FAKE JOB IDENTIFICATION

Thesis to be submitted in partial fulfillment of the requirements for the degree

of

M.TECH

by

SHUBHAPRIYA GHOSH
20ET61R05

Under the guidance of

MANJIRA SHINA

ATDC

INDIAN INSTITUTE OF TECHNOLOGY KHARAGPUR

Department of ATDC
Indian Institute of Technology,
Kharagpur
India - 721302

CERTIFICATE

This is to certify that we have examined the thesis entitled **FAKE JOB IDENTIFICATION**, submitted by **SHUBHAPRIYA GHOSH**(Roll Number: *20ET61R05*) a postgraduate student of **Department of ATDC** in partial fulfillment for the award of degree of M.TECH. We hereby accord our approval of it as a study carried out and presented in a manner required for its acceptance in partial fulfillment for the Post Graduate Degree for which it has been submitted. The thesis has fulfilled all the requirements as per the regulations of the Institute and has reached the standard needed for submission.

MANJIRA SHINA

Supervisor

Department of ATDC
Indian Institute of Technology,
Kharagpur

Place: Kharagpur
Date:04/05/2022

Abstract

In this project, machine learning and deep learning algorithms,LSTM,BERT are used so as to identify fake jobs and to differentiate them from real jobs.To prevent fraudulent post for job in the internet, an automated tool using machine learning and deep learning based classification techniques is proposed in the project. The data analysis part and data cleaning part are also proposed in this project, so that the classification algorithm applied is highly precise and accurate. The classification and detection of fake jobs can be done with high accuracy and high precision. Hence the machine learning and deep learning algorithms have to be applied on cleaned and preprocessed data in order to achieve a better accuracy. Finally all these classification models are compared with each other to find the classification algorithm with highest accuracy and precision. Keywords : Logistic Regression,Random Forest , Na¨ıve Bayes, Long Short Term Memory,Bert,Decision Tree,Cnn,Svm.

Contents

List of Figures

Chapter 1

Introduction

1.1 Introduction

In the domain of Online Recruitment Frauds (ORF)Employment job scam is one of the serious issues in recent times.Current poor market condition leads to high unemployment.The main reasons of reduction of job availability and the loss of jobs for many individuals is economic stress and covid pendamic.It affects the market condition and the situation become worse day by day.Because of this job scammers have appropriate opportunity to attract people. In such situation job scammers posted a fraud jobs in various online job platforms against reputate companies to violates their credibility. Their intention only to attract the attention of job seekers to provide such fraud jobs and taking money from job seekers.Fraud job ads draws a attention to job seekers so to prevent this problem we need to automated tool to identify fraud jobs and report to the people. For such cases we build classifier using various machine learning algorithms,neural network ,lstm and more to identify fraud jobs and help to save people from job scammers.

1.2 Problem Statement

TASK : The main goal is to build a classifier that will have the ability to recognize fake and real jobs. The output result will be evaluated based on different models used in this project. Since the data involves has both numeric and text features so the various models used these features and help to find out the final output. The final model will take in any relevant job posting data and gives a final result determining whether the job is real or fake one. For such cases we build classifier using various

machine learning algorithms,neural network ,lstm and more to identify fraud jobs and help to save people from job scammers.

Chapter 2

Related Work

2.0.1 Introduction

In this chapter I will detail a bit more, some related works that are worth investigating. According to several studies, Review spam detection, Email Spam detection, Fake news detection have drawn special attention in the domain of Online Fraud Detection ”Review Spam Detection People most likely post their reviews online forum regarding the products they purchase in this context, spammers can manipulate reviews for gaining profit and hence it is required to develop techniques that detects these spam reviews. this can be implemented by extracting features from the reviews by extracting features using Natural Language Processing (NLP) next, machine learning techniques are applied on these features.”[1].

”Unwanted bulk mails, belong to the category of spam emails, often arrive to user mailbox and this may lead to unavoidable storage crisis as well as bandwidth consumption to eradicate this problem, Gmail, Yahoo mail and Outlook service providers incorporate spam filters using Neural Networks while addressing the problem of email spam detection, content based filtering, case based filtering, heuristic based filtering, memory or instance based filtering, adaptive spam filtering approaches are taken into consideration”[1].

”Fake news in social media characterizes malicious user accounts, echo chamber effects the fundamental study of fake news detection relies on three perspectives- how fake news is written, how fake news spreads, how a user is related to fake news features related to news content and social context are extracted and a machine learning models are imposed to recognize fake news” [1]. For such cases we need

help various machine learning algorithm to recognized fraud jobs which is posted in various online platforms.

Chapter 3

Data Analysis

3.1 DATASET

Real or Fake Job Posting Dataset is used in this project. The dataset provides both textual information and numerical about the jobs. The dataset consists of 17,880 observations and 18 features. Fake or real jobs is the main dataset, the data exploration will start with this dataset and is to count the number of items per class.

3.1.1 Data Analysis and Preprocessing

We have plotted various bar graphs for getting comparative information about real job postings as well as fake job postings.

Using Google Collaboratory environment, we transferred data and initialization it, then pragmatic features extraction. We split the output of initialization data vectors into training and testing data and then pragmatic various machine learning algorithms neural network,lstm,bert methods we mapped using python‘s libraries and evaluated them.

”Pre-processing data is a phase involves the text cleaning function and converting the text in the form suitable to the classification method includes extracting noise and uninformative characters and words in the text, such as HTML tags, where such words do not influence the general orientation of text”[4]. ”There are a lot of missing values in the dataset,, the researcher used IF methods to fill all missing value in the dataset file the value “None” was used to fill the blank cells in the department category, the value “Not specified” was used to fill the blank cells in the employment type category and the value “No Information” was used to fill the blank cells in the

required experience, required education, Industry, and function categories and after completing the missing values, specific text mining tools were required to handle the string data type”[4].

