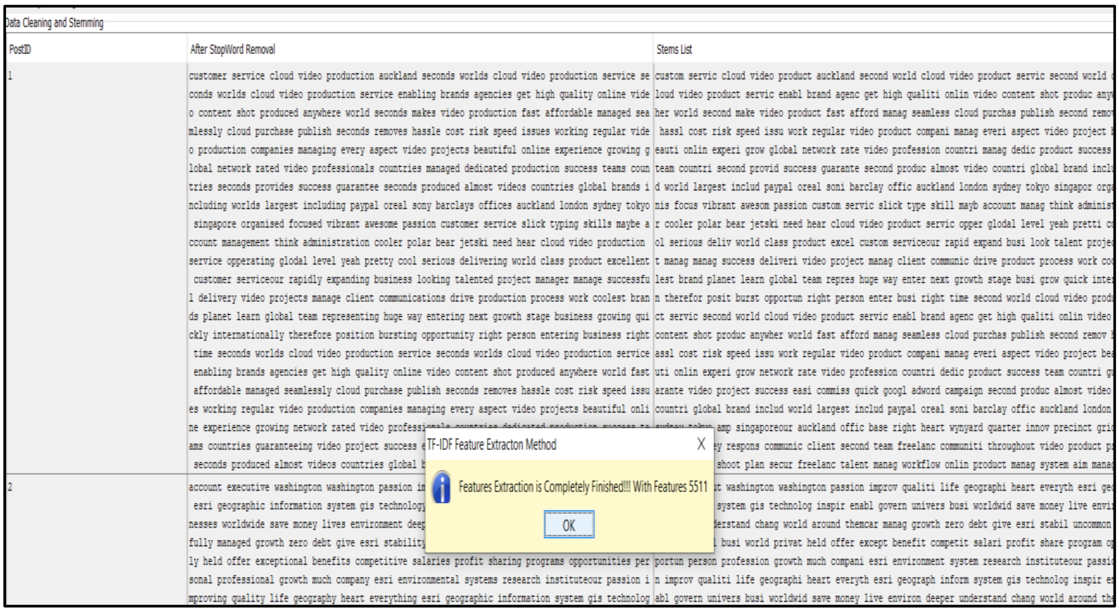


PostID	After StopWord Removal	Stems List
1	customer service cloud video production auckland seconds worlds cloud video production service seconds worlds cloud video production service enabling brands agencies get high quality online video content shot produced anywhere world seconds makes video production fast affordable managed seamlessly cloud purchase publish seconds removes hassle cost risk speed issues working regular video production companies managing every aspect video projects beautiful online experience growing global network rated video professionals countries managed dedicated production success teams countries seconds provides success guarantee seconds produced almost videos countries global brands including worlds largest including paypal oreal sony barclays offices auckland london sydney tokyo singapore organised focused vibrant awesome passion customer service slick typing skills maybe account management think administration cooler polar bear jetski need hear cloud video production service opperating glodal level yeah pretty cool serious delivering world class product excellent customer serviceour rapidly expanding business looking talented project manager manage successful delivery video projects manage client communications drive production process work coolest brands planet learn global team representing huge way entering next growth stage business growing quickly internationally therefore position bursting opportunity right person entering business right time seconds worlds cloud video production service seconds worlds cloud video production service enabling brands agencies get high quality online video content shot produced anywhere world fast affordable managed seamlessly cloud purchase publish seconds removes hassle cost risk speed issues working regular video production companies managing every aspect video projects beautiful online experience growing network rated video professionals countries dedicated production success teams countries guaranteeing video project success easy commissioning quick google adwords campaign seconds produced almost videos countries global brands including worlds largest including paypal	custom servic cloud video product auckland second world cloud video product servic second world c loud video product servic enabl brand agenc get high qualiti onlin video content shot produc anyw her world second make video product fast afford manag seamless cloud purchas publish second remov hassl cost risk speed issu work regular video product compani manag everi aspect video project b eauti onlin experi grow global network rate video profession countri manag dedic product success team countri second provid success guarante second produc almost video countri global brand inclu d world largest includ paypal oreal soni barclay offic auckland london sydney tokyo singapor orga nis focus vibrant awesom passion custom servic slick type skill mayb account manag think administ r cooler polar bear jetski need hear cloud video product servic opper glodal level yeah pretti co ol serious deliv world class product excel custom serviceour rapid expand busi look talent projec t manag manag success deliveri video project manag client communic drive product process work coo lest brand planet learn global team repres huge way enter next growth stage busi grow quick inter n therefor posit burst opportun right person enter busi right time second world cloud video produ ct servic second world cloud video product servic enabl brand agenc get high qualiti onlin video content shot produc anywher world fast afford manag seamless cloud purchas publish second remov h assl cost risk speed issu work regular video product compani manag everi aspect video project bea uti onlin experi grow network rate video profession countri dedic product success team countri gu arante video project success easi commiss quick googl adword campaign second produc almost video countri global brand includ world largest includ paypal oreal soni barclay offic auckland london sydney tokyo amp singaporeour auckland offic base right heart wynyard quarter innov precinct grid akl expect key respons communic client second team freelanc communiti throughout video product pr ocess includ shoot plan secur freelanc talent manag workflow onlin product manag system aim manag
2	account executive washington washington passion improving quality life geography heart everything	account execut washington washington passion improv qualiti life geographi heart everyth esri geo

Figure 4.4 Training Data Cleaning and Stemming

4.1.4 Feature Extraction using TF-IDF

The term is weighted using TF-IDF after text preprocessing is finished. Figure 4.5 is shown how many features and Figure 4.6 is shown the weight calculation on Training data. The extracted features are stored as feature vectors in a JAVA file with the extension for use in a fake job categorization.



