Sentiment Analysis to Determine Employee Job Satisfaction Using Machine Learning Techniques

by

Nazifa Mouli
18201171
Protiva Das
18101382
Munim Bin Muquith
20201228
Aurnab Biswas
19101249
MD Dilshad Kabir Niloy
18101548

A thesis submitted to the Department of Computer Science and Engineering
in partial fulfillment of the requirements for the degree of
B.Sc. in Computer Science

Department of Computer Science and Engineering
Brac University
January 2023

Declaration

It is hereby declared that

1. The thesis submitted is my/our own original work while completing degree at Brac University.
2. The thesis does not contain material previously published or written by a third party, except where this is appropriately cited through full and accurate referencing.
3. The thesis does not contain material which has been accepted, or submitted, for any other degree or diploma at a university or other institution.
4. We have acknowledged all main sources of help.

Student's Full Name & Signature:

Nazifa Mouli

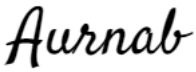
18201171

Protiva Das

18101382

Munim Bin Muquith

20201228



Aurnab Biswas

19101249

MD. Dilshad Kabir Niloy

18101548

Approval

The thesis titled "Sentiment Analysis to Determine Employee Job Satisfaction using Machine Learning Techniques" submitted by

1. Nazifa Mouli (18201171)
2. Protiva Das (18101382)
3. Munim Bin Muquith (20201228)
4. Aurnab Biswas (19101249)
5. MD Dilshad Kabir Niloy (18101548)

of Fall,2022 has been accepted as satisfactory in partial fulfillment of the requirement for the degree of B.Sc. in Computer Science on January 17, 2023.

Examining Committee:

Supervisor:
(Member)

Dewan Ziaul Karim

Lecturer
Department of Computer Science and Engineering
BRAC University

CO-Supervisor:
(Member)

Md Faisal Ahmed

Lecturer
Department of Computer Science and Enginnering
BRAC University

Program Coordinator:
(Member)

Md. Golam Rabiul Alam

Associate Professor
Department of Computer Science and Engineering
Brac University

Head of Department:
(Chair)

Sadia Hamid Kazi, PhD

Chairperson and Associate Professor
School of Data and Science
Department of Computer Science and Engineering
Brac University

Abstract

Over the past three years, the COVID-19 epidemic had a significant impact on the labor market. Employees have been laid off and the majority of them have changed careers. If they can collect more datasets in the future, the researchers will be able to apply fine-tuning approaches to achieve perfect accuracy and precision. Incorporating hybrid models such as optimization techniques, multi-modal models, transfer learning models, hybrid deep learning models, sentiment models, etc. also broadens the scope of this study. These models can employ a variety of learning approaches, such as deep learning or traditional machine learning, and they can use many different types of data, such as text, images, or audio. The corpus was an additional strategy for improvement. These models consider lengthier texts in addition. 10% of US workers who keep their existing jobs are dissatisfied with them. Employee happiness is mostly influenced by business culture, but there are also certain economic and social elements that are interconnected. To ascertain the level of employee satisfaction and associated factors, significant study has been conducted. One of the most popular channels for opinion expression is social media. People now discuss the advantages and disadvantages of their work on the US-based social media site Glassdoor. For this study, total 1,56,428 data has been collected from Glassdoor.First, the data is correctly pre-processed after collection. The understanding of employee work satisfaction is provided by user ratings. For the purpose of making future predictions, the data was divided into binary class dataset and multiclass dataset. Moreover, this data is subjected to machine learning algorithms and deep learning algorithms. The best way to reach the ultimate conclusion is to use Bi-GRU for binary class dataset which has an overall accuracy of 97% and Bert model for multiclass dataset which has an accuracy of 95%.

Keywords: Machine Learning, Naive Bayes, K-Nearest Neighbors (KNN), Deep Learning, Long Short Term Memory(LSTM), Gated Recurrent Unit (GRU), Convolutional Neural Network(CNN), Tokenization, Recall.

Dedication (Optional)

This dissertation is devoted to our adored parents and esteemed teachers.We owe them our thanks.We would not have gotten this far without their cooperation, concern, and support. Many thanks to them.

Acknowledgement

All glory to God, who has enabled us to finish our thesis without any significant setbacks.We appreciate the general direction provided by our advisor Dewan Ziaul Karim Sir and co-advisor MD. Faisal Ahmed Sir and Without their help, we would be unable to complete our project.

Table of Contents

Declaration i

Approval ii

Abstract iv

Dedication v

Acknowledgment vi

Table of Contents vii

List of Figures ix

List of Tables xi

Nomenclature xii

1 Introduction 1
- 1.1 Research Problem . . . 2
- 1.2 Research Objectives . . . 3

2 Background 4
- 2.1 Literature Review . . . 4
- 2.2 Algorithms . . . 10

3 Dataset 12
- 3.1 Binary Class Dataset : . . . 13
- 3.2 Multiclass Dataset: . . . 19

4 Proposed Methodology 26
- 4.1 Data Preprocessing . . . 26
- 4.2 Model Description . . . 27
  - 4.2.1 Binary Classification Models: . . . 27
  - 4.2.2 Multiclass Classification Models: . . . 29

5 Experimentation 31

6 Result Analysis **34**
6.1 Binary Classification Models 35
6.1.1 Deep Learning Approach: 35
6.1.2 Machine Learning Approach 40
6.2 Multiclass Classification Model 45
6.2.1 Deep Learning Approach 45
6.2.2 Machine Learning Approach 50
6.3 Combined Analysis 55
6.4 Discussion 58

7 Conclusion and Future Work **60**
7.1 Conclusion 60
7.2 Future Work 60

Bibliography **63**

List of Figures

3.1 Top 10 Job Status based on Count of Rating . . . 15
3.2 Top 10 Job Status based on Count of Rating . . . 16
3.3 Count of Sentiments . . . 17
3.4 Word Count of Binary Class Dataset . . . 18
3.5 Word Cloud of Binary Class Dataset . . . 18
3.6 Rating based on number of reviews . . . 21
3.7 Top 10 Job Status based on Count of Rating . . . 22
3.8 Count of Sentiments . . . 23
3.9 Word count based on sentiments . . . 24
3.10 Word Cloud of Multi Class Dataset . . . 24

4.1 Architecture of Bi-GRU . . . 29
4.2 Architecture of BERT . . . 30

5.1 Workflow of the experimentation . . . 31
5.2 Accuracy comparison among the best performing models for binaryclass dataset . . . 32
5.3 Accuracy comparison among the best performing models for multiclass dataset . . . 33

6.1 Confusion Matrix of Bi-GRU . . . 35
6.2 Confusion Matrix of BiCuDNNLSTM . . . 36
6.3 Confusion Matrix of Simple GRU . . . 37
6.4 Confusion Matrix of CNN . . . 38
6.5 Confusion Matrix of 1dCNN-BiCuDNNLSTM . . . 39
6.6 Confusion Matrix of Multinomial Naive Bayes . . . 40
6.7 Confusion Matrix of Bernoulli Naive Bayes . . . 41
6.8 Confusion Matrix of Logistic Regression . . . 42
6.9 Confusion Matrix of Support Vector Machine . . . 43
6.10 Confusion Matrix of Random Forest . . . 44
6.11 Confusion Matrix of K-Nearest Neighbors . . . 45
6.12 Confusion Matrix of BERT . . . 46
6.13 Confusion Matrix of Bi-CuDNNLSTM . . . 46
6.14 Confusion Matrix of Bi-GRU . . . 47
6.15 Confusion Matrix of Simple GRU . . . 48
6.16 Confusion Matrix of CNN . . . 48
6.17 Confusion Matrix of 1dCNN-BiCuDNNLSTM . . . 49
6.18 Confusion Matrix of Multinational Naive Bayes . . . 50
6.19 Confusion Matrix of Bernoulli Naive Bayes . . . 51

6.20 Confusion Matrix of Logistic Regression 52
6.21 Confusion Matrix of Support Vector Machine 52
6.22 Confusion Matrix of Random Forest 53
6.23 Confusion Matrix of K-Nearest Neighbors 54

List of Tables

3.1 Dataset for Binary Classification 14
3.2 Rating Based on Number of Reviews 14
3.3 Top 10 Companies Based on Sentiment Count 15
3.4 Top 10 Job Status based on Count of Rating 16
3.5 Count of Sentiments 17
3.6 Word count based on sentiments 18
3.7 Dataset for Multiclass Classification 20
3.8 Rating Based on Number of Reviews 20
3.9 Top 10 Companies Based on Sentiment Count 21
3.10 Top 10 Job Status based on Count of Rating 22
3.11 Count of Sentiments 23
3.12 Word Count based on Sentiments 23

5.1 Accuracy and training time comparison among the best performing algorithms for binary class dataset 32
5.2 Traning time comparison among the best performing algorithms for multi class dataset 33

6.1 Classification Report of Bi-GRU 36
6.2 Classification Report of Bi-CuDNNLSTM 37
6.3 Classification Report of Simple GRU 38
6.4 Classification Report of CNN 38
6.5 Classification Report of 1dCNN-BiCuDNNLSTM 39
6.6 Classification of Multinomial Naive Bayes 41
6.7 Classification of Bernoulli Naive Bayes 41
6.8 Classification of Logistic Regression 42
6.9 Classification of Support Vector Machine 43
6.10 Classification of Random Forest 44
6.11 Classification of K-Nearest Neighbors 45
6.12 Classification Report of Bart 46
6.13 Classification Report of Bi-CuDNNLSTM 47
6.14 Classification Report of Bi-GRU 47
6.15 Classification Report of Simple GRU 48
6.16 Classification Report of CNN 49
6.17 Classification Report of 1dCNN-BiCuDNNLSTM 50
6.18 Classification Report of Multinomial Naive Bayes 51
6.19 Classification Report of Bernoulli Naive Bayes 51
6.20 Classification Report of Logistic Regression 52
6.21 Classification Report of Support Vector Machine 53

6.22 Classification Report of Random Forest 53
6.23 Classification Report of K-Nearest Neighbors 54
6.24 Classification Report of proposed model, other machine learning and deep learning model for Binary Classification 55
6.25 Classification Report of proposed model, other machine learning and deep learning model for Multiclass Classification 57

Chapter 1

Introduction

The goal of sentiment analysis is to ascertain if a statement's representation of feelings or attitudes regarding the text is positive, negative, or neutral. It is utilized for brand reputation, consumer comprehension, and social media sentiment analysis. Sentiment analysis is a sort of text research or text mining that uses natural language processing (NLP), machine learning, and statistics to extract information from a text. People experience a wide spectrum of emotions, including sadness, happiness, interest and disinterest, positivity and negativity, yes or maybe no, and others. A machine can mechanically learn to recognize emotion with the right dataset (in text). A corporation can use sentiment analysis to assess the aspect-based sentiment, the orientation specificity of such a company's product or service, and to identify and examine emotion and intent. Customers use sentiment analysis to monitor brand recognition and make decisions, and businesses use it to evaluate customer satisfaction and assess comments on social media and other information. For a person, sentiment analysis aids in decision-making. For instance, a person looking to purchase a motorbike at great velocity can use an algorithm to determine whether a text discussing fast velocity is good, bad, or neutral. Additionally, when marketers comprehend their clients, they may answer in the greatest approach feasible. When a corporation is aware of the attitudes of clients about a product, it may guarantee better service and deliver the special equipment. This research will leverage a dataset compiled from Glassdoor and apply several machine learning algorithms to the dataset comprising employee evaluations of their jobs. Each word in a sentence will be transformed into phrases or other words using tokenization techniques. To ascertain the sentiments of the audience, the technique examines the order and context of the situation of each phrase. After conducting studies on several sentiment analysis on different platforms. For example- Nasa Zata Dina and Nyoman Juniarta conducted a study on Aspect based Sentiment Analysis of Employee Review Experience carried out in April 2020 had accuracy reaching more than 90This paper is based on supervised machine learning to assist us with a better sentiment analysis technique or result. The collection of our dataset will be discussed first. After that, different pre-processing techniques will be used, such as data cleaning, tokenization, lemmatization, removing punctuation, and stop words. After data preprocessing, different machine learning techniques like Naive Bayes Classifier, K-Neighbors Classifier(KNN), Random Forest, Support Vector Machine (SVM), Logistic Regression and deep learning techniques like Long Short Term Memory(LSTM), Gated Recurrent Unit(GRU), Convolutional Neural Network(CNN) will be used. For evaluation,

different evaluation methods will be used, such as Precision, Recall and F1-Score.