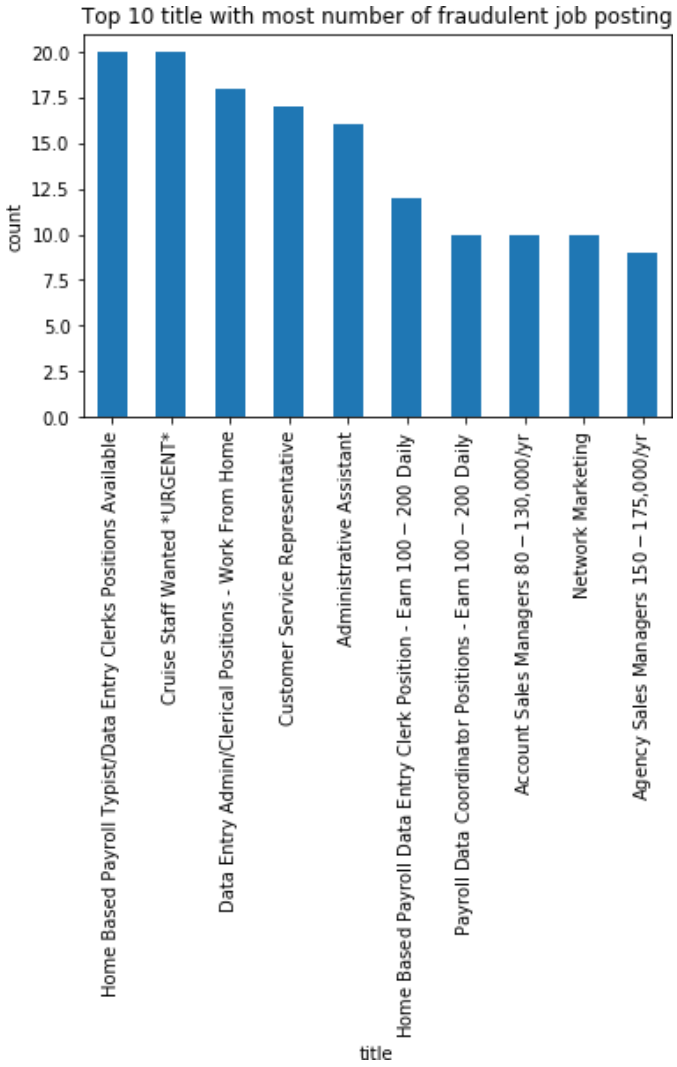

Figure 3.1: Top 10 title with most fraudulent job posting

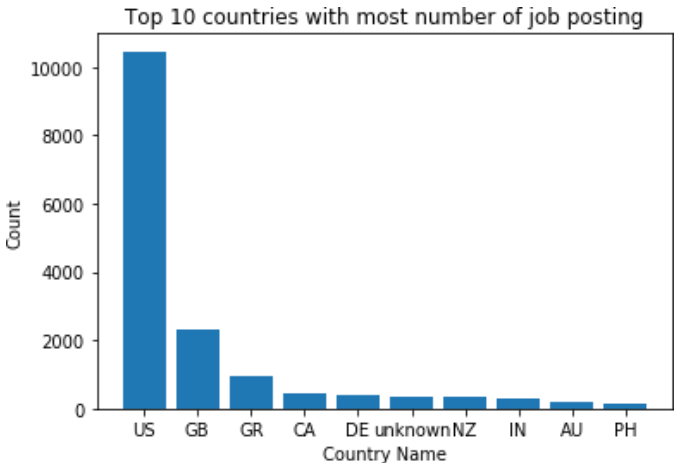

Figure 3.2: Top 10 countries with most job posting

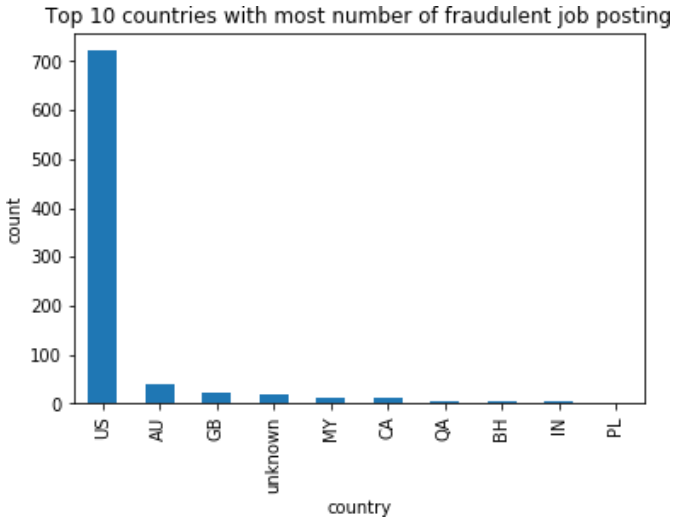

Figure 3.3: Top 10 title with most number of fraudulent job posting

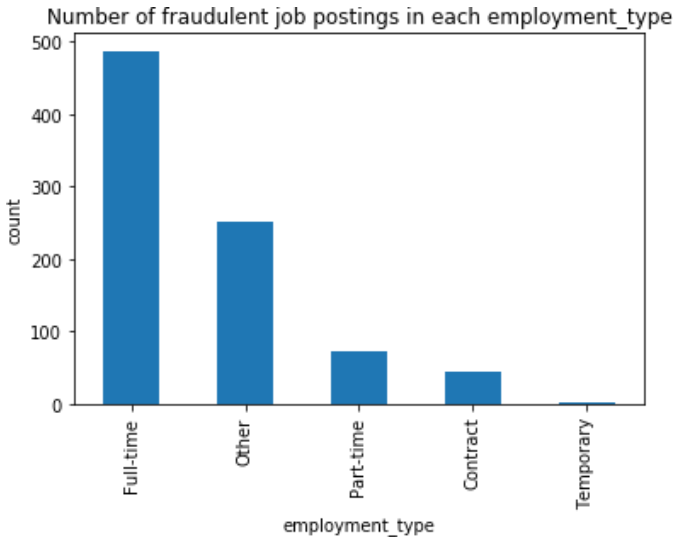

Figure 3.4: Number of fraudulent job posting in each employment type

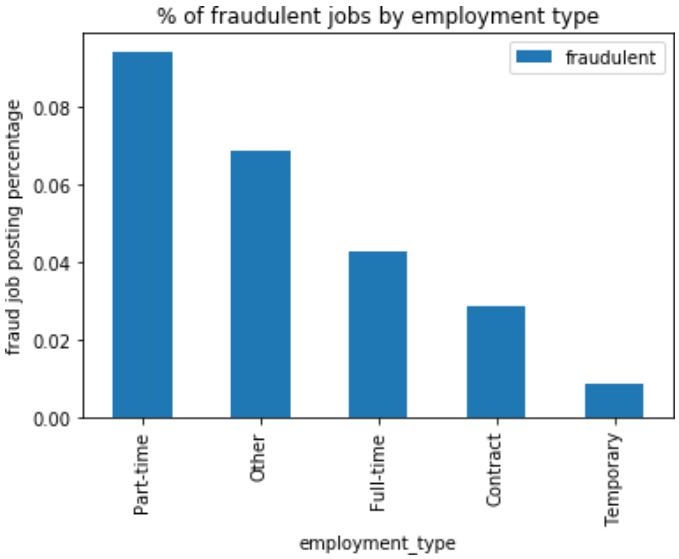

Figure 3.5: Fraudulent jobs by employment type

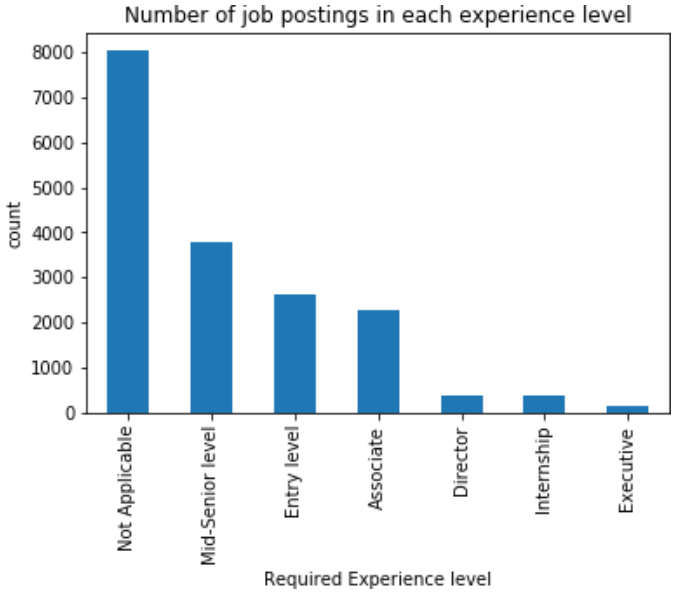

Figure 3.6: Number of job posting in each experience level

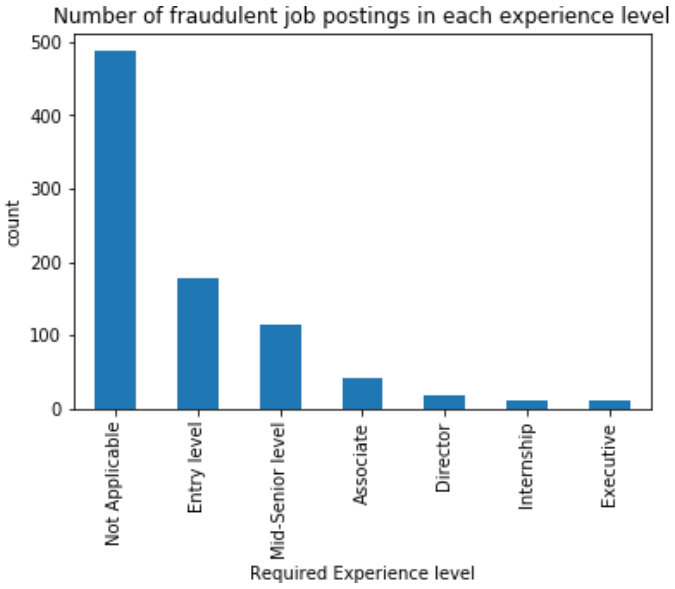

Figure 3.7: Number of fraudulent job postings in each experience level

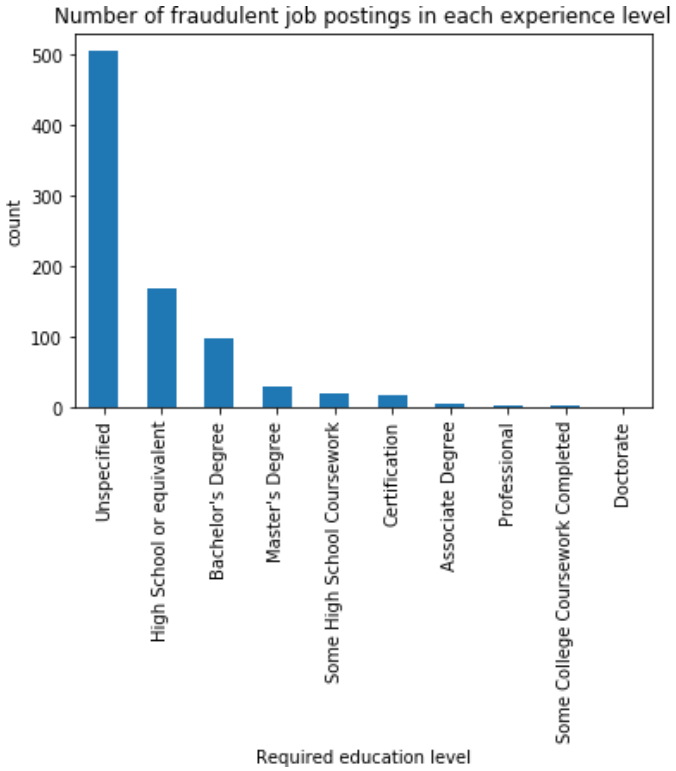

Figure 3.8: Number of fraudulent job postings in each education level

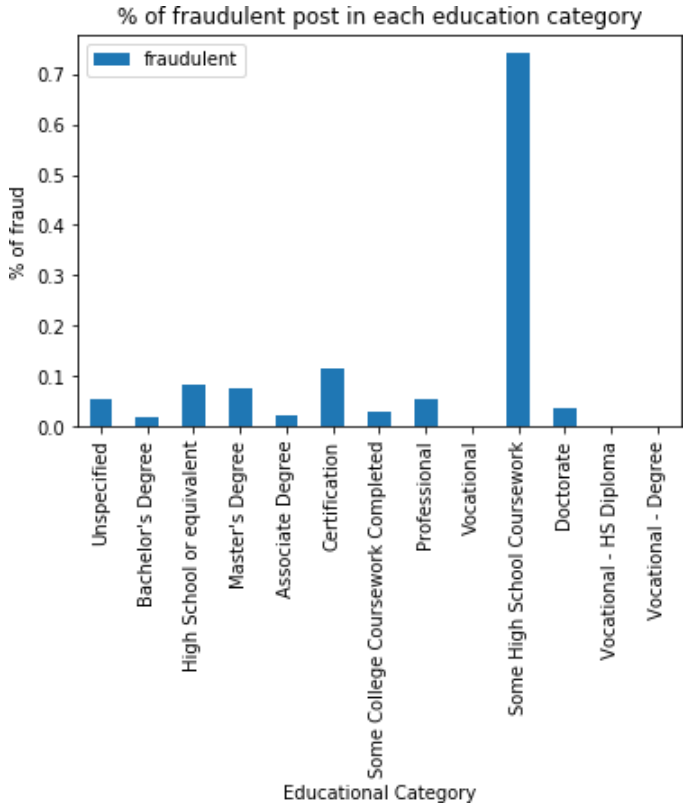

Figure 3.9: Percentage of fraudulent job postings in each education level

Figure 3.10: Word Cloud for fraudulent jobs of description

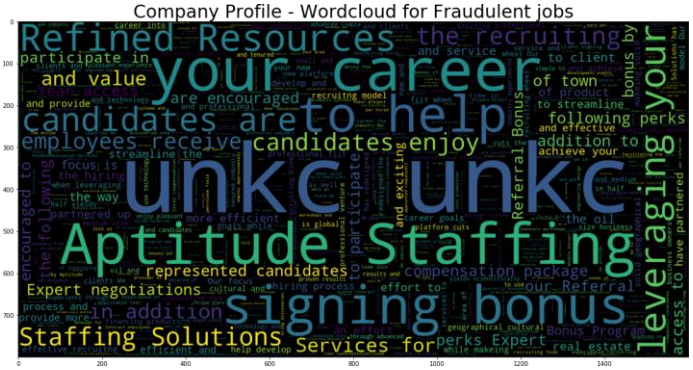

Figure 3.11: Word Cloud for fraudulent jobs of company profile

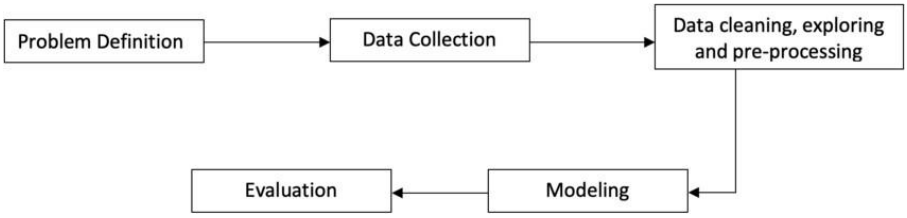

Figure 3.12: Stages of preprocessing method

Chapter 4

Evaluation Metrics

The main motivation to use evaluation metrics to evaluating performance of a model, it is necessary to pragmatic evaluation metrics to predict the evaluation. For this purpose, following metrics are taken in this project which are given below:

4.1 Accuracy

Accuracy : TP + TN / TP + FP + TN + FN

4.2 Precision

”Precision identifies the ratio of correct positive results over the number of positive results predicted by the classifier.”[1]

4.3 Recall

Recall : TP / TP + FN ”It denotes number of correct positive results divided by the number of all relevant samples.”[1]

4.4 F1-Score