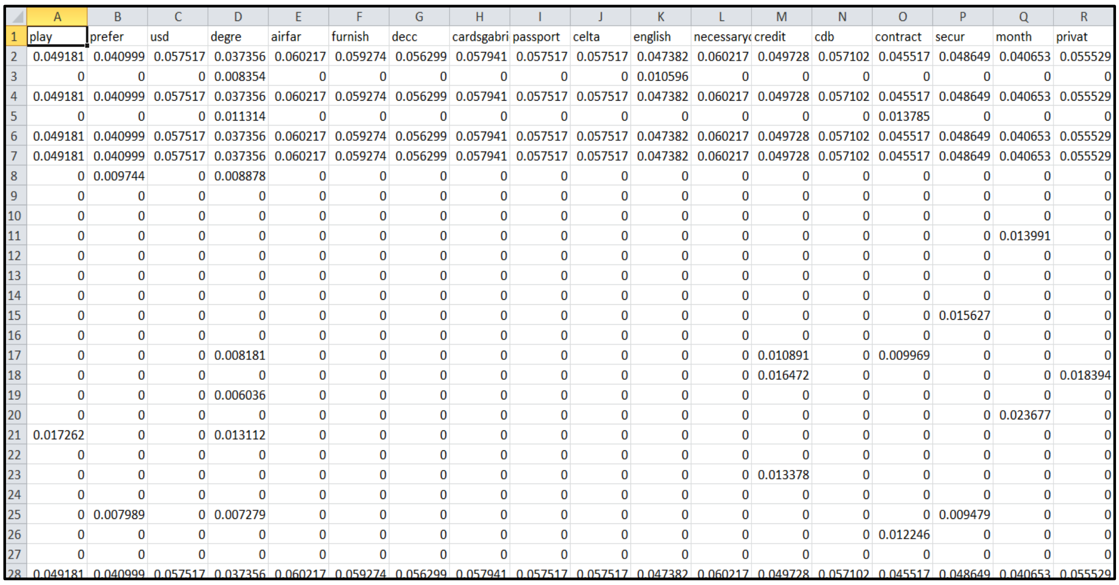
Data Cleaning and Stemming

PostID	After StopWord Removal	Stems List
1	customer service cloud video production auckland seconds worlds cloud video production service se conds worlds cloud video production service enabling brands agencies get high quality online vide o content shot produced anywhere world seconds makes video production fast affordable managed sea mlessly cloud purchase publish seconds removes hassle cost risk speed issues working regular vide o production companies managing every aspect video projects beautiful online experience growing g lobal network rated video professionals countries managed dedicated production success teams coun tries seconds provides success guarantee seconds produced almost videos countries global brands i ncluding worlds largest including paypal oreal sony barclays offices auckland london sydney tokyo singapore organised focused vibrant awesome passion customer service slick typing skills maybe a ccount management think administration cooler polar bear jetski need hear cloud video production service opperating glodal level yeah pretty cool serious delivering world class product excellent customer serviceour rapidly expanding business looking talented project manager manage successfu l delivery video projects manage client communications drive production process work coolest bran ds planet learn global team representing huge way entering next growth stage business growing qui ckly internationally therefore position bursting opportunity right person entering business right time seconds worlds cloud video production service seconds worlds cloud video production service enabling brands agencies get high quality online video content shot produced anywhere world fast affordable managed seamlessly cloud purchase publish seconds removes hassle cost risk speed issu es working regular video production companies managing every aspect video projects beautiful onli ne experience growing network rated video professionals countries dedicated production success te ams countries guaranteeing video project success e seconds produced almost videos countries global b	custom servic cloud video product auckland second world cloud video product servic second world loud video product servic enabl brand agenc get high qualiti onlin video content shot produc any her world second make video product fast afford manag seamless cloud purchas publish second remov hassl cost risk speed issu work regular video product compani manag everi aspect video project b eauti onlin experi grow global network rate video profession countri manag dedic product success team countri second provid success guarante second produc almost video countri global brand inclu d world largest includ paypal oreal soni barclay offic auckland london sydney tokyo singapor orga nis focus vibrant awesom passion custom servic slick type skill mayb account manag think administ r cooler polar bear jetski need hear cloud video product servic opper glodal level yeah pretti co ol serious deliv world class product excel custom serviceour rapid expand busi look talent projec t manag manag success deliveri video project manag client communic drive product process work coo lest brand planet learn global team repres huge way enter next growth stage busi grow quick inter n therefor posit burst opportun right person enter busi right time second world cloud video produ ct servic second world cloud video product servic enabl brand agenc get high qualiti onlin video content shot produc anywher world fast afford manag seamless cloud purchas publish second remov h assl cost risk speed issu work regular video product compani manag everi aspect video project bea uti onlin experi grow network rate video profession countri dedic product success team countri gu arante video project success easi commiss quick googl adword campaign second produc almost video countri global brand includ world largest includ paypal oreal soni barclay offic auckland london sydney tokyo amp singaporeour auckland offic base right heart wynyard quarter innov precinct gri ey respons communic client second team freelanc communiti throughout video product p shoot plan secur freelanc talent manag workflow onlin product manag system aim manag
2	account executive washington washington passion im esri geographic information system gis technology nesses worldwide save money lives environment deep fully managed growth zero debt give esri stability ly held offer exceptional benefits competitive salaries profit sharing programs opportunities per sonal professional growth much company esri environmental systems research instituteour passion i mproving quality life geography heart everything esri geographic information system gis technolog	t washington washington passion improv qualiti life geographi heart everyth esri geo system gis technolog inspir enabl govern univers busi worldwid save money live envi derstand chang world around themcar manag growth zero debt give esri stabil uncommon l busi world privat held offer except benefit competit salari profit share program o portun person profession growth much compani esri environment system research instituteour passi n improv qualiti life geographi heart everyth esri geograph inform system gis technolog inspir e abl govern univers busi worldwid save money live environ deeper understand chang world around the