1.1 Research Problem

Finding a job has become more difficult as a result of COVID-19 since there is more competition now than there used to be. Many workers lost their jobs during the epidemic, some of them changed careers, and many more. 21% of people in Japan are dissatisfied with their current job [13]. Many people believe that their employment is a means of support. Company culture is an important factor in influencing employee satisfaction, according to 74% of American workers [13]. The nature of the task, the level of the work, the connection with the supervisor, and a variety of other factors affect an employee's satisfaction. If a worker is happy, an organization will be in business for a very long period. Increasing employee satisfaction may boost morale and contentment with the company, which increases overall organizational effectiveness. Job satisfaction is essential for an employee's professional advancement, physical and emotional well-being, and a host of other factors. Job happiness is crucial for both the employee and the employer. Employee reviews need to be evaluated, and extensive studies are needed to determine work satisfaction and the elements associated with it.
Glassdoor is America's top destination when it comes to looking for a job. According to Internet Glassdoor, there are 114 million employer reviews, CEO approval ratings, salary reports, and other employment-related data in their database. Current or former employees don't have to worry about violating company rules when they post anonymously on Glassdoor. Here, workers are asked to assess the benefits and drawbacks of a specific organization. Since it is anonymous, they are unlikely to find their honest point of view. A human individual, however, is unable to go through all of these reviews to determine whether or not they are favorable. It is necessary to examine all of this data systematically as a result. As a result, it is possible to obtain relevant information by analyzing Glassdoor data. The employee's voice may be heard in online evaluations and the qualities that the employee values most should be examined to determine the level of satisfaction brings. Satisfaction internet statistics do have certain limitations. The quality of this review, for instance, cannot be guaranteed as everyone is free to share their opinion online.
The first step is to get information from Glassdoor. Web scraping is used to gather data, which is then saved in a database. The entirety of the data is presented in human-readable language. All data is processed using Natural Language Processing (NLP) so that the computer can interpret it. Phases of Natural Language Processing include data processing and algorithm development. Data must be pre-processed after it has been collected. Pre-processing transforms raw data into a format that is both convenient and effective. To decide if the data is favorable, negative, or neutral, the emotional approach is employed. Furthermore, the sentimental analysis can be used to identify emotions like joy, rage, despair, or dissatisfaction. This helps to clarify employee satisfaction.Researchers and entrepreneurs have made several attempts to extract relevant information from product reviews Much earlier research emphasized external client happiness or product reviews. In the meantime, only a few studies are interested in surveying internal customer (employee) satisfaction. Employee motivation, performance, and other factors are all closely tied to employee satisfaction. Employee happiness and company performance have previously been

studied, as well as the relationship between organizational culture and workplace satisfaction and other factors. However, there are relatively few studies that use internet reviews and draw conclusions about those variables to examine job satisfaction. Therefore, the goal of this project is to use machine learning techniques like Naive Bayes, Decision Tree, K- Nearest Neighbors and some deep learning methods like Long Short Term Memory(LSTM), Gated Recurrent Unit(GRU), Convolutional Neural Network(CNN) to train and categorize data.

1.2 Research Objectives

The sentiment is a feeling-driven attitude, idea, or judgment extracted for reviews. The method of sentiment analysis can evaluate sentiment polarity like whether a person reacts positively or negatively or neutrally on a topic using the text or dataset. Almost every brand, company, and organization uses sentiment analysis to get an idea of their brand value, product improvement, etc.The Internet is a useful resource for gathering sentiment data, as users may share their ideas on a variety of social networking sites. With the growth of information and communication technology (ICT), sentiment analysis has become a necessary technique as a result of user-created material on internet platforms to disentangle consumer sentiment information about a product or service. Researchers believe that exposing application programming interfaces (APIs) and encouraging data collection on social networking sites would help expand the scope of sentiment analysis. The objective here is to collect employee reviews about their job on the glassdoor.com website and then analyze those to determine whether their attitude towards their job was satisfactory or not using machine learning techniques. So, the objectives are:

- Understanding the application of algorithms like Naive Bayes, Decision tree, LSTM, GRU, word embedding technique, stemming, and overall natural language processing (NLP).
- Sentiment classification of targeted reviews from glassdoor
- Developing a model for evaluating the sentiment with the probable best accuracy
- Suggest improvements to the current model

Chapter 2

Background

2.1 Literature Review

According to evaluations on glassdoor, this document offers the results of a survey of employees' job satisfaction. Over 600,000 companies have ratings and reviews available online. Recently, many individuals have lost their employment, many are dissatisfied with their work conditions and perks, and there has been continued price inflation. Through this study, various businesses, groups, and polarity in the work satisfaction rate will be identified.
ELM (Emitted Light Modulation) is used in the study [12] for aspect-sentiment embeddings. The authors first gathered data from Glassdoor.com, then they afterward tokenized the reviews into sentences using the NLTK Tokenize Package. The raw text is then transformed into review-level summary embeddings. They finally used the ELM algorithm. In their dataset, the workers' incomes, locations, work-life balance, etc. were utilized. The accuracy rate for the study was 95%.
The authors of the research [29] looked outside the office setting and employed browsing and text analytics to establish the indoor work environment. Additionally, it advocated enhancing the working circumstances of those who are most negatively impacted by poor indoor environmental quality. Its findings also identify the most important IEQ factors across a range of industrial sectors and job roles, a finding that can be of great use to businesses, particularly in those areas where IEQ accusations are found to be most widespread.However, automatic information extraction using an iterative cleaning method is not always able to retrieve all the relevant data. Additionally, the repeated cleaning method used for automated information extraction might not always yield accurate results. Future research on the significance of both the soundscape within workspaces, as well as acoustic measurements and investigations, are all actively encouraged. With a 99.99% accuracy rate, it worked utilizing 1,158,706 English employer review outcomes.
The purpose of the study [5] is to offer a fresh perspective on how to examine employees ' job satisfaction and how it relates to organizational performance. This study pulls anonymous employee reviews from glassdoor.com for textual analysis. It looked at the connection between worker happiness and business performance using user text mining. The major focus was on the connection between customer satisfaction and company financial success, and it was discovered that there is a positive correlation between employee contentment and share prices.It inspires other scholars to think about the expansive setting that a text processing technique enables.

However, companies with fewer than 10 reviews are subsequently removed from the list. It gathered a total of 274,061 reviews between 2008 and 2014. Furthermore, the authors were unable to use more sophisticated regression analysis, control the industrial sector, and examine each industry separately. Additionally, it used nine categories and keywords for this study, but it is possible to use additional categories and keywords and construct more sophisticated extraction methods. For its study, three models were employed.

According to the study [34], regardless of the type of investment, businesses with significant operating cash flow growth have a high level of employee satisfaction with "accomplishment" and "promotion." This study used the KH (Koichi Higuchi) coder and sentiment classification to analyze the text of online reviews. It also divided each organization into groups depending on its corporate performance and ran a regression analysis to determine the relationship between employee happiness and corporate performance. Using the "growth rate of operating cash flow" and "ratio of making an investment cash flow to operating cash flow" indicators, this study classified the target companies into four dimensions using the NIKKEI VALUE SEARCH system, a powerful business intelligence tool that offers comprehensive corporate financials, economic data sets, and news and industry reports. The study did identify a link between employee happiness, company operating company's financial quality, and top management attitude toward investment, but it did not fully confirm the mechanism behind this connection. The accuracy percentage for this research article was 81.7%.

The job seeker may pick which firm best meets his needs and abilities with the aid of the study of the research review [17]. To remove any incomplete data, the raw data were processed during the data pre-processing step. Additionally, Stanford POS Tagger was used to remove any terms other than the noun keywords. This work employed user review data that was crawled from Glassdoor and saved in a database. The noun keywords were thereafter divided into each category. Finally, using the aspect-based sentiment analysis, the aspect score was determined. However, it was unable to tell if the evaluation was objective or only intended to harm some businesses. The accuracy percentage for this research article was 92.3%. This paper worked on EB that intelligence [20] can play an essential role in understanding the brand image and sentiments of current and old employees.It means that the company may utilize a variety of methods to learn about how employees feel about the many EVPs (Executive Vice Presidents) they offer and how they feel about the company's brand. The current research offers HR managers information on how to keep up with emerging employer branding tools and tactics, but it does not offer specific development plans for certain EVPs. Additionally, it offers suggestions on how an employer might raise the EB's social or interest value. On the other hand, the study can examine the tendencies in a particular sector where specialized expertise is required, like IT or knowledge management.

For this study's [18] rating and review data, which were accessible from 2012 to 2017 on the employment site Indeed, as well as financial information, which was gathered from the financial records of publicly traded organizations, were used. For 2,738 firms, there were 1.24 million reviews. The results showed a substantial positive association between job rating and financial success as determined by considering all three relevant factors. Although the data were conflicting when looking at relative within-firm impacts, it was discovered that reviews and financial results had a

positive association in the cross-section. This essay has an accuracy percentage of 88%.
A semi-open question appears to be effective for gauging work satisfaction in the paper, according to the study [26]. Furthermore, the study's findings suggest that depending on the context, particular words would naturally have different sentiment ratings.
On the other hand, this study simply created a measure of work satisfaction from the textual replies, demonstrating the imperfect dependability of text measures derived by computer-aided sentiment analysis. The development of more dependable methods to create text measures and get closer to the measurement-error-free sentiment measure might be the focus of future studies.
To study online employee reviews of their employers and find work satisfaction characteristics concealed in the reviews for the research, methods employed the modeling approach, one of the common text mining techniques [13]. It was able to gather important information that helped it decide how to motivate workers to have more job satisfaction than before. Though the reliability test results for the topic modeling technique demonstrate acceptable levels of agreement, compared to humans, it still has relatively low levels of understanding. Because supervised learning algorithms require human supervision and may successfully identify work satisfaction components from the reviews, future studies may utilize them to extract job satisfaction factors from online employee reviews. More data may need to be gathered and used in future studies to provide more broadly applicable results. It achieved a 99.99% accuracy rate.
According to the paper [6], which used these datasets to examine the effects of relative earnings within an occupation and an employer, relative income within an occupation—rather than absolute income or relative income within a firm—indicates the key factor influencing job satisfaction resulting from changes in income. Although their findings are in line with prior research predictions, they differ in terms of how much salary affects work satisfaction. As the precise reasons influencing various components of a job remain unclear, this gave future studies on worker happiness a direction. The accuracy of this study is 99%.
The authors of this study show actual findings on four text classification issues using various iterations of the multinomial naive Bayes classifier. They also discuss a method for enhancing the classifier using locally weighted learning. They demonstrated that some of the adjustments contained in TWCNB may not be required to get the best performance on particular datasets by contrasting traditional multinomial naive Bayes with the recently developed distorted weight-normalized complement naive Bayes classifier (TWCNB) [1]. Additionally, the researchers demonstrated the value of TFIDF conversion and document length normalization. Additionally, it demonstrates how multinomial naive Bayes may do better utilizing least squares learning and how support vector machines might occasionally beat both approaches by a large margin.
Unlabeled documents were implemented by the authors of the work [2], but their usage, in reality, is frequently constrained because of their difficulty to construct, inconsistent prediction results, or high computational cost when employing Multinomial Naive Bayes (MNB). In terms of AUC and accuracy, they attempted to enhance MNB with new data (labeled or unlabeled), which is not the case when combining MNB with Expectation Maximization (EM).