F1-Score : 2* Precision * Recall / Precision + Recall The f1-score merge the recall and the precision.

Chapter 5

Machine Learning techniques

5.1 Introduction

In this chapter, we emphasis on the more traditional methods used in natural language processing such as Na¨ıve-Bayes, Decision Trees, Logisitic regression, SVM ,Random Forest. After this we serve advanced methods in this projects which are LSTM and BERT.

5.2 Models

5.2.1 Logistic Regression

Logistic regression forecast and produce the categorical dependent variable. So the output must be a categorical or discrete value. The main aim behind logistic regression is maximize probability of data to gain linear separator between classes in training data. Logistic regression given to sigmoid function which is called logistic function to estimate the probability. .

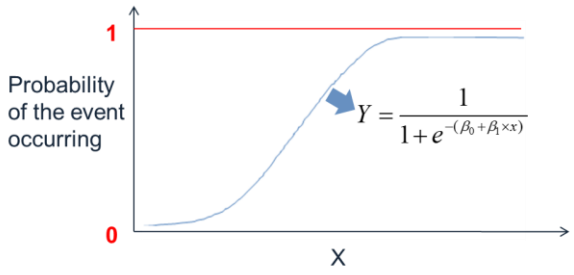

Figure 5.1: Logistic Regression

5.2.2 Naive Bayes

”The Naive Bayes classifier is a supervised classification tool that exploits the concept of Bayes Theorem of Conditional Probability and the decision made by this classifier is quite effective in practice even if its probability estimates are inaccurate and this classifier obtains a very promising result in the following scenario- when the features are independent or features are completely functionally dependent so the accuracy of this classifier is not related to feature dependencies rather than it is the amount of information loss of the class due to the independence assumption is needed to predict the accuracy”[1].

5.2.3 SVM