TF-IDF Feature Extracton Method

Features Extraction is Completely Finished!!! With Features 5511

Figure 4.5 Feature Extraction of each word

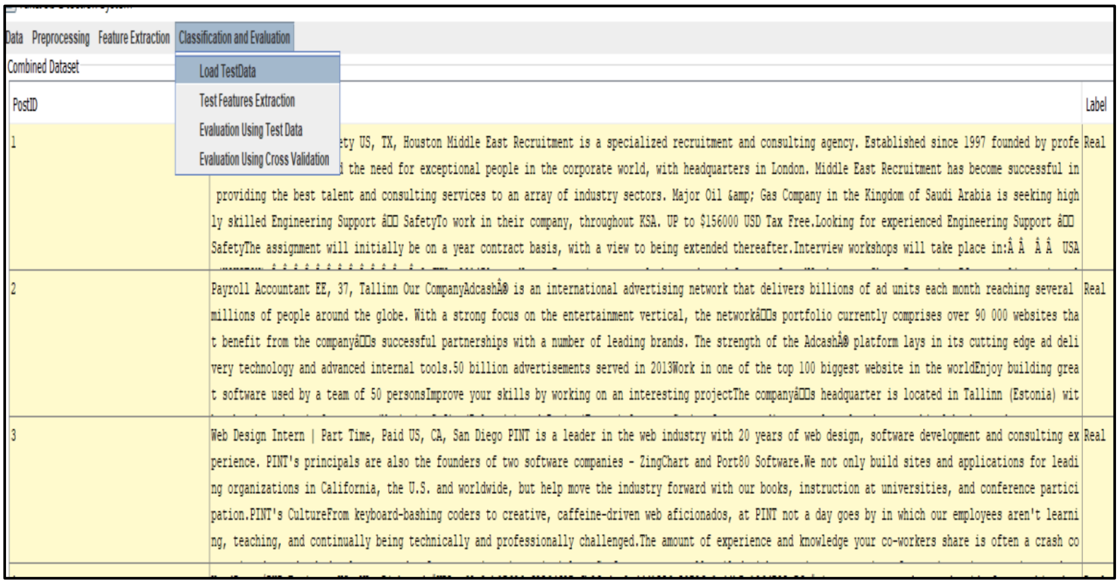


	A	B	C	D	E	F	G	H	I	J	K	L	M	N	O	P	Q	R
1	play	prefer	usd	degre	airfar	furnish	decc	cardsgabri	passport	celta	english	necessary	credit	cdb	contract	secur	month	privat
2	0.049181	0.040999	0.057517	0.037356	0.060217	0.059274	0.056299	0.057941	0.057517	0.057517	0.047382	0.060217	0.049728	0.057102	0.045517	0.048649	0.040653	0.055529
3	0	0	0	0.008354	0	0	0	0	0	0	0.010596	0	0	0	0	0	0	0
4	0.049181	0.040999	0.057517	0.037356	0.060217	0.059274	0.056299	0.057941	0.057517	0.057517	0.047382	0.060217	0.049728	0.057102	0.045517	0.048649	0.040653	0.055529
5	0	0	0	0.011314	0	0	0	0	0	0	0	0	0	0	0.013785	0	0	0
6	0.049181	0.040999	0.057517	0.037356	0.060217	0.059274	0.056299	0.057941	0.057517	0.057517	0.047382	0.060217	0.049728	0.057102	0.045517	0.048649	0.040653	0.055529
7	0.049181	0.040999	0.057517	0.037356	0.060217	0.059274	0.056299	0.057941	0.057517	0.057517	0.047382	0.060217	0.049728	0.057102	0.045517	0.048649	0.040653	0.055529
8	0	0.009744	0	0.008878	0	0	0	0	0	0	0	0	0	0	0	0	0	0
9	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
10	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
11	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0.013991	0
12	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
13	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
14	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
15	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0.015627	0	0
16	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
17	0	0	0	0.008181	0	0	0	0	0	0	0	0	0.010891	0	0.009969	0	0	0
18	0	0	0	0	0	0	0	0	0	0	0	0	0.016472	0	0	0	0	0.018394
19	0	0	0	0.006036	0	0	0	0	0	0	0	0	0	0	0	0	0	0
20	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0.023677	0
21	0.017262	0	0	0.013112	0	0	0	0	0	0	0	0	0	0	0	0	0	0
22	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
23	0	0	0	0	0	0	0	0	0	0	0	0	0.013378	0	0	0	0	0
24	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
25	0	0.007989	0	0.007279	0	0	0	0	0	0	0	0	0	0	0	0.009479	0	0
26	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0.012246	0	0	0
27	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
28	0.049181	0.040999	0.057517	0.037356	0.060217	0.059274	0.056299	0.057941	0.057517	0.057517	0.047382	0.060217	0.049728	0.057102	0.045517	0.048649	0.040653	0.055529

Figure 4.6 Weight calculations on Training data

4.1.5 Load Test Data

After loading the testing data, preprocessing is required for tokenization, remove stop words, and stemming like training data. Figure 4.7 is shown loading for test data.

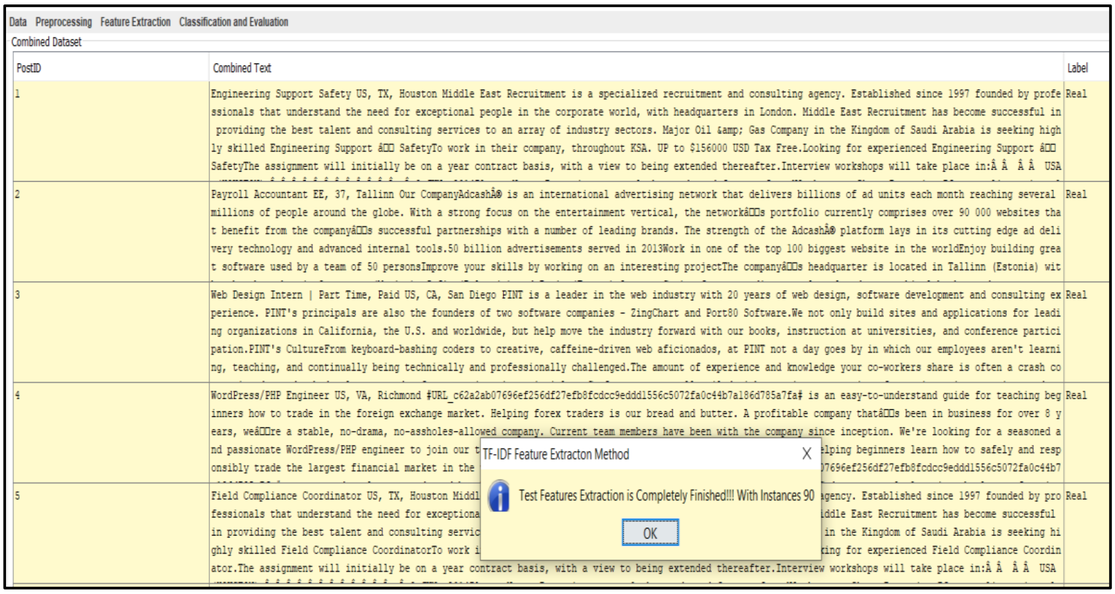

Data
Preprocessing
Feature Extraction
Classification and Evaluation
Load TestData
Test Features Extraction
Evaluation Using Test Data
Evaluation Using Cross Validation
Combined Dataset