The Multivariate Bernoulli Nave Bayes Classification and the Multinomial Nave Bayes Classification are the two widely used Nave Bayes Text Categorization methods that the authors of [14] used to determine if the sentiment of the news story is positive or negative. The research also tries to determine which of the two methodologies presented works better for the dataset in question.
The goal of the study was to apply a machine learning method called Bernoulli's Naive Bayes Classifier to identify false news. This algorithm is an extended form of Multinomial Naive Bayes and uses predictors that are Boolean variables with values of 0 and 1. Gaussian Naive Bayes was used in earlier investigations [24]. Their suggested technique divides the input information into two categories, 00 for fake news and 10 for real news articles. Additionally, it is noted that the outcomes are improved in comparison to Gaussian Naive Bayes. According to the trials, Bernoulli's Naive Bayes Classifier produces better classification results than Gaussian Naive Bayes. Accuracy, precision, recall, as well as the F1 measure, are all compared. The precision is improved by 10%, precision by 15%, and F1 measure by 6%.
Three binary decision trees, each trained using a deep learning model with a convolution neural network focused on the PyTorch frame, were used by the authors of [27] in an effort to categorize data. The CXR pictures are categorized as normal or abnormal in the first decision tree. The third tree does the same function for COVID-19 whereas the second tree detects the aberrant pictures that contain symptoms of TB. The first and second choice trees' accuracy rates are 98 and 80%, consecutively, while the third decision tree's accuracy rate is 95% on average. Pre-screening patients for triage and quick decision-making may be done using the suggested deep learning-based decision-tree classifier.
When analyzing the effects of wrapper and filter selection methods on classification performance, the authors of [23] attempted to compare their findings. The Correlation Feature Selection (CFS), Information Gain (IG), and Chi-Square (CS) filter techniques have all been taken into consideration. The Best First Search (BFS), Linear Forward Selection (LFS), and Greedy Step Wise Search (GSS) wrapper approaches have all been taken into consideration. The WEKA tool has been used to create a Decision Tree algorithm as a classifier for this investigation.
By using split criteria at each node to separate the employee data among sections with exogenous variables belonging to the same class, decision trees are constructed iteratively. The procedure begins at the decision tree's root node and moves forward by applying split criteria through each non-leaf node to produce homogeneous subsets. However, according to the researchers [25], it is impossible to create pure homogeneous subsets. They suggested using metrics like the GINI index and gain ratio to gauge how good the split was. Additionally, they attempted to compare the GINI index versus knowledge gain empirically. Application of the Index value and Information acquired separately results in the construction of classification models that used a decision tree classifier technique. The models' classification accuracy was estimated utilizing different metrics such as Confusion matrix, Overall accuracy, Per-class accuracy, Recall, and Precision.
In order to evaluate the performance (as analyzed by correctness, precision, and recall) of both the KNN using a large number of parameters, assessed on a variety of real-world data sets, while and without adding different levels of noise. the authors of the paper [10] make an attempt to address this question. The experimental findings demonstrate that the KNN classifier's performance substantially depends

on the distance employed, with considerable performance gaps across different distances.

The authors of [22] determined the location of the nearest neighbor by applying the Euclidean distance formula, as opposed to earlier ways that maximized the euclidean distance by evaluating it with other related formulae to reach perfect results. Their work investigated the calculation of something like the distance measure formula in KNN in comparison both with normalized distance measure, manhattan, plus normalized manhattan in order to acquire the best results or best value when calculating the distance to the nearest neighbor.

After processing the data, [33] the authors are then identified and use a supervised KNN classification technique. The algorithm divides the information into neutral, bad, and positive categories. These seminars speak to the broad public whose Tweets are taken in for examination. They performed sentiment analysis using the LDA machine learning method on this data. It is discovered that the discussion of COVID-19 includes a large amount of dread.

The best categorization technique is the decision tree. The outputs of the decision tree, according to [32] experts, might reveal mistakes brought on by overfitting or noisy data. As a result, the tree could grow overly large and have extra nodes and branches. Pruning is performed within the decision tree to deal with the mistake rate.

Long Short-Term Networks (LSTM) were tested by the researchers [11] for the automatic fake-news detection job. 36 model configurations are tested on two real-world datasets for the binary, end-to-end classification goal of automatically identifying false news. According to the experimental findings, bidirectional LSTM models with generative model word embeddings and, whenever appropriate, an adjusted multiplier factor exhibit strong discriminative ability in automatically classifying fabricated news.

After being present, a hypernymy link between compound entities is detected using the authors' [8] attention-based Bi-GRU-CapsNet model. They have included numerous significant elements in their model. English words or Chinese characters from compounded entities are supplied into the bidirectional gated recurrent units in order to circumvent the out-of-vocabulary issue. To concentrate on the distinctions between two compound entities, an attention mechanism is used.

The authors of [9] suggested a bi-directional hierarchical multi-input and output model-oriented recurrent neural network that takes into account both the lexical and semantic content of emotional expression. Their approach generates sentence and portion of speech representation using two separate Bi-GRU layers. The result of the softmax activation on the section of the speech representation is then considered while paying attention to the lexical information.

The researchers [19] used statistics, individual biographical data, and combined sequential behavior data from a VLE to try and predict students' success in a certain

course as it is being taught. In order to do this, a brand-new RNN-gated GRU combined neural network is developed, in which the data completion method is also used to fill in the missing stream data. This network can fit both static and sequential data. Three different time-series deep neural network algorithms—simple RNN, GRU, and LSTM—are initially taken into account to consider the sequential relationship of learning data.

The authors of [3] concentrated on the assessment of each of the traditional gated architectures for language modeling with voice recognition with a big vocabulary. They assess the highway network, lateral network, LSTM, as well as GRU specifically. Additionally, LSTM and GRU can benefit from the same drive that underlies the highway network. It has recently been suggested to add an extra highway link between the memory cells of neighboring LSTM layers in an expansion that is exclusive to the LSTM.

The authors of the research [28] suggested using a more holistic approach to create a more adequate foundation from which to construct a comprehensive knowledge of DL. In particular, this analysis aims to give a more thorough overview of the most crucial DL components, taking into account any recent advancements. Specifically, their study describes the significance of DL and offers the various DL networks and methodologies. The most common DL network type, convolutional neural networks (CNNs), is next introduced. Their evolution and key characteristics are described, for example, starting with the AlexNet network and ending with the High-Resolution network (HR.Net).

The authors of [31] outlined CNN's tenets and distilled the reasons why they were especially well suited for vegetative remote sensing. The primary section summarized current trends and advancements, taking into account factors like spectral resolution, spatial granularity, various sensor types, modalities of generating reference data, sources of already-existing reference data, and CNN techniques and architectures. The analysis of the documents revealed that CNN may be used to solve a variety of issues, such as the identification of specific plants or the pixel-by-pixel segmentation of vegetation types, and that it performs better than shallow machine learning techniques in multiple studies. According to several studies, the utility of data with the extremely high spatial resolution is notably facilitated by CNN's capacity to exploit spatial patterns. The typical deep learning frameworks' modularity provides for a significant degree of flexibility for the adaptation of architectures, whereby especially multi-modal or multi-temporal applications can benefit.

In order to operationalize OC as a word vector representation, the researchers employed a variety of job characteristics. They [16] utilize text from 650k distinct Glassdoor reviews to confirm this model. They then provide a way for applying their concept to Glassdoor evaluations in order to measure the OC of workers by industry. Additionally, they validate our OC measure using a dataset of 341 employees by offering empirical proof that it contributes to the explanation of job performance. They talked about how their research may be used to develop tools and guide actions aimed at enhancing worker performance.

The researchers [30] look at how businesses react to the greater workplace openness brought on by Glassdoor.com, which gathers and shares employee happiness evaluations. They were using a difference-in-differences design to take advantage of the staggered timing of the first reviews on Glassdoor and discover that after receiving reviews on the site, companies improve their workplace practices as indicated by corporate social responsibility results on human resources and diversity. They discover that this increase is concentrated in companies with poor initial assessments and high labor intensity, which is consistent with businesses upgrading their workplace procedures to maintain their competitiveness in the labor market. The rise is concentrated in businesses with significant institutional ownership, which is consistent with the idea that businesses are making more disclosures regarding workplace policies in order to satisfy regulators.

2.2 Algorithms

Sentiment analysis is identifying the emotional undertone of a text in order to discover whether it is favorable, negative, or neutral. As a starting point for sentiment analysis, the straightforward and understandable model of logistic regression is frequently utilized. The likelihood that a given text belongs to a specific sentiment class is modeled using a logistic function.

Support Vector Machines (SVMs): SVMs are known for their good performance on small and high-dimensional datasets. They use the idea of finding a hyperplane that maximally separates the different sentiment classes in feature space.

Naive Bayes:A notable probabilistic technique for sentiment analysis is naive bayes. It is based on the Bayes theorem and may be applied in a variety of ways, including Bernoulli, Multinomial, and Gaussian Naive Bayes.

Random Forests: As an ensemble approach, random forests mix different decision trees to create a more robust model. They are suitable for sentiment analysis and are especially helpful for handling unbalanced datasets.

KNN: The supervised learning method K-nearest neighbors (KNN) is used for the classification and regression applications. Finding the k-number of data points that are the closest to a particular data point allows the method to identify the object depends on the classifier or mean value of the nearby points. It's a straightforward approach that works well with both normal and quasi data. KNN's key benefit is that it is simple to comprehend and use, but when the data set is big, it may be prohibitively costly.

Bi-CuDNNLSTM: A recurrent neural network (RNN) architecture that makes use of the bidirectional processing technique and the CUDA-accelerated implementation of LSTM (CuDNN) is known as Bidirectional CuDNNLSTM (Bidirectional CuDNN Memory (lstm Memory).

Bi-GRU: The gated recurrent unit (GRU) architecture and the bidirectional processing method are combined in the bidirectional gated recurrent unit (Bi-GRU) kind of recurrent neural network (RNN) architecture.A Bi-GRU, like a standard GRU, employs gating methods to regulate the flow of data through the network, enabling it to choose whether data from earlier period steps to keep or reject.