The main aim behind support vector machine is maximize margin that is the small distance from decision boundary to any training point and make a hyperplane that linearly separates the training data.The point nearest to hyperplane are called support vectors. Decision boundary of SVM is nonlinear unlike logistic regression.SVM transform data into high dimensional space which rely on kernel. SVM works efficient in high dimensional space on when hyperplane divide the points. Conversely, SVM not good with or work well enormous data sets when computational complexity is high or the target classes merged.SVM works well with nonlinear decision boundaries or small amount of data.

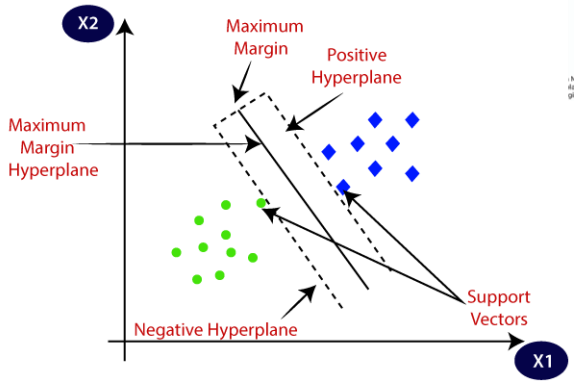

Figure 5.2: SVM

5.2.4 Random Forest

”Random forest (RF) exploits the concept of ensemble learning approach and regression technique applicable for classification based problems and this classifier assimilates several tree-like classifiers which are applied on various sub-samples of the data set and each tree casts its vote to the most appropriate class for the input”[1]. ”Boosting is an efficient technique where several unstable learners are assimilated into a single learner in order to improve accuracy of classification so the boosting technique applies classification algorithm to the reweighted versions of the training data and chooses the weighted majority vote of the sequence of classifiers as well as adaBoost is a good example of boosting technique that produces improved output even when the performance of the weak learners is inadequate”[1].

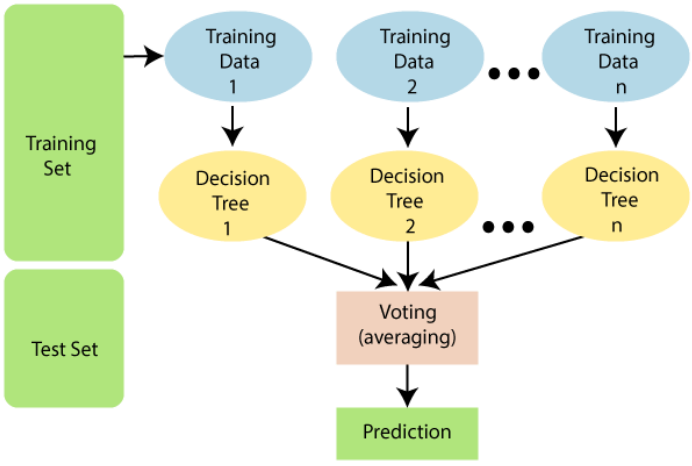

Figure 5.3: RF

5.2.5 Decision Tree

”A Decision Tree (DT) is a classifier that exemplifies the use of tree-like structure and it gains knowledge on classification so each target class is denoted as a leaf node of DT and non-leaf nodes of DT are used as a decision node that indicates certain test then the outcomes of those tests are identified by either of the branches of that decision node.”[1]. Below diagram explains the general structure of a decision tree:

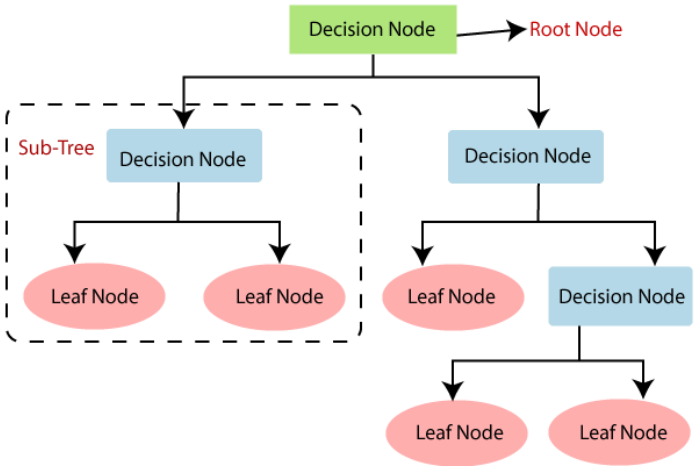

Figure 5.4: Decision Tree

5.2.6 CNN

A Neural Network (CNN) catch multiply of matrix that afford results to embody for training process. This method is defined as convolution and that type of neural network is called a neural network. In NLP, words in a sentence or a fake job project are denoted as word vectors. These word vectors are acclimated for training a neural network. The training is drifting out by establish a kernel size and filters. A CNN can be multi-dimensional.In the case of text classification, a one-dimensional neural network (Conv1D) is acclimated. Conv1D pledge with 1d array denoted as word vectors. In neural network, a filter of same window side renewal through the training data, which at each step multiply the given filter weights and provides an result which is saved in result array. This result array is a feature map of information. In this manner features is identified from the given training data.The size of the filter is define as kernel size and the number of filters define the sequences of features map to be acclimated. In this manner neural network can be employ to gain feature that acquire from the training data.

5.3 Results Analysis of Machine Learning Neural Network Model

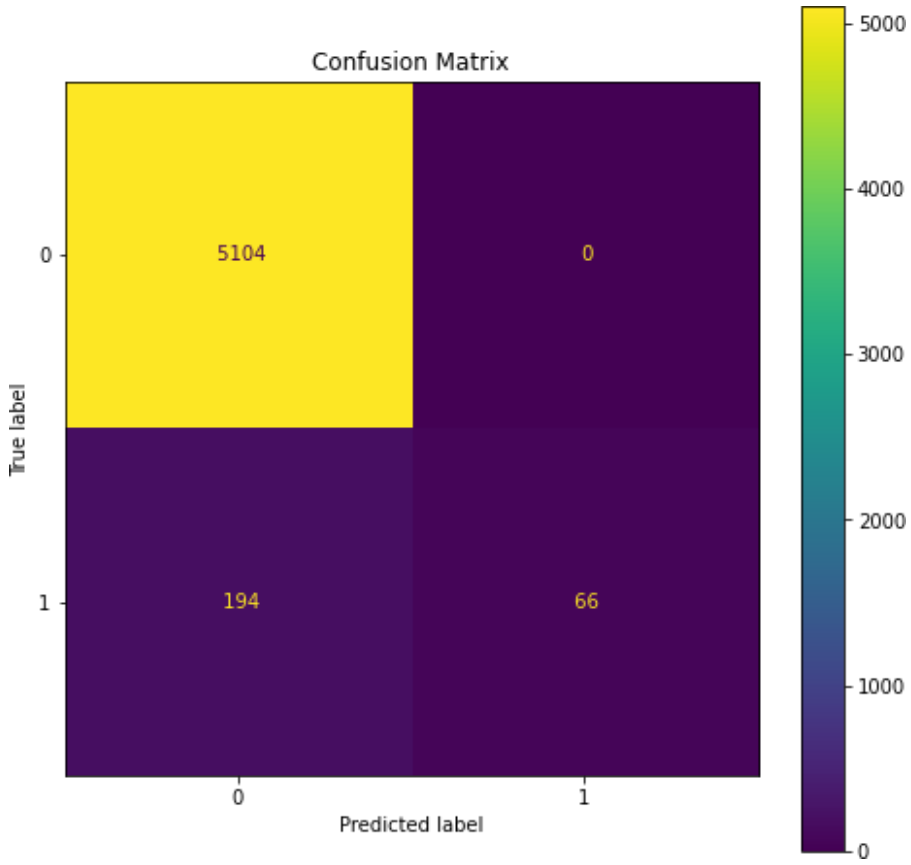

Figure 5.5: Confusion Matrix for Logisitic Regression

Classification Report

	precision	recall	f1-score	support
0	0.96	1.00	0.98	5104
1	1.00	0.25	0.40	260
accuracy			0.96	5364
macro avg	0.98	0.63	0.69	5364
weighted avg	0.97	0.96	0.95	5364