PostID		Label
1	ety US, TX, Houston Middle East Recruitment is a specialized recruitment and consulting agency. Established since 1997 founded by profe d the need for exceptional people in the corporate world, with headquarters in London. Middle East Recruitment has become successful in providing the best talent and consulting services to an array of industry sectors. Major Oil &amp; Gas Company in the Kingdom of Saudi Arabia is seeking highly skilled Engineering Support â□□ SafetyTo work in their company, throughout KSA. UP to $156000 USD Tax Free.Looking for experienced Engineering Support â□□ SafetyThe assignment will initially be on a year contract basis, with a view to being extended thereafter.Interview workshops will take place in:Â Â Â Â USA	Real
2	Payroll Accountant EE, 37, Tallinn Our CompanyAdcashÂ® is an international advertising network that delivers billions of ad units each month reaching several millions of people around the globe. With a strong focus on the entertainment vertical, the networkâ□□s portfolio currently comprises over 90 000 websites that benefit from the companyâ□□s successful partnerships with a number of leading brands. The strength of the AdcashÂ® platform lays in its cutting edge ad delivery technology and advanced internal tools.50 billion advertisements served in 2013Work in one of the top 100 biggest website in the worldEnjoy building great software used by a team of 50 personsImprove your skills by working on an interesting projectThe companyâ□□s headquarter is located in Tallinn (Estonia) wit	Real
3	Web Design Intern | Part Time, Paid US, CA, San Diego PINT is a leader in the web industry with 20 years of web design, software development and consulting experience. PINT's principals are also the founders of two software companies - ZingChart and Port80 Software.We not only build sites and applications for leading organizations in California, the U.S. and worldwide, but help move the industry forward with our books, instruction at universities, and conference participation.PINT's CultureFrom keyboard-bashing coders to creative, caffeine-driven web aficionados, at PINT not a day goes by in which our employees aren't learning, teaching, and continually being technically and professionally challenged.The amount of experience and knowledge your co-workers share is often a crash co	Real

Figure 4.7 Load Test Data

4.1.6 Test Feature Extraction with Instances

Figure 4.8 is the step of feature extraction how many instances of test data.


Data
Preprocessing
Feature Extraction
Classification and Evaluation
Combined Dataset
TF-IDF Feature Extracton Method
X
OK

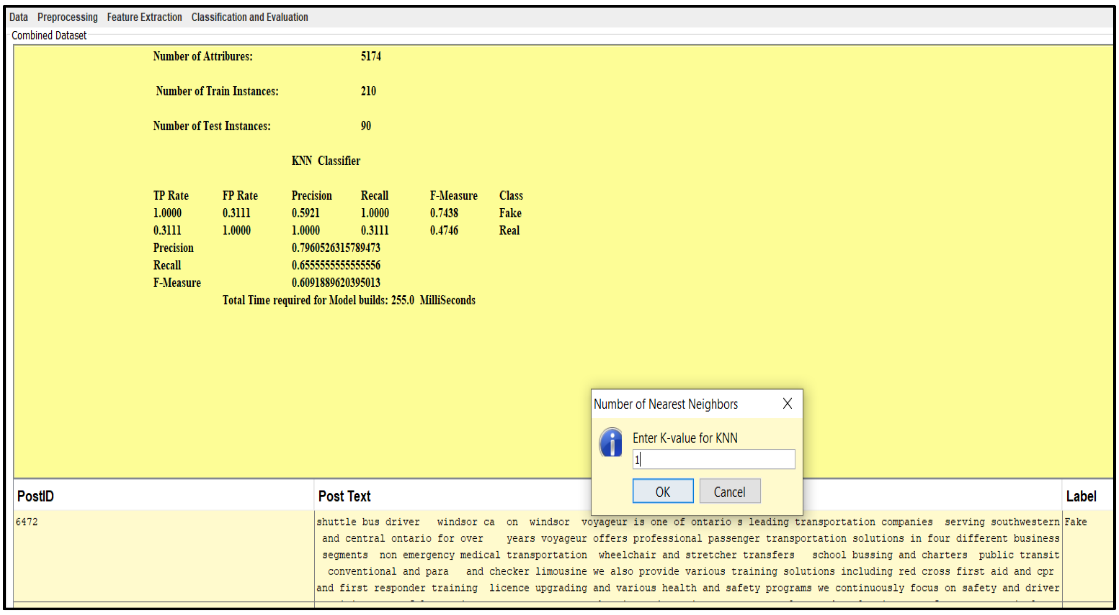
Test Features Extraction is Completely Finished!!! With Instances 90