Simple Gated Recurrent Unit (GRU): A Gated Recurrent Unit (GRU) is an example of a recurrent neural network (RNN) design that use gating methods to regulate the flow of input through the network. As a result, the network may process data sequences more effectively and efficiently by choosing what data from earlier time steps to keep or discard.
CNN: Deep learning neural network architectures known as convolutional neural networks (CNNs) are particularly effective in processing images and videos. CNNs are built to dynamically and efficaciously learn provides advanced of characteristics from input data and are prompted by the organization of the visual cortex.
1dCNN-BiCuDNNLSTM:1dCNN-BiCuDNNLSTM is a form of neural network architecture that combines the advantages of CNNs with bidirectional LSTMs. It consists of a 1D Convolutional Neural Network (1D CNN) and a bidirectional CuDNN LSTM (Bidirectional CuDNN Long Short-Term Memory). 1D CNNs are specialized CNN architectures that are designed to process one-dimensional data, such as time series, audio, or text. They use convolutional layers that are able to scan the input data in one dimension and learn different filters that can detect patterns such as trends, cycles, and anomalies. Transformer-based models (BERT): For sentiment analysis, transformer-based algorithms like BERT may be modified using a dataset of tagged text. These models are highly suited for comprehending the sentiment of the word since they have already been pre-trained on a sizable corpus of text and can capture the text's overall context.
These are just a few examples of machine learning models and deep learning models that can be used for sentiment analysis, and there are many other models available, each with its own strengths and weaknesses. The model used will be determined by the nature of the issue and the characteristics of the data.

Chapter 3

Dataset

DATA DESCRIPTION AND ANALYSIS:

Job satisfaction is a complex concept that can be measured and analyzed in various ways. One common approach is to survey employees and ask them to rate their level of satisfaction with various aspects of their job, such as their workload, compensation, relationships with coworkers and supervisors, and the overall culture of the organization. These surveys can be administered on a regular basis, such as annually or quarterly, to track changes in employee satisfaction over time. Job satisfaction data is self-reported and thus may be susceptible to bias. Furthermore, job satisfaction can change based on a multitude of factors, hence it is important to take a holistic approach and combine the analysis of survey data with other sources of information such as exit interviews, engagement surveys, or performance data. For this study, Raw data is collected for analysis by using the scraping method. The data collected was roughly over 156000 before processing and divided into two significant variations: 1. Binary Classification and 2. Multi-Class Classification.

The collected dataset was imbalanced. Imbalanced datasets are datasets where the classes are not equally represented. This can be problematic for machine learning algorithms because they may be biased toward the majority class. To address this problem, techniques such as undersampling, oversampling, and synthetic data generation can be used to balance the dataset. Additionally, techniques such as cost-sensitive learning and class-weighted algorithms can be employed to mitigate the effects of the imbalance. The information was gathered primarily from IT firms. Before undersampling, the total number of records collected was approximately 156,428. Undersampling is often used when the dataset is imbalanced. Undersampling is a data pre-processing technique that involves reducing the majority class by randomly removing some observations so that the ratio between several classes is balanced. This is done to avoid bias in the model due to the class imbalance and ensure an unbiased accuracy metric. Understanding the data resulted in 1,20,098 records for binary classification and 98,247 for three-layer multiclass classification. Later, model training was conducted for both binary and multiclass classification. Amazon, Apple, Concentrix, Conduent, Google, HCL Technologies, IBM, Infosys, Microsoft, and TATA are the top 10 IT firms from whom the data is gathered. To improve analysis, the five main categories are divided into three sections based on ratings ranging from Negative to Positive.

3.1 Binary Class Dataset :

A binary classification dataset is a collection of data that is used to train a machine learning model for binary classification. It typically consists of two parts: the input data and the corresponding labels.
Depending on the issue, several methods represent the input data. For instance, the input data for classification tasks may be a collection of images, whereas the input data for natural language processing could consist of a collection of text documents. The labels are used to indicate the correct classification for each input data point. In binary classification, there are only two possible labels, such as "positive" and "negative" for sentiment analysis or "spam" and "not spam" for email filtering. A binary classification dataset is often divided into two parts: training and testing data. The training set is used to develop the model, while the test set is used to assess it. This makes it possible to estimate how well the model performs on unknown data. The size of the dataset may vary based on the complexity of the problem, and the number of features in the data. A larger dataset with more features is more robust to overfitting and has better generalization capabilities. But at the same time, it also means more data to preprocess and train, which may be more computationally expensive.
There are several benefits to using a binary classification dataset for training a machine learning model:
Simplicity: Binary classification problems are relatively simple compared to multi-class classification problems, making them a good starting point for developing and testing machine learning models.

Intuitive: Binary classification problems often have clear and intuitive outcomes, making it easy to understand the results of the model and how to improve it.
Widely applicable: Binary classification models can be applied to a wide range of problems, from natural language processing and computer vision to healthcare and finance.

Less data requirement: As the model has to decide between 2 outcomes, it requires less data to train in comparison to multi-class classification.
Easier to interpret: The model results are easy to interpret, as it gives a clear output of one among the two possible outcomes.

Good for online or real-time systems: The models are fast in making decisions which are good for online or real-time systems where time is a constraint.It is worth noting that while binary classification models have many advantages, they may not be suitable for every problem, and more complex multi-class classification models may be needed in some cases.

Table 3.1: Dataset for Binary Classification

Company Name	**Rating**	**Job Status**	**Content**	**Sentiments**
TATA	3	Current Employee	Low salary increment, politics, poor management Good environment, less pressure, good canteen	Positive
Amazon	4	Current Employee, more than 3 years	managers have too much power good pay and stock options	Positive
HCL Technologies	2	Current Employee	Good learning experience but salary is not satisfied Good learning experience but salary is not satisfied and also no benefits	Negative

In this Binary class dataset, which is defined as 0 and 1 for true and false since it highlights two different types of employment reviews, all of the ratings are divided into two halves. The job situation and key word content are shown below, along with the favorable rating of 3 from the TATA group, which is also noted. Similar to Amazon, it has a rating of 4, indicating that the content as well as other factors are favorable. However, the last assessment, which is from HCL Technologies, has a rating of 2, the term "Negative," and numerous tables of contents to support it.

Table 3.2: Rating Based on Number of Reviews

Rating	Count of Rating
1	29492
2	30557
3	20397
4	19032
5	20620
Grand Total	120098

The dataset compiles data from five distinct rating counts with a star rating range of 1 to 5. The ratings 29492, 30557, 20397, 19032, and 20620 are all included in each rating. A total of 120098 ratings have been collected and processed once all processing is complete. The numbers which follow show this.

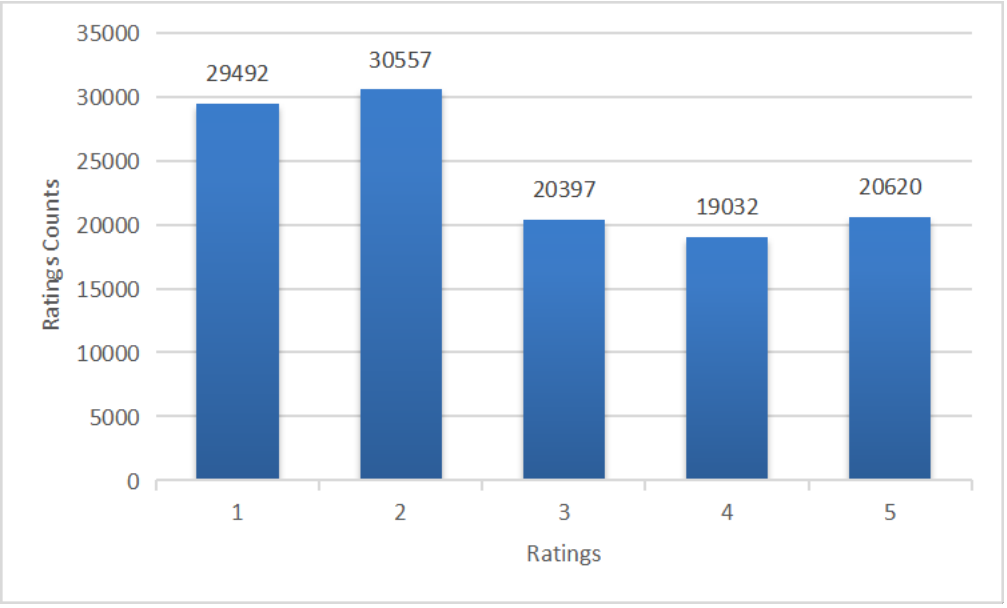


Figure 3.1: Top 10 Job Status based on Count of Rating

Table 3.3: Top 10 Companies Based on Sentiment Count

Company Name	**Negative**	**Positive**	**Grand Total**
Amazon	13000	11767	24767
Apple	847	5971	6818
Concentrix	2328	6047	8375
Conduent	2240	1835	4075
Dell Technologies	2790	1090	3880
HCL Technologies	5317	1672	6989
IBM	7963	0	7963
Tech Mahindra	4366	48	4414
Microsoft	210	3466	3676
TATA	980	2780	3760
Grand Total	40041	34676	74717

Here are some evaluations of companies like Amazon, Apple, Conduent, Dell, HCL Tech, IBM, Microsoft, Tech Mahindra, Infosys, and others, together with their detailed numerical data that includes both negative and positive datasets. Negative data total is 40041, while positive data total is 34676.

Table 3.4: Top 10 Job Status based on Count of Rating

Job Status	Count of Rating
Current Employee	25314
Current Employee, less than 1 year	10398
Current Employee, more than 1 year	14173
Current Employee, more than 10 years	3177
Current Employee, more than 3 years	9490
Current Employee, more than 5 years	5440
Former Employee	16217
Former Employee, less than 1 year	8968
Former Employee, more than 1 year	11273
Former Employee, more than 3 years	6482
Grand Total	110932

An accurate and nuanced knowledge of text sentiment may be obtained by employing a multi-class dataset, which can also lead to more sophisticated sentiment analysis algorithms. To describe the employment evaluation more appropriately and correctly, several job statuses are gathered to specify the job review more accurately and proper. the rating count of the job status is: Current Employee - 25314, Current Employee, less than 1 year - 10398, Current Employee, more than 1 year - 14173, Current Employee, more than 10 years - 3177, Current Employee, more than 3 years - 9490, Current Employee, more than 5 years - 5440, Former Employee - 16217, Former Employee, less than 1 year - 8968, Former Employee, more than 1 year -11273, Former Employee, more than 3 years -6482 and finally in total 110932 collections have been made. As per the figure shows.