Accuracy: 0.9638329604772558
TNR: 1.0
NPV: 0.25384615384615383

Figure 5.6: Classification Report of Logisitic Regression

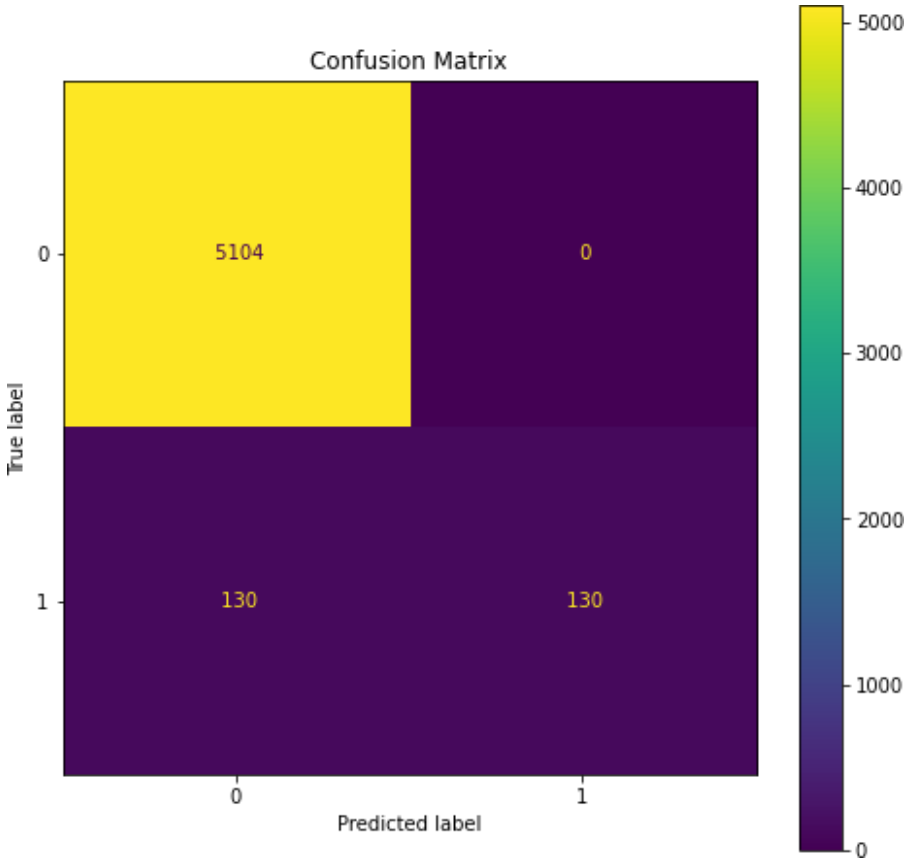

Figure 5.7: Confusion Matrix for Random Forest

```
Classification Report
              precision    recall  f1-score   support

           0       0.98      1.00      0.99      5104
           1       1.00      0.50      0.67       260

    accuracy                           0.98      5364
   macro avg       0.99      0.75      0.83      5364
weighted avg       0.98      0.98      0.97      5364

Accuracy: 0.9757643549589858
TNR: 1.0
NPV: 0.5
```

Figure 5.8: Classification Report of Random Forest

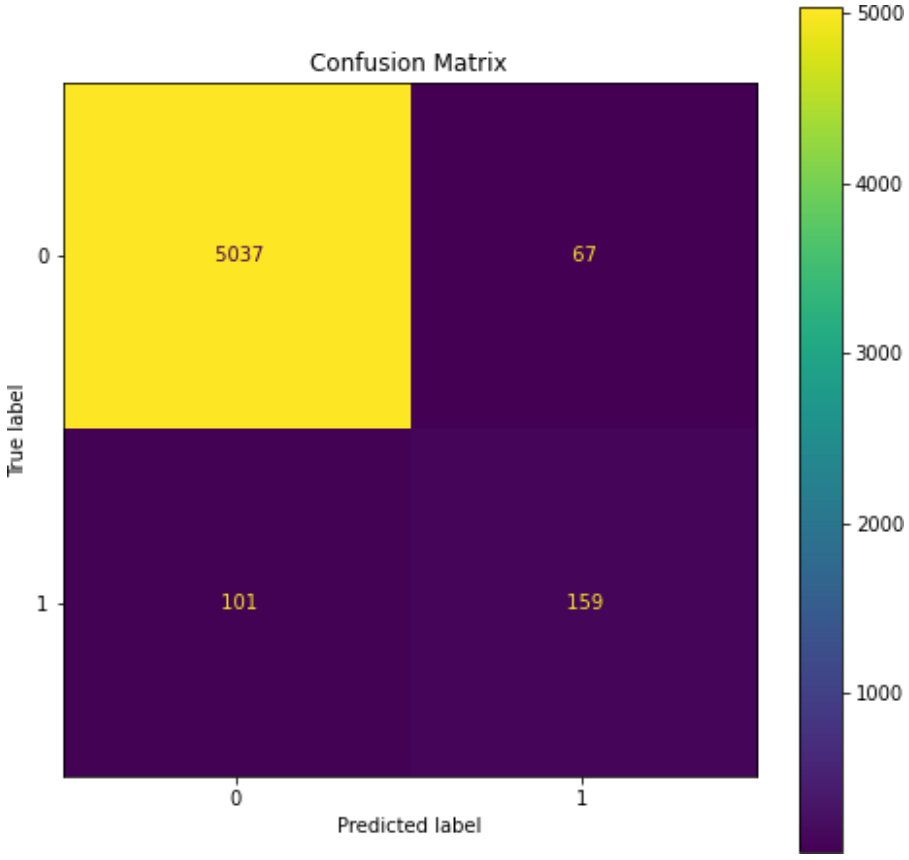

Figure 5.9: Confusion Matrix for Decision Tree

Classification Report

	precision	recall	f1-score	support
0	0.98	0.99	0.98	5104
1	0.70	0.61	0.65	260
accuracy			0.97	5364
macro avg	0.84	0.80	0.82	5364
weighted avg	0.97	0.97	0.97	5364

Accuracy: 0.9686800894854586
TNR: 0.7035398230088495
NPV: 0.6115384615384616

Figure 5.10: Classification Report of Decision Tree

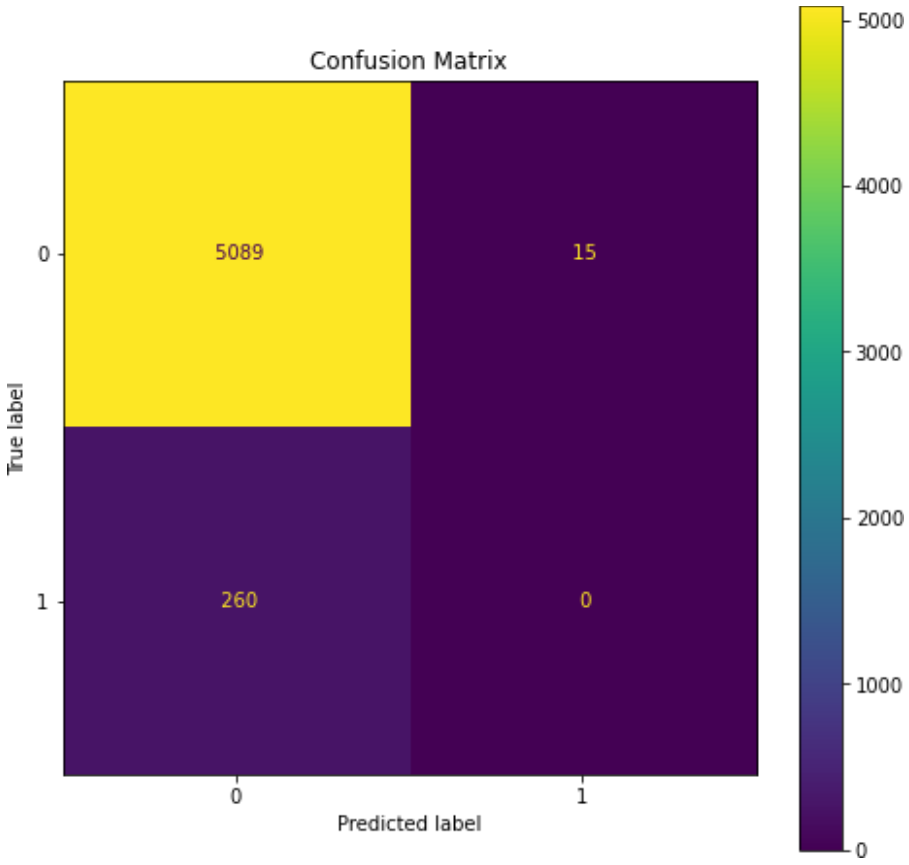

Figure 5.11: Confusion Matrix for Naive Bayes

```
Classification Report
              precision    recall  f1-score   support

           0       0.97      1.00      0.99      5112
           1       1.00      0.43      0.60       252

    accuracy                           0.97      5364
   macro avg       0.99      0.72      0.80      5364
weighted avg       0.97      0.97      0.97      5364

Accuracy: 0.9733407904548844
TNR: 1.0
NPV: 0.43253968253968256
```