PostID	Combined Text	Label
1	Engineering Support Safety US, TX, Houston Middle East Recruitment is a specialized recruitment and consulting agency. Established since 1997 founded by professionals that understand the need for exceptional people in the corporate world, with headquarters in London. Middle East Recruitment has become successful in providing the best talent and consulting services to an array of industry sectors. Major Oil &amp; Gas Company in the Kingdom of Saudi Arabia is seeking highly skilled Engineering Support â□□ SafetyTo work in their company, throughout KSA. UP to $156000 USD Tax Free.Looking for experienced Engineering Support â□□ SafetyThe assignment will initially be on a year contract basis, with a view to being extended thereafter.Interview workshops will take place in:Â Â Â Â USA	Real
2	Payroll Accountant EE, 37, Tallinn Our CompanyAdcashÂ® is an international advertising network that delivers billions of ad units each month reaching several millions of people around the globe. With a strong focus on the entertainment vertical, the networkâ□□s portfolio currently comprises over 90 000 websites that benefit from the companyâ□□s successful partnerships with a number of leading brands. The strength of the AdcashÂ® platform lays in its cutting edge ad delivery technology and advanced internal tools.50 billion advertisements served in 2013Work in one of the top 100 biggest website in the worldEnjoy building great software used by a team of 50 personsImprove your skills by working on an interesting projectThe companyâ□□s headquarter is located in Tallinn (Estonia) wit	Real
3	Web Design Intern | Part Time, Paid US, CA, San Diego PINT is a leader in the web industry with 20 years of web design, software development and consulting experience. PINT's principals are also the founders of two software companies - ZingChart and Port80 Software.We not only build sites and applications for leading organizations in California, the U.S. and worldwide, but help move the industry forward with our books, instruction at universities, and conference participation.PINT's CultureFrom keyboard-bashing coders to creative, caffeine-driven web aficionados, at PINT not a day goes by in which our employees aren't learning, teaching, and continually being technically and professionally challenged.The amount of experience and knowledge your co-workers share is often a crash co	Real
4	WordPress/PHP Engineer US, VA, Richmond #URL_c62a2ab07696ef256df27efb8fcdcc9eddd1556c5072fa0c44b7a186d785a7fa# is an easy-to-understand guide for teaching beginners how to trade in the foreign exchange market. Helping forex traders is our bread and butter. A profitable company thatâ□□s been in business for over 8 years, weâ□□re a stable, no-drama, no-assholes-allowed company. Current team members have been with the company since inception. We're looking for a seasoned and passionate WordPress/PHP engineer to join our t elping beginners learn how to safely and responsibly trade the largest financial market in the 7696ef256df27efb8fcdcc9eddd1556c5072fa0c44b7	Real
5	Field Compliance Coordinator US, TX, Houston Middl agency. Established since 1997 founded by professionals that understand the need for exceptiona iddle East Recruitment has become successful in providing the best talent and consulting servic in the Kingdom of Saudi Arabia is seeking highly skilled Field Compliance CoordinatorTo work i ing for experienced Field Compliance Coordinator.The assignment will initially be on a year contract basis, with a view to being extended thereafter.Interview workshops will take place in:Â Â Â Â USA	Real

Figure 4.8 Instances of Test Feature Extraction

4.1.7 Evaluation Using Test Data

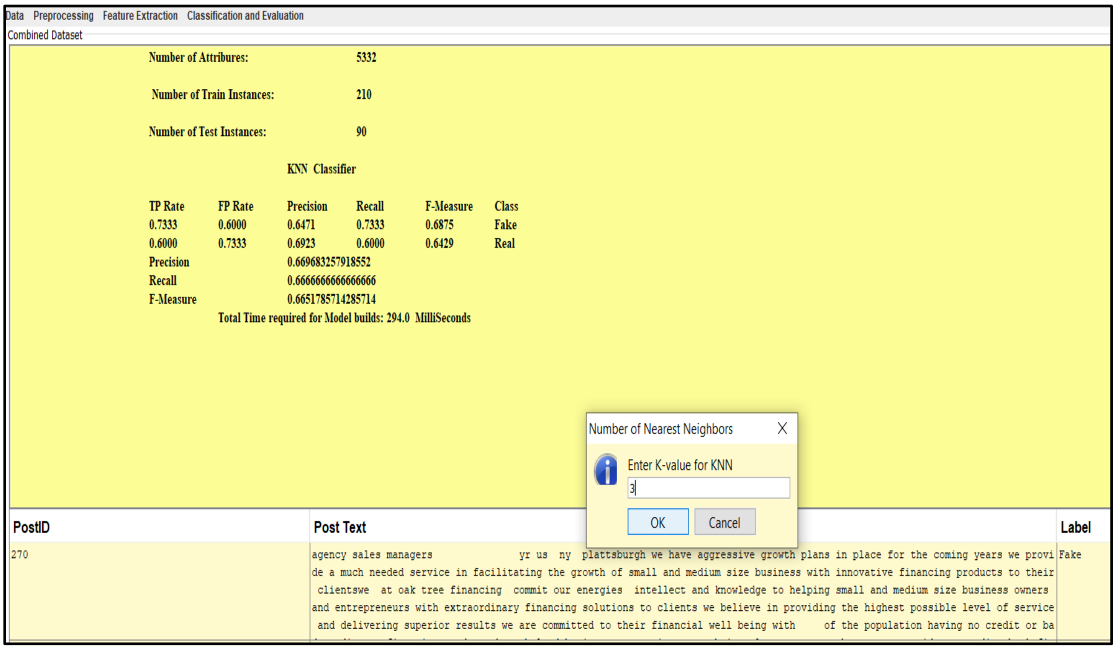
The performance evaluation of test data with different K values of KNN Classifier is described in Figure (4.9), (4.10), (4.11). This outcome is used to illustrate the binary class's confusion matrix as well as the precision, recall, and f-measure values.



```
Number of Attribures:              5174

 Number of Train Instances:        210

Number of Test Instances:          90

                    KNN  Classifier

TP Rate    FP Rate    Precision    Recall    F-Measure    Class
1.0000     0.3111     0.5921       1.0000    0.7438       Fake
0.3111     1.0000     1.0000       0.3111    0.4746       Real
Precision             0.7960526315789473
Recall                0.6555555555555556
F-Measure             0.6091889620395013
           Total Time required for Model builds: 255.0  MilliSeconds
```

PostID	Post Text	Label
6472	shuttle bus driver windsor ca on windsor voyageur is one of ontario s leading transportation companies serving southwestern and central ontario for over years voyageur offers professional passenger transportation solutions in four different business segments non emergency medical transportation wheelchair and stretcher transfers school bussing and charters public transit conventional and para and checker limousine we also provide various training solutions including red cross first aid and cpr and first responder training licence upgrading and various health and safety programs we continuously focus on safety and driver	Fake