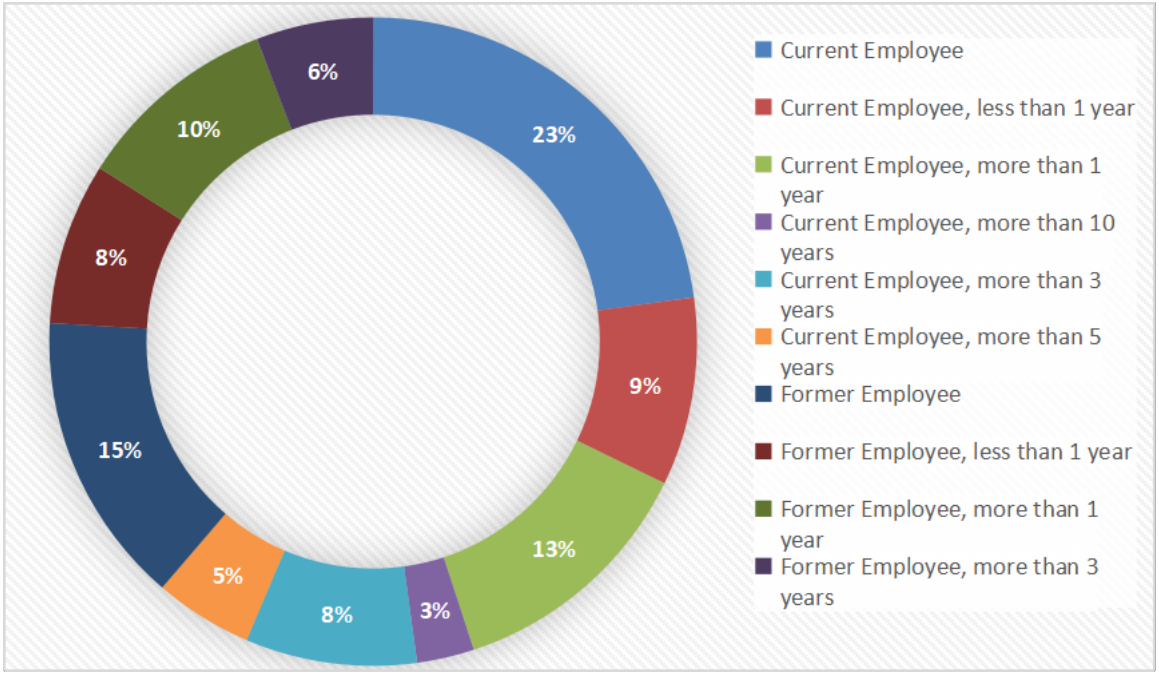


Figure 3.2: Top 10 Job Status based on Count of Rating

The rating count of the job status of multiclass classification is: Current Employee - 23%, Current Employee, less than 1 year - 9%, Current Employee, more than 1 year - 13%, Current Employee, more than 10 years - 3%, Current Employee, more than 3 years - 8%, Current Employee, more than 5 years - 5%, Former Employee - 15%, Former Employee, less than 1 year - 8%, Former Employee, more than 1 year - 10%, Former Employee, more than 3 years - 6% all along in 100% of collections have been made. As per the figure shows.
Negative sentiment analysis data refers to text data that has a negative sentiment or emotion associated with it, such as reviews or comments that express dissatisfaction or disappointment. Positive sentiment analysis data, on the other hand, refers to text data that has a positive sentiment or emotion associated with it, such as reviews or comments that express satisfaction or pleasure. These types of data can be used to train machine learning models for sentiment analysis, which can then be used to automatically classify text data as having positive or negative sentiment. For binary classification, all the data are split into two categories: positive and negative. Each one of them has 60049 review data points, for a total of 120098 as per in the figure shows.

Table 3.5: Count of Sentiments

Sentiments	Count of Rating
Negative	60049
Positive	60049
Grand Total	120098

Figure 3.3: Count of Sentiments

Table 3.6: Word count based on sentiments

Sentiments	Count of Rating
Negative	3463491
Positive	1876811
Grand Total	5340302

As Binary Classification is divided into two segment, each of them has the ratio of consecutively Positive: 35% and Negative: 65%.

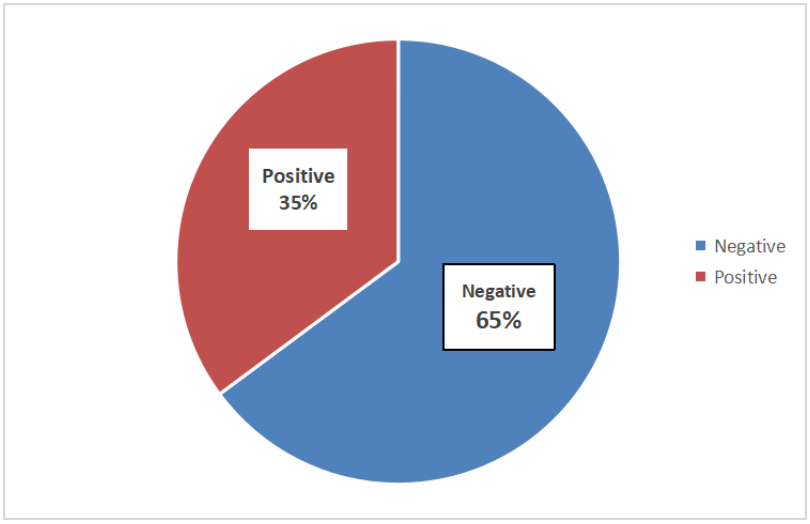


Figure 3.4: Word Count of Binary Class Dataset

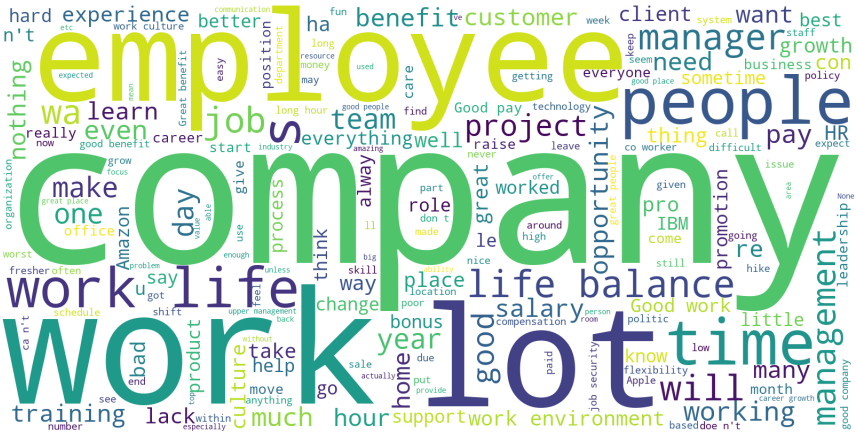


Figure 3.5: Word Cloud of Binary Class Dataset

The most used word clouds are Work, Company, Lot, People, Manager, Need, Pay, Will, Working, Training, Good benefits, and others. Finally, These are how this dataset has been cleaned and used for further model analysis.

3.2 Multiclass Dataset:

A multi-class classification dataset is a dataset used for training machine learning models for classification tasks with more than two classes. Each example or sample in such a dataset includes a collection of attributes that define it, as well as a label that indicates the class it belongs to. The model's objective is to predict the class of a new sample based on its attributes. Examples of multi-class classification problems include image classification with many different object classes, text classification with multiple categories of documents, and speech recognition with different spoken words or phonemes.A multi-class classification dataset for sentiment analysis can be beneficial for several reasons:
Handling nuanced sentiment: Sentiment analysis is not always a binary task of determining whether a text is positive or negative. Using a multi-class dataset can allow for the classification of texts into more nuanced sentiment categories such as positive, neutral, or negative.
Better performance: By providing more classes to classify into, a multi-class dataset can increase the model's ability to capture more subtle differences in sentiment. With more classes, the model can learn more complex relationships between the features of the text and the sentiment it expresses.
Handling mixed sentiment: A text may express multiple sentiments at once, or the sentiment may change over the course of the text. With a multi-class dataset, the model can learn to identify and classify these mixed sentiments.
Better understanding of the problem: When working with a multi-class dataset, it's easier to understand how different sentiments are expressed in language and how they differ from each other. This can be useful for understanding the nuances of the problem, and for debugging and evaluating the performance of the model.
The ratings between 1 and 5 are separated into three portions in this Multiclass class dataset because three different forms of multi-class are being employed. Here, the job condition and key phrase content are displayed along with the Amazon group's negative rating of 2, which is indicated. Similar to Infosys, the rating there is 5, which means that the content and other aspects are positive. The last review comes from Infosys and has a rating of 4, has the word "Positive," and is illustrated with more tables of contents.

Table 3.7: Dataset for Multiclass Classification

Company Name	Rating	Job Status	Content	Sentiments
Amazon	2	Former Employee, less than 1 year	Pay,Benefits,Discounts, Covid Testing, Good Managers Production Hours, Work Life/Sleep Balance	Negative
Infosys	5	Current Employee, more than 8 years	I don't think there is any cons You will get great exposure	Positive
Infosys	4	Current Employee, more than 10 years	Green Card slots are less, Promotion slots are less sometimes Stability, Long term assignments, On par with market salary, Learning programs	Positive

The dataset gathers information from five separate rating counts ranging from 1 to 5 for multiclass categorization. Each rating includes the ratings 16136, 16658, 32794, 15744, and 17050 in order. After all processing, a total of 98382 ratings have already been gathered and processed. This is depicted in the following figures.

Table 3.8: Rating Based on Number of Reviews

Rating	Count of Rating
1	16136
2	16658
3	32794
4	15744
5	17050
Grand Total	98382

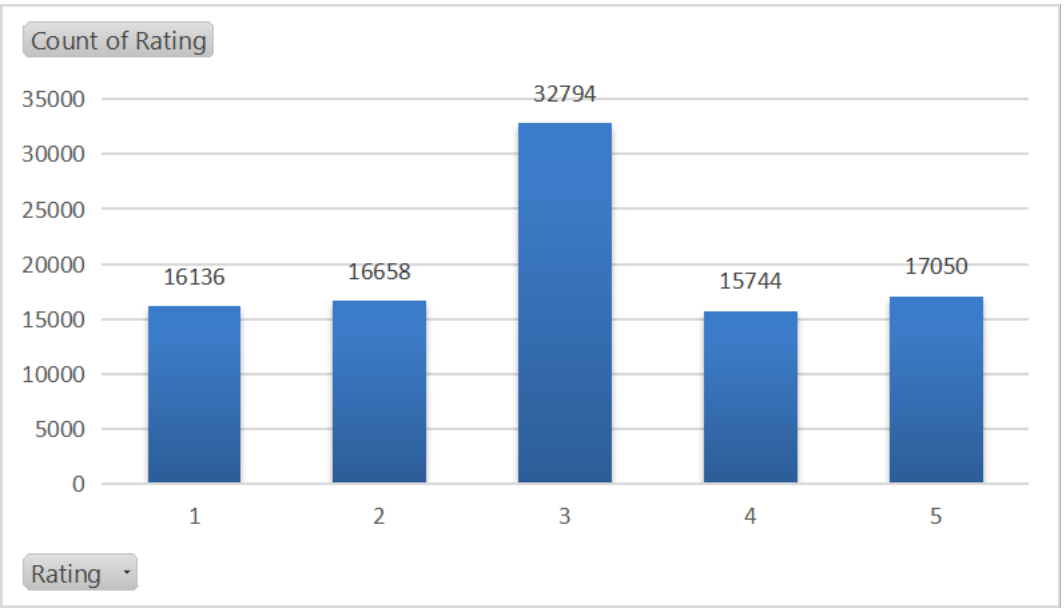


Figure 3.6: Rating based on number of reviews

Here are some reviews of Amazon, Apple, Conduent, Dell, HCL Tech, IBM, Infosys, and some others along with their elaborated data of numbers containing Negative, Neutral, and Positive dataset

Table 3.9: Top 10 Companies Based on Sentiment Count

Company Name	**Negative**	**Neutral**	**Positive**	**Grand Total**
Amazon	7012	4084	7626	18722
Apple	462	1385	4219	6066
Concentrix	1283	2687	3653	7623
Conduent	1218	1353	810	3381
Dell Technologies	1520	1742	0	3262
HCL Technologies	2918	2663	0	5581
IBM	4369	0	0	4369
Infosys	587	1616	1142	3345
Microsoft	118	521	2538	3177
TATA	531	4470	0	5001
Grand Total	20018	20521	19988	60527

Several job statuses are gathered to specify the job review more accurately and proper. the rating count of the job status is: Current Employee - 21571, Current Employee, less than 1 year - 8507, Current Employee, more than 1 year - 11874, Current Employee, more than 10 years - 2556, Current Employee, more than 3

years - 8127, Current Employee, more than 5 years - 4680, Former Employee - 12889, Former Employee, less than 1 year - 6645, Former Employee, more than 1 year - 8870, Former Employee, more than 3 years - 5282 and finally in total 91001 collections have been made. As per the figure shows.