Figure 5.12: Classification Report of Naive Bayes

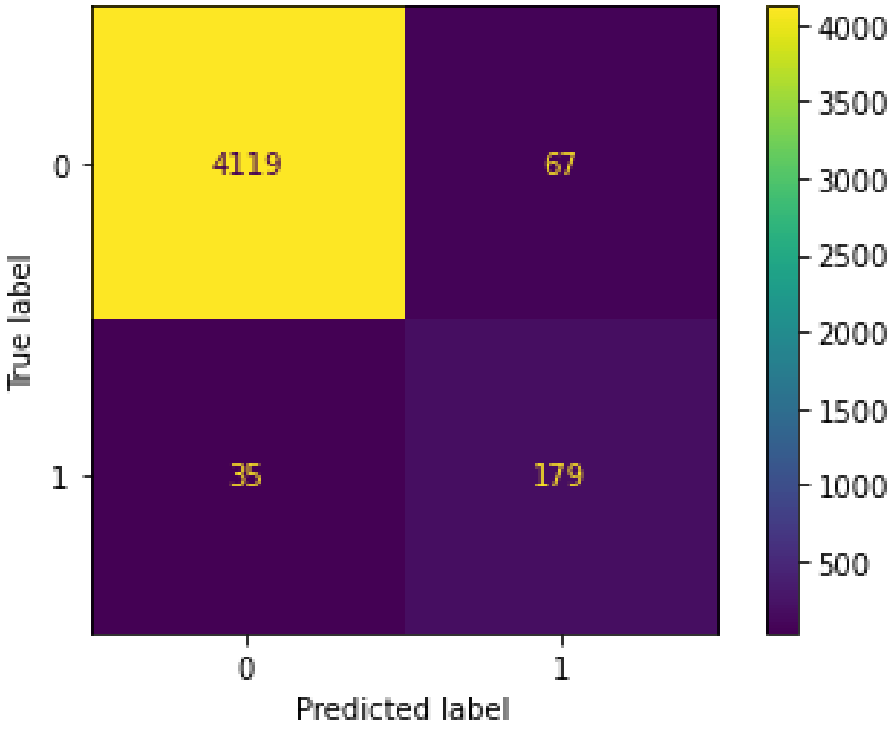

Figure 5.13: Confusion Matrix for Support Vector Machine

Classification Report

	precision	recall	f1-score	support
0	0.95	1.00	0.97	5104
1	0.00	0.00	0.00	260
accuracy			0.95	5364
macro avg	0.48	0.50	0.49	5364
weighted avg	0.91	0.95	0.93	5364

Accuracy: 0.9487322893363161
TNR: 0.0
NPV: 0.0

Figure 5.14: Classification Report of Support Vector Machines

	precision	recall	f1-score	support
0	0.97	1.00	0.98	3531
1	0.97	0.35	0.51	170
accuracy			0.97	3701
macro avg	0.97	0.67	0.75	3701
weighted avg	0.97	0.97	0.96	3701

Figure 5.15: Classification Report of CNN

Chapter 6

Self Attention Method

6.1 Introduction

In this section, we will focus on the deep learning models, the first one being a bidirectional LSTM and the second one an attention layer is added to this LSTM.

6.1.1 LSTM

A RNN is a type of neural network where the hidden state is fed in a loop with the sequential inputs.Long Short Term Memory is a class of recurrent neural network that suited well to temporal or sequential input such as texts. Recurrent Neural Networks not efficient well with long-term dependencies, that is why LSTM have been employ. It comprise of an input gate, an output gate and a forget gate . Following steps involves training and test data into a neural network : i) Tokenize the text data which is convert in word index vectors. ii) Padding text sequence which are of same length of all text vectors. iii) Map embedding vector to every word index then multiply with embedding matrix. iv) Get the result as neural network.