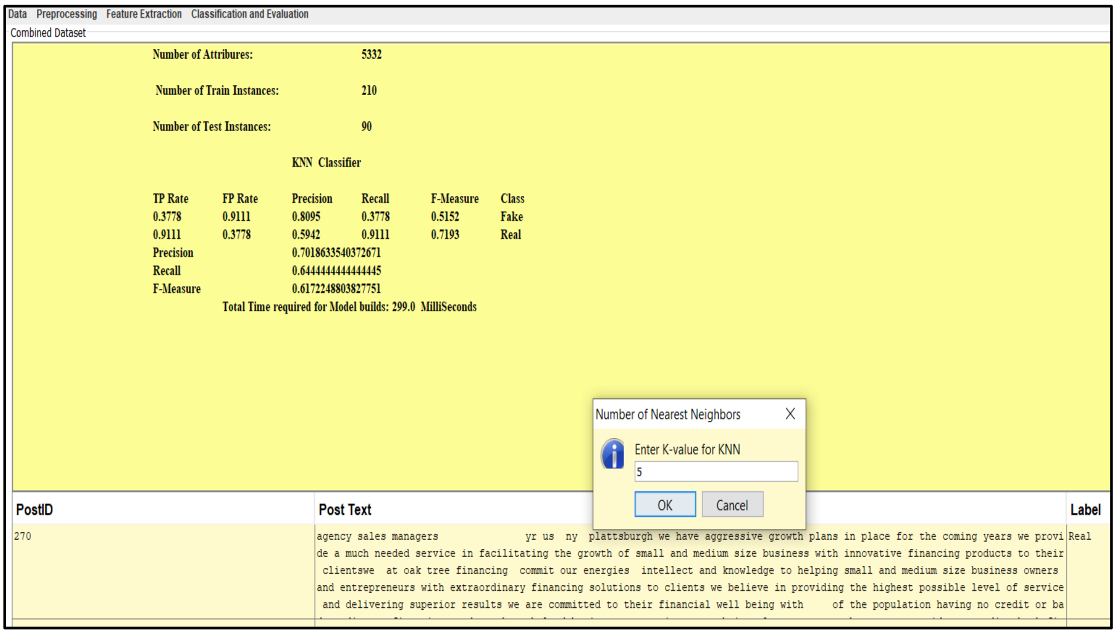
Figure 4.9 Evaluation Using Test Data with K=1



```
Number of Attribures:              5332

 Number of Train Instances:        210

Number of Test Instances:          90

                    KNN  Classifier

TP Rate    FP Rate    Precision    Recall    F-Measure    Class
0.7333     0.6000     0.6471       0.7333    0.6875       Fake
0.6000     0.7333     0.6923       0.6000    0.6429       Real
Precision             0.669683257918552
Recall                0.6666666666666666
F-Measure             0.6651785714285714
           Total Time required for Model builds: 294.0  MilliSeconds
```

PostID	Post Text	Label
270	agency sales managers yr us ny plattsburgh we have aggressive growth plans in place for the coming years we provide a much needed service in facilitating the growth of small and medium size business with innovative financing products to their clientswe at oak tree financing commit our energies intellect and knowledge to helping small and medium size business owners and entrepreneurs with extraordinary financing solutions to clients we believe in providing the highest possible level of service and delivering superior results we are committed to their financial well being with of the population having no credit or ba	Fake

Figure 4.10 Evaluation Using Test Data with K=3



```
Number of Attribures:                    5332

 Number of Train Instances:              210

Number of Test Instances:                90

                          KNN  Classifier

TP Rate    FP Rate    Precision    Recall    F-Measure    Class
0.3778     0.9111     0.8095       0.3778    0.5152       Fake
0.9111     0.3778     0.5942       0.9111    0.7193       Real
Precision             0.7018633540372671
Recall                0.6444444444444445
F-Measure             0.6172248803827751
           Total Time required for Model builds: 299.0  MilliSeconds
```

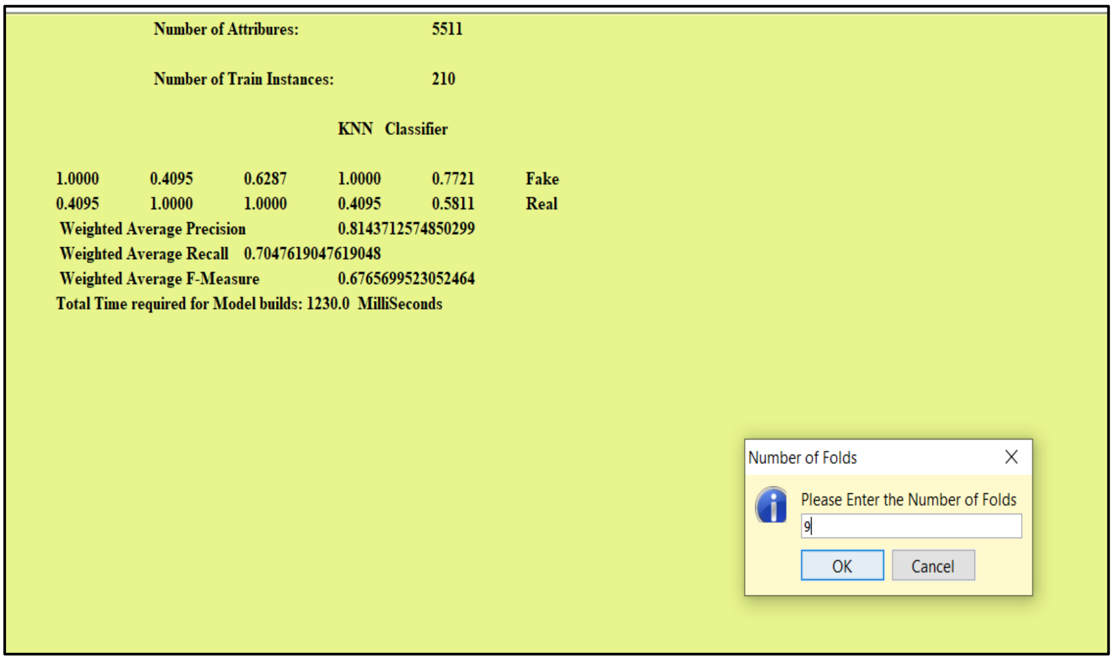
Enter K-value for KNN: 5

PostID	Post Text	Label
270	agency sales managers yr us ny plattsburgh we have aggressive growth plans in place for the coming years we provide a much needed service in facilitating the growth of small and medium size business with innovative financing products to their clientswe at oak tree financing commit our energies intellect and knowledge to helping small and medium size business owners and entrepreneurs with extraordinary financing solutions to clients we believe in providing the highest possible level of service and delivering superior results we are committed to their financial well being with of the population having no credit or ba	Real

Figure 4.11 Evaluation Using Test Data with K=5

4.1.8 Evaluation Using Cross Validation

The performance evaluation of test data 300 by using 9-Cross Validation result is shown in Figures (4.12).



```
          Number of Attribures:                5511

          Number of Train Instances:           210

                             KNN   Classifier

1.0000     0.4095     0.6287     1.0000     0.7721     Fake
0.4095     1.0000     1.0000     0.4095     0.5811     Real
 Weighted Average Precision            0.8143712574850299
 Weighted Average Recall   0.7047619047619048
 Weighted Average F-Measure            0.6765699523052464
Total Time required for Model builds: 1230.0  MilliSeconds
```