Table 3.10: Top 10 Job Status based on Count of Rating

Job Status	Count of Rating
Current Employee	21571
Current Employee, less than 1 year	8507
Current Employee, more than 1 year	11874
Current Employee, more than 10 years	2556
Current Employee, more than 3 years	8127
Current Employee, more than 5 years	4680
Former Employee	12889
Former Employee, less than 1 year	6645
Former Employee, more than 1 year	8870
Former Employee, more than 3 years	5282
Grand Total	91001

The rating count of the job status of multiclass classification is: Current Employee - 24%, Current Employee, less than 1 year - 9%, Current Employee, more than 1 year - 13%, Current Employee, more than 10 years - 3%, Current Employee, more than 3 years - 9%, Current Employee, more than 5 years - 5%, Former Employee - 14%, Former Employee, less than 1 year - 7%, Former Employee, more than 1 year - 10%, Former Employee, more than 3 years - 6% all along in 100% of collections have been made. As per the figure shows.

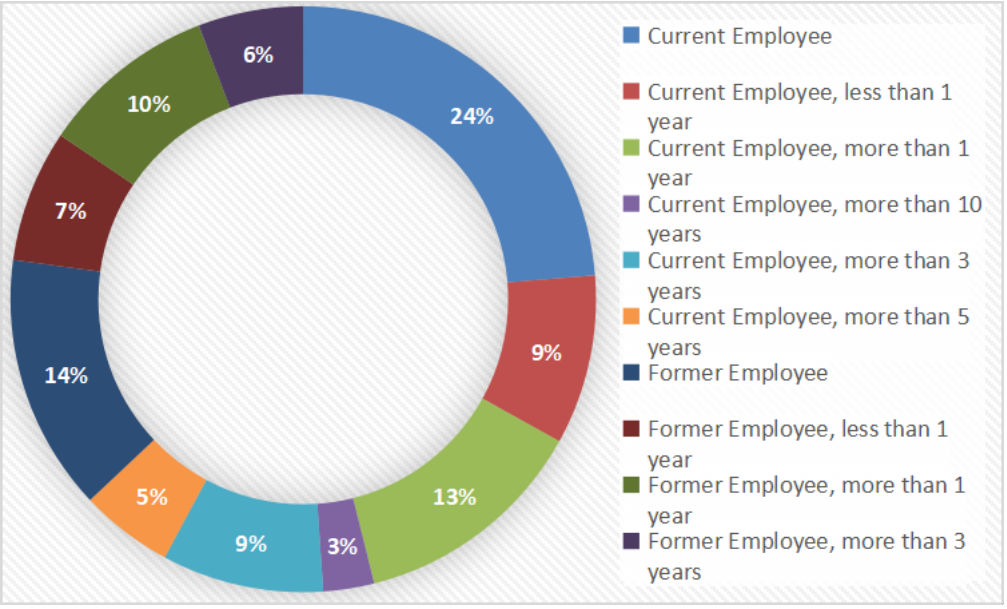


Figure 3.7: Top 10 Job Status based on Count of Rating

Table 3.11: Count of Sentiments

Sentiments	Count of Rating
Negative	32794
Neutral	32794
Positive	32794
Grand Total	98382

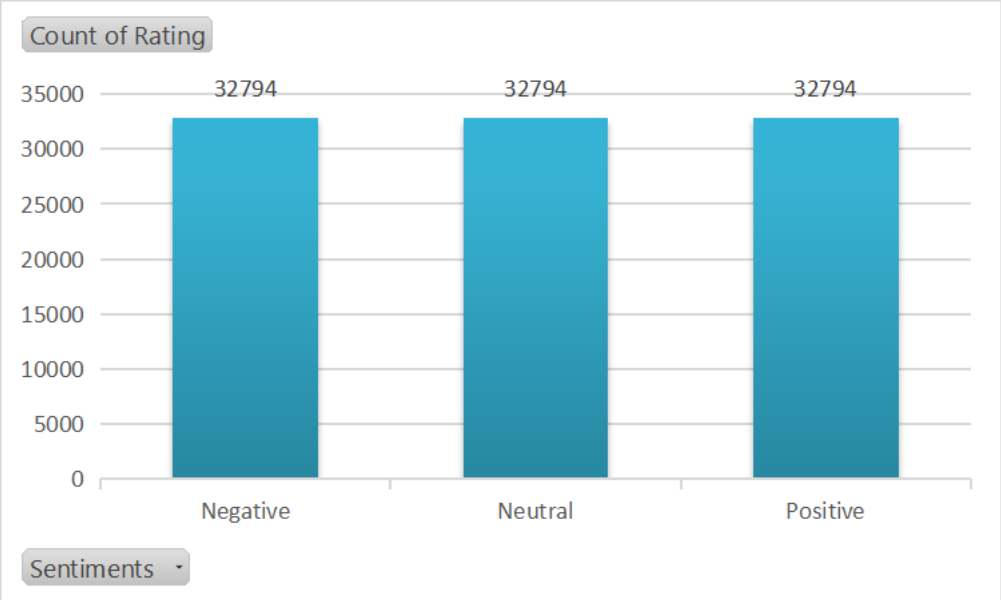


Figure 3.8: Count of Sentiments

Table 3.12: Word Count based on Sentiments

Sentiments	Sum of number of words
Negative	1906791
Neutral	1048725
Positive	1009845
Grand Total	3965361

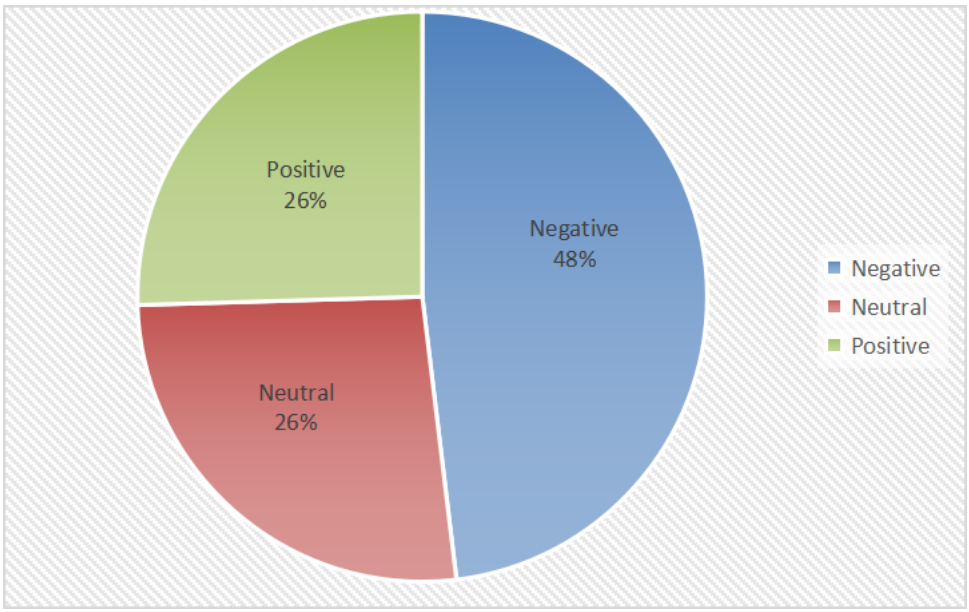


Figure 3.9: Word count based on sentiments

Among all as the rating is divided into three segment each segment like Positive, Neutral and Negative has the ratio of 26%, 26% and 48% consecutively. The word cloud is divided mainly into two main categories such as pros and cons. The words that are mostly used in Positive and Very Positive and sometimes Neutral reviews are collected together.

Figure 3.10: Word Cloud of Multi Class Dataset

Therefore, the researchers collected the dataset on their own using data scraping. In this code, BeautifulSoup is used to collect the data from the companies.
This dataset also contains nobility in such a way that,

- This dataset is completely new, raw, and collected by the researchers..
- This dataset is categorized and labeled by using AI applications to make more impactful and accurate outcomes.
- This dataset is cleaned and scrubbed. The errors, duplicates, and irrelevant data are identified from this dataset and fixed properly.
- No Human Generated Decision has been taken for labeling, rating, or any value of this dataset. This data-collecting and cleaning procedure are fully automated.
- All the invalid reviews are removed from this dataset
- The researchers used the voting system to come up with a combined answer or decision for removing irrelevant and invalid reviews.

In such a way, this dataset represents new, raw, and nobility that this research only uses for the first time.
For the analysis and pre-processing of the data, there are some consequential processes are used in this research. Firstly, the process started by importing the dependencies such as NumPy, panda, matplotlib, seaborn, etc. Later on, stop words, punkt, wordnet, omw-1.4 are downloaded. Secondly, the data has been loaded by the researchers in a way where these key points are maintained which are: Company name, Given Rating, Job status, Pros, and Cons. Thirdly, the data cleaning part has been started where the null values are removed as a column. In the first checking, 3 in pros and cons null values were found. After removing the null values, in the second search, there are all values without null values. So, in that way, this is a user-friendly datasheet where the researchers can work. Fourthly, the datasheet has been copied to excel and the analysis part started to occur.

Chapter 4

Proposed Methodology

The proposed method starts with peparing data and creating an embedding layer for deep learning models. Bi-GRU was used for the binary dataset and Bert model for the multiclass dataset.

4.1 Data Preprocessing

Text cleaning or text pre-processing is a crucial step when working with text in Natural Language Processing (NLP). Because real-life human-written text frequently contains words with wrong spellings, short words, special symbols, emoticons, and so on. Therefore, the authors must clean this type of noisy text data before feeding it to the machine learning model.
For implementing Machine Learning Algorithms and Deep Learning Algorithms, the authors used Remove Punctuations, Tokenization, Removing Stop Words and Lemmatization.

A. Remove Punctuations:

The list of punctuation that will be disregarded by the writers must be determined based on the use case.They will succeed in their goal of eliminating punctuation from the text since the Python string module provides a list of punctuation.One of the finest tech businesses in the world, for instance, is Google. The dataset reads, "Google one of the finest IT companies in the world," after the punctuation has been removed.

.

B. Tokenization:

A statement is tokenized when it is broken up into individual words.A major divider could be appropriate here.However, a separator won't separate abbreviations with spaces between the letters or special characters, as U.A.R.T.The challenges become harder as you add additional languages.The bulk of these problems may be resolved with the nltk library.The normalization and cleaning operations use the tokens that the tokenize module creates as input.In order for machine learning models to understand a text string, it may also convert it to numerical data. The authors

also converted uppercase to lowercase at the same time they performed tokenization.