6.2 Results Analysis of LSTM

Here we analyze the result of LSTM Model

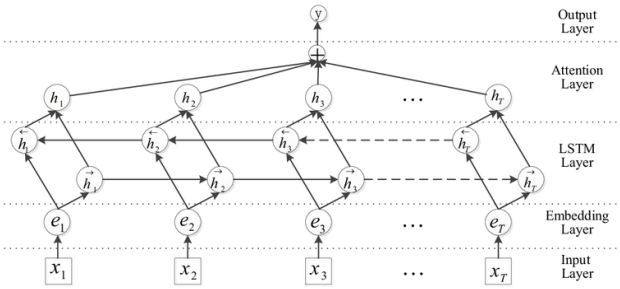

Figure 6.1: Bidirectional-LSTM-model-with-Attentio

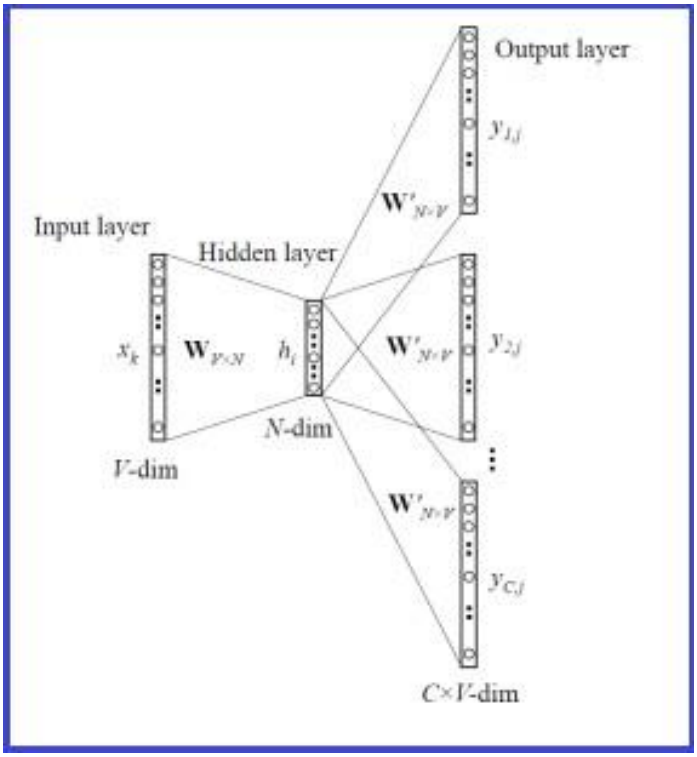

Figure 6.2: skip-gram-model

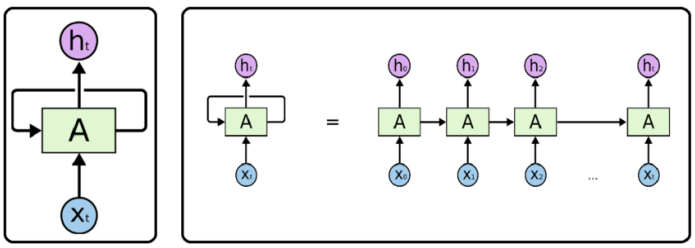

Figure 6.3: Schematic-representation-of-RNN-a-in-rolled-form-b-in-unrolled-form

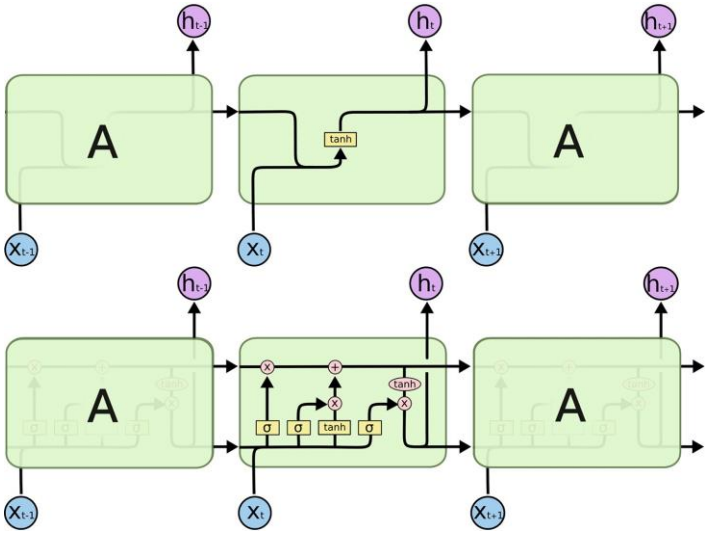

Figure 6.4: LSTM

	precision	recall	f1-score	support
0	0.97	0.99	0.98	3585
1	0.79	0.52	0.63	212
accuracy			0.97	3797
macro avg	0.88	0.76	0.81	3797
weighted avg	0.96	0.97	0.96	3797

Figure 6.5: Classification Report of LSTM

Chapter 7

Introduction

7.1 BERT

BERT denotes for Bidirectional Encoder Representations from Transformers.The pre-trained BERT model can be fine-tuned with just one additional output layer to create state-of-the-art models for a wide range of NLP tasks.It is designed to pre-train deep bidirectional representations from unlabeled text by jointly conditioning on both left and right context.

7.2 Why BERT?

These models are all unidirectional or shallowly bidirectional, BERT is fully bidirectional. BERT gives it incredible accuracy and performance on smaller data sets which solves a huge problem in natural language processing.BERT was made upon recent work and better ideas in pre-training contextual representations including Semi-supervised Sequence Learning ,transformer and other models.

7.3 Parts of Bert

i) Input layer : ”It aims to build an input sequence model via constructing auxiliary sentence and turn the task into a sentence-pair one and after that, the WordPiece embeddings with a 30,000 token vocabulary are used for segmenting the input sequence and the split word pieces are so the position embeddings, word embed-dings and segmentation embeddings for each token are then summed to yield the final input representations”[3].

ii) BERT encoder:”It consists of 12 Transformer block sand 12 self-attention heads by taking an input of a sequence of no more than 512 tokens and outputting the representations of the sequence”[3].

iii) Output layer: ”It consists of a simple softmax classifier on the top of BERT encoder for calculating the conditional probability distributions over pre-defined categorical labels”[3].

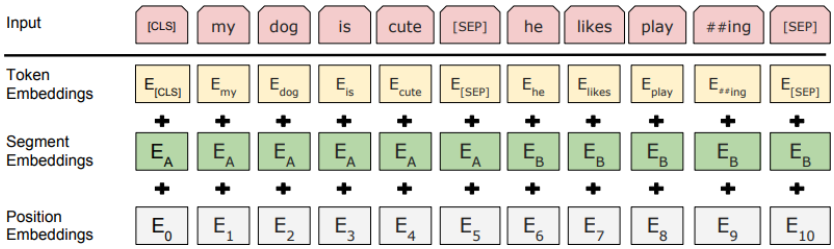

Figure 7.1: Bert Embedding

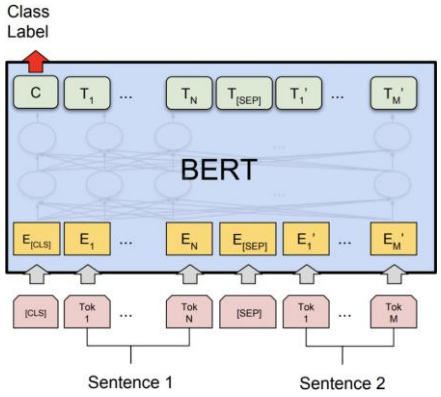

Figure 7.2: BERT Classification

7.4 Result Analysis of BERT Model

Here we provide the classification report of BERT model.

	precision	recall	f1-score	support
0	0.98	1.00	0.99	17014
1	0.99	0.68	0.81	866
accuracy			0.98	17880
macro avg	0.99	0.84	0.90	17880
weighted avg	0.98	0.98	0.98	17880

Figure 7.3: Classification report on bert model

Chapter 8

Result and Conclusion

Various types of models used in this project and applied every models on fake job dataset.So we find that few model work very well like bert and other model like svm does not work very well not give better accuracy. By applying the classifiers we come to know that the BERT Model gives the better accuracy result in identify the fake jobs compared to other models.

8.1 CONCLUSION

In this project we have applied machine learning and deep learning algorithms to classify and detect fake jobs from real jobs in a large dataset of job posts. After analysis these performances, we conclude that having global information of the dataset helps the model in avoid fraud jobs.

8.2 ACKNOWLEDGEMENT

I would like to express my special thanks of gratitude to Professor Manjira Shina who gave me this opportunity to work on this project, also who guided me through the whole course of the project.

Chapter 9

REFERENCES

[1]. Shwani, D., Samir, K.B. (2020). Fake Job Recruitment Detection Using Machine Learning Approach. International Journal of Engineering Trends and Technology (IJETT).

[2]. Shivam Bansal (2020, February). [Real or Fake] FakeJobPosting Prediction,Version 1.Retrieved March 29,2020 https://www.kaggle.com/shivamb/real-or-fakejobposting Prediction.

[3].Yu, Shanshan Jindian, Su Luo, Da. (2019). Improving BERT-Based Text Classification With Auxiliary Sentence and Domain Knowledge. IEEE Access. PP. 1-1. 10.1109/ACCESS.2019.2953990.

[4].Journal of Information Security, 2019, 10, 155-176 http://www.scirp.org/journal/jis ISSN Online: 2153-1242 ISSN Print: 2153-1234.

.

Bibliography