Please Enter the Number of Folds: 9

Figure 4.12 9 fold cross-validation result

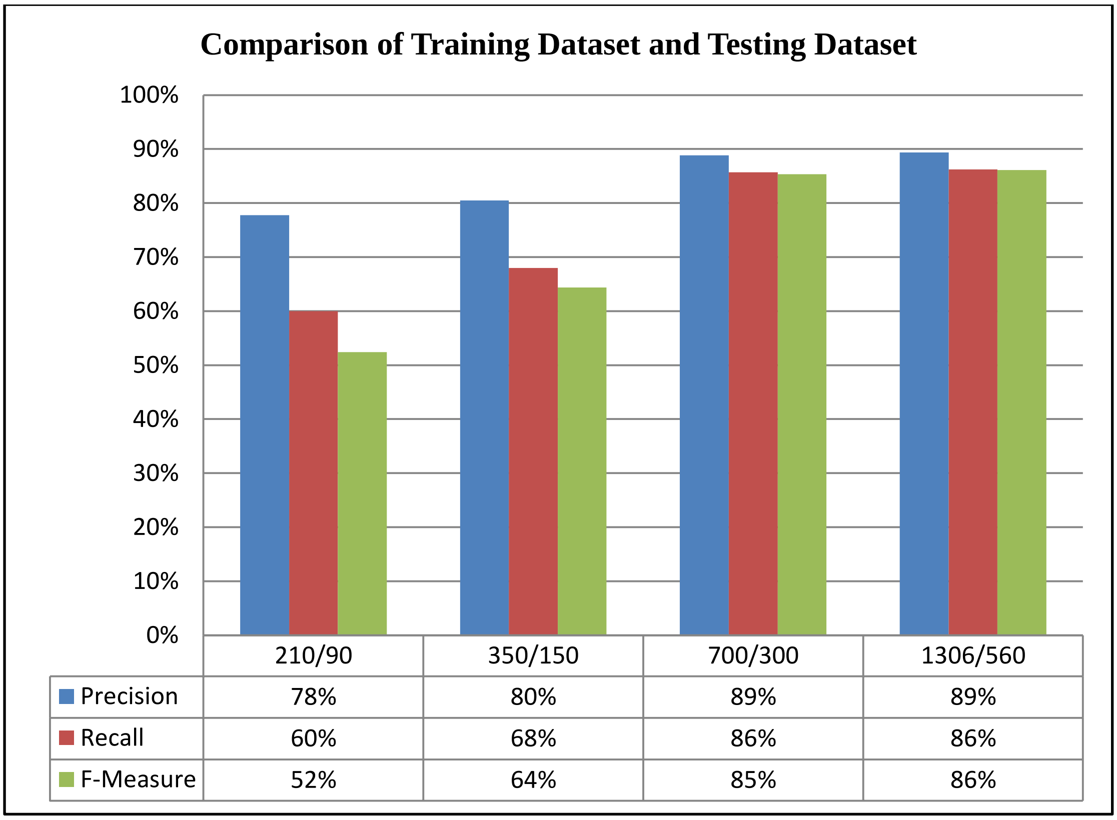
4.2 Experimental Results

The analysis results of different dataset are shown in table 4.1 and Figure 4.13.

Table 4.1 Experimental Results of Analysis

Testing No	Training and Testing data Proportion	Precision Result	Recall Result	F-Measure
Test 1	210/90	78%	60%	52%
Test 2	350/150	80%	68%	64%
Test 3	700/300	89%	86%	85%
Test 4	1306/560	89%	86%	86%

If more trained data can be fed into the system, based on the analysis, can produce better results.



	210/90	350/150	700/300	1306/560
🟦 Precision	78%	80%	89%	89%
🟥 Recall	60%	68%	86%	86%
🟩 F-Measure	52%	64%	85%	86%

Figure 4.13 Experimental Results with Different Datasets

Table 4.2 is showed the sample example of 300 jobs for experimental result with different K values.

Table 4.2 Sample example of 300 jobs with different K values

Performance Measure Metric	K=3	K=5	K=7	K=9
Precision	86%	86%	87%	88%
Recall	80%	84%	87%	86%
F-measure	79%	84%	87%	85%

By the comparison of different k values with 300 datasets, K = 9 is better; when comparing K values, the more data, the higher the K values.

CHAPTER 5
CONCLUSION

Employment scam is one of the serious issues in recent times addressed in the domain of online recruitment frauds. Nowadays, a lot of businesses decide to post job vacancies online so that job seekers can easily and swiftly access. The detection of fake job posts can make a job-seeker only apply for legitimate companies. Specifically, supervised learning classifier is used for real or fake job detection. This system mainly focuses on implementing the classifier for Fake Job Posts detection using the K-Nearest Neighbor (KNN) classifier with feature extraction (TF-IDF) method. The K-Nearest Neighbor algorithm is one of the most extensively used and successful text categorization methods.

The KNN algorithm is used to evaluate the performance of the proposed system for real or fake job posts. According to the experimental results, the K-Nearest Neighbor algorithm with Euclidean Distance has better performance. The performance accuracy results were calculated with different K values. The proposed system has many experimental results by using the K-Nearest Neighbor algorithm with k values (K = 1, 3, 5, 7, 9). In this chapter, the main contents of the research work are concluded. Moreover, the advantages and future wok are also discussed in this chapter.

5.1 Advantages of the System

By applying this method, candidates can avoid the fraudulent post for job in the internet. For better outcomes before training the model, TF-IDF is often applied to the text. Stop words is eliminated and stemmed before TF-IDF produced in more accurate text classification. The TF-IDF feature extraction method can provide suitable feature vectors to improve the classifier's performance. According to the evaluation of the system, after completing the training phase, the system can successfully complete classification tasks. The accuracy results can be improved by using more training data.

5.2 Further Extensions

Furthermore, in fake job post detection system, instead of using KNN as a classifier, more powerful classifiers in machine learning like Naive Bayes and Random Forest must be applied to fake job posts detection.

AUTHOR'S PUBLICATIONS

[1] Khin Mar Htay, Yu Yu Than, "The Detection of Fake Job Posts by Using K-Nearest Neighbor (KNN)", National Journal of Parallel and Soft Computing, University of Computer Studies, Yangon, 2022.