C. Remove Stop Words:

Stop word removal is a common pre-processing step used in a variety of NLP applications. The primary concept is to exclude terms that appear frequently in all of the texts in the corpus. Pronouns and articles are regularly used to classify stop words. For Example, "I like reading, so I read" this will be converted to "like", "reading", or "read". These keywords have low importance in certain Natural language processing tasks such as information extraction and classification. However, in other cases, stop-word removal may not make a significant effect.

D. Lemmatization:

Lemmatization produces normalization via the study of word morphology and vocabulary. The goal of lemmatization is to restore the lemma word's basic form by deleting only inflectional endings.Despite being slower than stemming, it is a considerably more sophisticated and powerful text analysis method.It aims to preserve the structural relationships between the words.We also utilized the WordNetLemmitizer() method since it was the earliest and most widely used.

All the above techniques has been used for deep learning models. While training, the deep learning models learn to extract useful information from the input data and convert it into a format that can be used for the task at hand.

4.2 Model Description

Deep learning models are known for their notable work in image and speech recognition, natural language processing, and game playing. They have also been used in various industries such as healthcare, finance, and transportation. One of the key advantages of deep learning models is their ability to learn hierarchical representations of data, where the model learns to extract increasingly complex features at each layer. Here Bi-GRU is used to detect sentiments from binaryclass dataset and Bert model is used to detect sentiments from multiclass dataset. Finally, to increase overall accuracy tuning and optimization techniques are used.

4.2.1 Binary Classification Models:

Binary deep learning models are neural network designs intended to output binary values, often 0 or 1, representing the presence or absence of a certain feature or class in the input data.Frequently, these models are used for binary classification problems, where the objective is to categorize an input into one of two classes.Various designs, including feedforward neural networks, convolutional neural networks, and recurrent neural networks, may be used to create binary deep learning models.They use sigmoid or other activation functions with binary output.Using a binary cross-entropy loss function, which assesses the dissimilarity between the predicted prob-

ability and the actual binary label, is a typical method for training a binary deep learning network.In applications such as picture classification, voice recognition, and natural language processing, where the objective is to recognize the presence or absence of specified characteristics or classes in the input data, these models are often used.

GloVe Embedding Layer: Words may be represented as dense vectors in a high-dimensional space using GloVe (Global Vectors for Word Representation), a pre-trained word embedding approach. In order to aid in natural language processing tasks like translation, text classification, and sentiment analysis, these vectors are educated to capture the meaning and context of words. These vector representations of the input words, called GloVes, are then sent along to subsequent layers of the network as input. Combining the GloVe embedding layer with various neural network designs is possible. These include feedforward neural networks, convolutional neural networks, and recurrent neural networks.The GloVe algorithm's ability to learn dense vector representations of text that capture associations between words and their use in multiple contexts is promising for a wide range of NLP applications.Using a pre-trained GloVe embedding layer may boost a neural network's efficiency by giving it a more precise representation of the input words. In this paper's case, we employed Glove to generate the weight matrix for the deep learning models' embedding layer.
The GloVe algorithm's ability to learn dense vector representations of text that capture associations between words and their use in multiple contexts is promising for a wide range of NLP applications. Using a pre-trained GloVe embedding layer may boost a neural network's efficiency by giving it a more precise representation of the input words. In this paper's case, we employed Glove to generate the weight matrix for the deep learning models' embedding layer.

Bi-GRU Classifier: To perform binary classification tasks, recurrent neural networks (RNNs) may be constructed using Bi-GRUs (bidirectional gated recurrent units). An input is processed by two GRUs in a Bi-GRU, one in the forward direction and one in the reverse direction.The final classification decision is made by combining the output of the two GRUs and running it through a fully linked layer.The Bi-capacity GRU's to categorize information appropriately may be enhanced by its ability to evaluate previous and future context for the input. Bi-GRU has a precision of 0.97, or 97%.

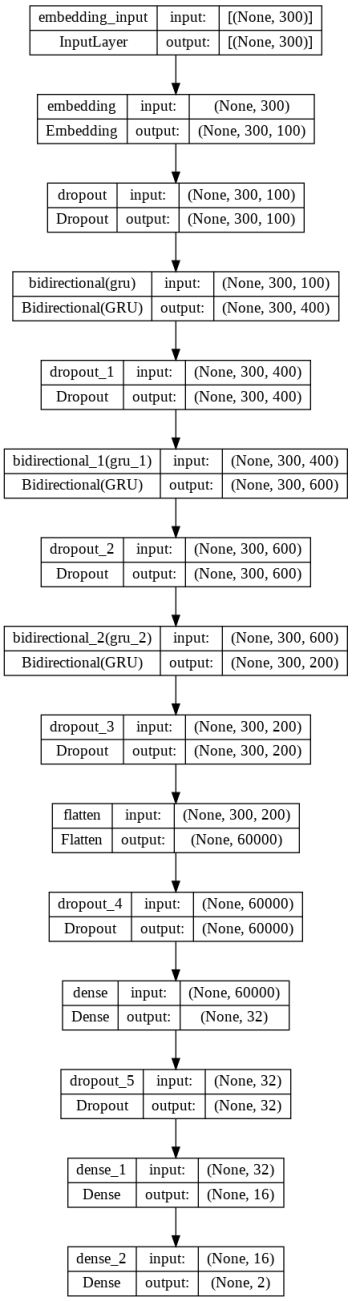


Figure 4.1: Architecture of Bi-GRU

4.2.2 Multiclass Classification Models:

Multiclass deep learning models are a form of machine learning models intended to categorize and manage many classes.These models, which are based on neural networks, are able to learn and generalize from vast volumes of data.Traditional machine learning techniques may struggle with challenges involving a high number of classes.Multiclass classification problems often use deep learning models, such as convolutional neural networks (CNNs), recurrent neural networks (RNNs), and long short-term memory (LSTM) networks.These models may be taught intricate interactions between inputs and outcomes using big datasets.They may also increase

their performance by using pre-trained models and transfer learning.

BERT Embedding Layer:

Natural language processing applications such as text categorization, named entity identification, and question answering may all benefit from the BERT (Bidirectional Encoder Representations from Transformers) embedding layer, a pre-trained neural network layer. [15] Because it creates deep, contextualized representations of words, phrases, and sentences, the BERT embedding layer is crucial. What's more, BERT is a bidirectional model, so it considers the words to the left and right of each one as well. We used a multiclass dataset to fine-tune it for this research.

BERT

Google has created a pre-trained neural network model called BERT (Bidirectional Encoder Representations from Transformers). Text containing both upper- and lowercase characters may be processed using a multi-cased BERT model, a version of the BERT model. BERT has a precision of 0.95, or 95%.

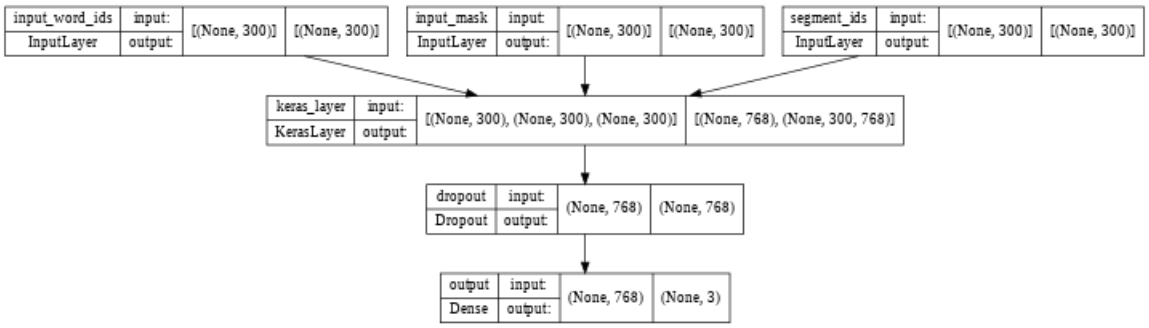


Figure 4.2: Architecture of BERT

Relu:

ReLU introduces non-linearity into the model, which is necessary for the model to be able to learn complex, non-linear decision boundaries. [7] Therefore, in this paper Relu was used to create fully connected layer. In fully connected layer we need complex, non-linear decision boundaries

SoftMax:

The softmax is a activation function used in the final layer of a neural network, the output layer. [4] It is used to transform the outputs of the neural network into probabilities In this study, softmax function is used at the output layer for both binary classification and multiclass classification.

Dropout:

Deep learning models frequently employ the regularization method known as dropout to reduce overfitting.[21] During each training iteration, a certain proportion of the network's neurons are randomly "dropped out" or set to zero.

Chapter 5

Experimentation

The step-by-step procedure of the experimentation conducted is represented in the following workflow:

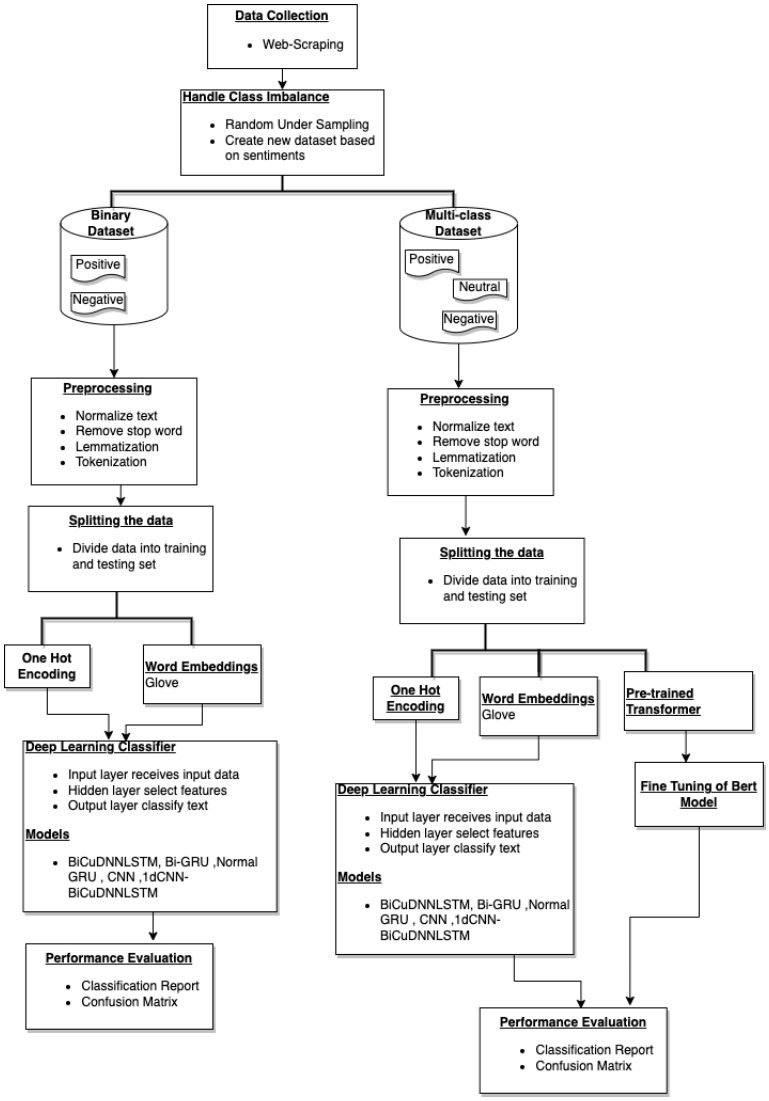


Figure 5.1: Workflow of the experimentation

Google collab was used to train all the models. Colaboratory is a Google research initiative that was created to contribute to the study of Artificial Intelligence. In

this study, data was collected from Glassdoor. Then class imbalance was handled by using random under sampling to create new datasets based on sentiments. The binary class dataset consists of 1,20,098 rows and 5 columns in total. Moreover, in the multiclass dataset there were 98,382 rows and 5 columns. Data processing was implemented in both datasets to improve overall accuracy. In order to understand data distribution, data visualization was performed in both datasets. Then, in a ratio of 80:20 both of the datasets were divided to create train and test data.
For the binary class dataset, Bi-GRU classifier was trained to generate predictions. 20 epochs were used to train this model. The Bi-GRU model has a accuracy of 97%. The training time of Bi-GRU is 7051.69 sec. Bi-CuDNNLSTM model has a testing accuracy of 97% and training time of 7142.32 sec. Bi-CuDNNLSTM and the proposed model have similar testing accuracy. However, Bi-CuDNNLSTM training time is more than Bi-GRU. Moreover, Simple GRU has a accuracy of 96% and a training time of 3465.25 sec, 1dCNN-BiCuDNNLSTM model has a accuracy of 96% and a training time of 1224.53 sec.From this table it can be concluded that Bi-GRU gives better accuracy compared to rest. The accuracy on the training set and the validation set are similar and reasonably high. Finally, this can be concluded that the proposed model has good generalization and it is able to learn the underlying patterns in the data. The table below shows the training accuracy, testing accuracy and training time of among all the best performing models.