REFERENCES

[1] Ahmed .S, An. Nabi. D, “Survey on Classification Algorithms for Data Mining: (Comparison and Evaluation)”, Computer Engineering and Intelligent Systems, 2013.

[2] Alandjani, “ Online Fake Job Advertisement Recognition and Classification Using Machine Learning”, 3C TIC. Cuadernos de desarrollo aplicados a las TIC, 11(1), 251-267, https://doi.org/10.17993/3ctic.2022.111.251-267, 2022.

[3] Andrea Gasparetto , Matteo Marcuzzo , Alessandro Zangari and Andrea Albarelli ,“A Survey on Text Classification Algorithms: From Text to Predictions”, Information, February 2022.

[4] Anita , Nagarajan; Aditya Sairam; Ganesh, Deepakkumar, “Fake Job Detection and Analysis Using Machine Learning and Deep Learning Algorithms”, International Journal of Engineering Trends and Technology (IJETT), February 2022.

[5] Anjali Ganesh Jivani, “A Comparative Study of Stemming Algorithms”, Department of Computer Science & Engineering, 2011.

[6] Auria, Laura; Moro, Rouslan A, “Support Vector Machines (SVM) as a technique for solvency analysis”, German Institute for Economic Research, 2008.

[7] Bandar Alghamdi, Fahad Alharby, “An Intelligent Model for Online Recruitment Fraud Detection”, Journal of Information Security, 2019.

[8] Batta Mahesh, “Machine Learning Algorithms - A Review”, International Journal of Science and Research (IJSR), 2020.

[9] S. Bansal, “[Real or Fake]: Fake Job Posting Prediction” Kaggle.
https://www.kaggle.com/shivamb/real-or-fake-fakejobposting-prediction
(accessed Jun. 03, 2020).

[10] Bruno Trstenjaka ,Sasa Mikacb , Dzenana Donkoc, “KNN with TF-IDF Based Framework for text Categorization”, 24th DAAAM International Symposium on Intelligent Manufacturing and Automation, 2013.

[11] Chetna Khanna "Text pre-processing: Stop words removal using different libraries", from https://towardsdatascience.com/text-pre-processing-stop-words-removal-using-different-libraries-f20bac19929a.

[12] "Feature extraction for machine learning and deep learning", from https://www.mathworks.com/discovery/feature-extraction.html.

[13] Hakim, Erwin, K. Eng, M. Galinium, and W. Muliady, "Automated document classification for news article in bahasa Indonesia based on term frequency inverse document frequency (TFIDF) approach", ICITEE, pp. 1-4, 2014.

[14] Han EH.., Karypis G., Kumar V. ,"Text Categorization Using Weight Adjusted k-Nearest Neighbour Classification", Advances in Knowledge Discovery and Data Mining, 2001.

[15] Ibrahim M. Nasser and Amjad H. Alzaanin, "Machine Learning and Job Posting Classification: A Comparative Study", International Journal of Engineering and Information Systems (IJEAIS), September, 2020.

[16] Kashvi Taunk, Sanjukta De, Srishti Verma, "Machine Learning Classification With K-Nearest Neighbours", Kiit University, 2019.

[17] Lior Rokach, Oded Maimon, "DECISION TREES", Department of Industrial Engineering, 2005.

[18] Mohammad Waseeem, " How To Implement Classification In Machine Learning?", from https://www.edureka.co/blog/classification-in-machine-learning.

[19] Onel Harrison, "Machine Learning Basics with the K-Nearest Neighbors Algorithm", from https://towardsdatascience.com/machine-learning-basics-with-the-k-nearest-neighbors-algorithm-6a6e71d01761, 2018.

[20] Osisanwo F.Y , Akinsola J.E.T, Awodele O , Hinmikaiye J. O , Olakanmi O ,Akinjobi J, "Supervised Machine Learning Algorithms: Classification and Comparison", International Journal of Computer Trends and Technology (IJCTT), July 2017.

[21] Richard S. Sutton and Andrew G. Barto, “Reinforcement Learning: An Introduction”, Second edition.

[22] Roshna Chettri, Shrijana Pradhan, Lekhika Chettri, “Internet of Things: Comparative Study on Classification Algorithms (k-NN, Naive Bayes and Case based Reasoning)” , International Journal of Computer Applications, 2015.

[23] Shawni Dutta, Prof.Samir Kumar Bandyopadhyay, “Fake Job Recruitment Detection Using Machine Learning Approach”, International Journal of Engineering Trends and Technology (IJETT), 2020.

[24] Shengyi Jiang,Guansong Pang,Meiling Wu,Limin Kuang, “An improved K-nearest-neighbour algorithm for text categorization”, Expert Systems with Applications, Elsevier 2012.

[25] “Text Classification: What it is And Why it Matters” from https://monkeylearn.com/text-classification.

[26] Vaibhav Kumar, “Classifying Fake and Real Job Advertisements using Machine Learning”, https://analyticsindiamag.com/classifying-fake-and-real-job-advertisements-using-machine-learning ,2020.

[27] Van Huynh, T., Van Nguyen, K., Nguyen, N. L. T., & Nguyen, A. G. T., “Job prediction: From deep neural network models to applications”, In RIVF International Conference on Computing and Communication Technologies (RIVF), 2020.

[28] “What is Online Recruitment and What are Its Advantages?”, from https://www.fountain.com/posts/what-is-online-recruitment-and-what-are-its-advantages.

[29] Wettschereck and D. Thomas G., “Locally Adaptive Nearest Neighbor Algorithms,” Adv. Neural Inf. Process. Syst., pg. 184–186, 1994.

[30] WiraBuana, Jannet D.R.M., and Ketut Gede Darma Putra, “Combination of K-Nearest Neighbour and K-Means based on Term Re-weighting for Classify Indonesian News,” Int. J. Comput. Appl., vol. 50, no. 11, pp. 37–42, July 2012.

[31] Xindong Wu · Vipin Kumar · J. Ross Quinlan , “Top 10 algorithms in data mining” Knowl Inf Syst, 2008.

[32] Zhuoqiao Hong , “Measuring Pro-social Message in Job Postings Using Machine Learning”, Massachusetts Institute of Technology, 2020.