Table 5.1: Accuracy and training time comparison among the best performing algorithms for binary class dataset

Algorithms	Training Accuracy	Testing Accuracy	Training time
Bi-GRU	97%	97%	7051.69 s
Bi-CuDNNLSTM	99%	97%	7142.32 s
Simple GRU	97%	96%	3465.25 s
1dCNN-BiCuDNNLSTM	97%	95%	1224.53 s

The bar chart for representing the accuracy and training time of best performing models is given below:

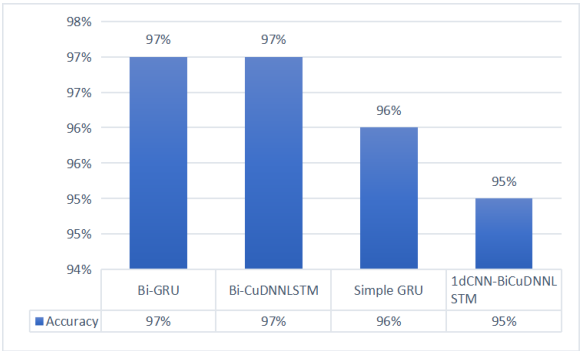


Figure 5.2: Accuracy comparison among the best performing models for binaryclass dataset

For multi class dataset, BERT classifier was trained to generate prediction. 3 epochs were used to train this model. The BERT model has accuracy of 95% and a training time of 17875.34 sec. Bi-CuDNNLSTM model has an accuracy of 97% and a training time of 3145.24 sec. Bi-GRU model has an overall accuracy of 97% and the training time of Bi-GRU was 2853.49 sec. Moreover, accuracy of simple GRU is 96% and a training time of 1475.05 sec, 1dCNN-BiCuDNNLSTM model has an accuracy of 96% and a training time of 866.78 sec. From this table it can be concluded that the BERT model gives better accuracy compared to rest. The table below shows the training accuracy, testing accuracy and training time of among all the best performing models. S

Table 5.2: Traning time comparison among the best performing algorithms for multi class dataset

Algorithms	Accuracy	Training time
Bert	95%	17875.34s
Bi-CuDNNLSTM	92%	3145.24s
Bi-GRU	91%	2853.49s
Simple GRU	91%	1475.05s
1dCNN-BiCuDNNLSTM	89%	866.78s

The bar chart for representing the accuracy and training time of best performing models is given below:

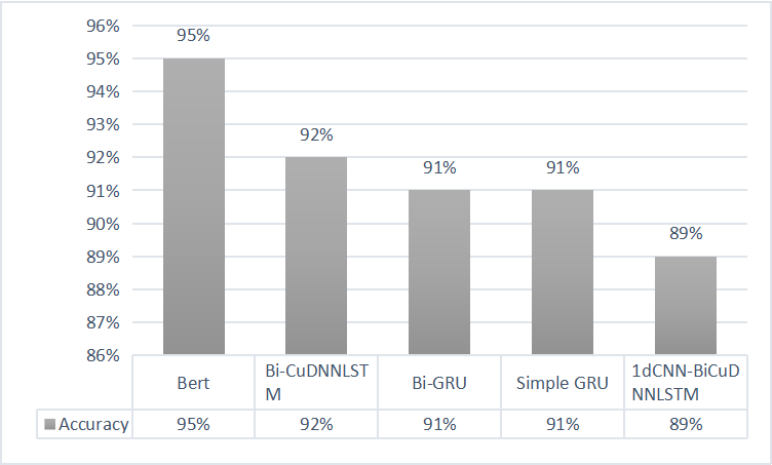


Figure 5.3: Accuracy comparison among the best performing models for multiclass dataset

Chapter 6

Result Analysis

Two datasets were produced for this study's subsequent investigation. The binary classification issue was resolved using the binary class dataset. In this work, a Bi-GRU model was suggested for this investigation.Moreover, the multiclass dataset was used to address multi classification problems.In this research, the pre - trained deep Bert model for multiclassification was suggested. In contrast, a number of deep learning and machine learning models, including Multinomial Naive Bayes, Bernoulli Naive Bayes, Logistic Regression, SVM, Random Forest, KNN, Bi-CuDNNLSTM, Bi-GRU, Simple GRU, CNN, and 1dCNN-BiCuDNNLSTM, For comparison analysis, Bert were utilized. To illustrate the accuracy of every class, a confusion matrix was developed for each algorithm. Furthermore, in this investigation, assessment measures such as Precision, Recall, and F1-score were taken into account. The classification Report, a performance evaluation statistic, displays the test data's accuracy, precision, recall, and f1-score.

Accuracy
The percentage of correct predictions is indicated by the classification issue's accuracy. It is determined by dividing the whole forecasted data by the overall estimated data that was right. A model's performance is measured by accuracy, which contrasts the proportion of accurate forecasts to all other predictions. It is frequently utilized as a summarizing statistic to evaluate a classification model's overall effectiveness.

Precision
Among all positive predictions produced by the model (true positives plus false positives), precision is the percentage of correctly predicted predictions (i.e., the quantity of right positive predictions). It is a gauge of how well the model can spot positive instances and reduce the amount of false positives. A model becomes less likely to mistakenly classify a counter example as positive when there is high accuracy since there are minimal false positives.

Recall
Recall, often referred to as sensitivity, is the percentage of accurate predictions made (i.e., the number of positive predictions that came true) out of all real good instances in the data (true positives plus false negatives). It evaluates the model's capability to accurately identify each positive example and reduce the amount of false negatives. High recall increases the likelihood that the model will correctly identify all positive cases since there are minimal false negatives.

F1-score
Since F1-score is the mean of these two values, it gives a full picture of Precision and Recall. Assuming Precision and Recall were equal, it is at its best. Accuracy can be a useful metric to evaluate a model's performance, but it can be misleading if the class distribution is imbalanced (i.e., one class has many more examples than the other).In these circumstances, a model that consistently predictions the majority class can be highly accurate but would not be a desirable model. Other measures, such accuracy, recall, and F1-score, are more instructive in this situation.

6.1 Binary Classification Models

6.1.1 Deep Learning Approach:

The percentage of accuracy of Bi-GRU is 97%, Bi-CuDNNLSTM is 97%, Simple GRU is 96%, CNN is 90% and 1dCNN-BiCuDNNLSTM is 95% in Binary Classification.

Bi-GRU

The proposed method can successfully predict 11,600 negative labels and 11,594 positive labels out of 12,016 times. The confusion matrix is given below:

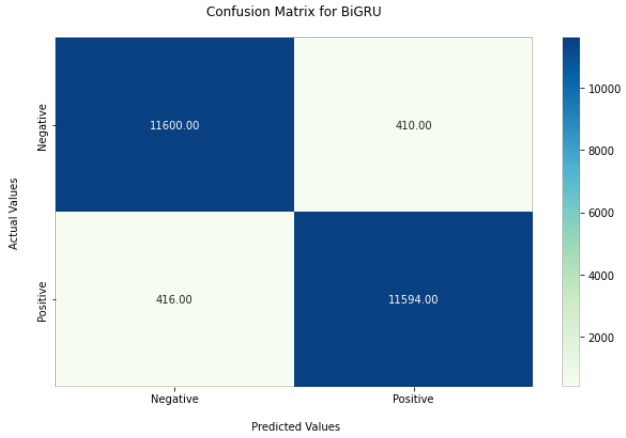


Figure 6.1: Confusion Matrix of Bi-GRU

In the figure below, the precision value, recall value, and f1-score are highest for both sentiments.

Table 6.1: Classification Report of Bi-GRU

	Precision	Recall	F1-score	Support
Negative	0.97	0.97	0.97	12016
Positive	0.97	0.97	0.97	12004
Accuracy			0.97	24020
Macro Avg	0.97	0.97	0.97	24020
Weighted Avg	0.97	0.97	0.97	24020

Bi-CuDNNLSTM

The accuracy of is Bi-CuDNNLSTM is 0.97 or 97Ḃi-CuDNNLSTM can successfully predict 11,586 negative labels and 11,610 positive labels out of 12,016 times. The confusion matrix is given below:

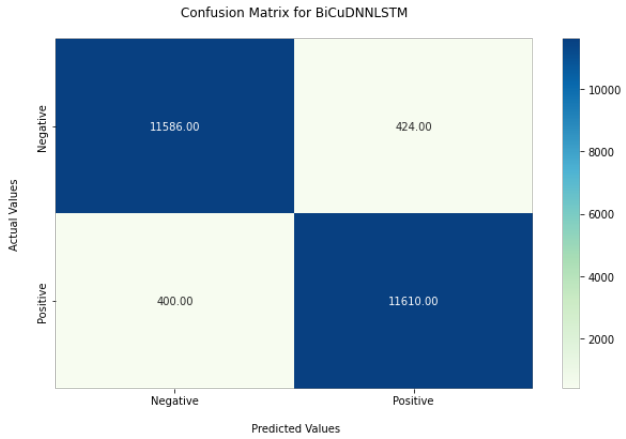


Figure 6.2: Confusion Matrix of BiCuDNNLSTM

In the figure below, the precision value, recall value, and f1-score varies for both sentiments.

Table 6.2: Classification Report of Bi-CuDNNLSTM

	Precision	Recall	F1-score	Support
Negative	0.96	0.97	0.97	11986
Positive	0.97	0.96	0.97	12034
Accuracy			0.97	24020
Macro Avg	0.97	0.97	0.97	24020
Weighted Avg	0.97	0.97	0.97	24020

Simple GRU

A GRU is a variant of RNN that can process sequential data by maintaining an internal memory state. In a GRU for binary classification, the input is passed through the GRU and the output is then passed through a fully connected layer that produces the final classification decision. The accuracy of Simple GRU is 0.96 or 96%. The confusion matrix of the Simple-GRU is given below:

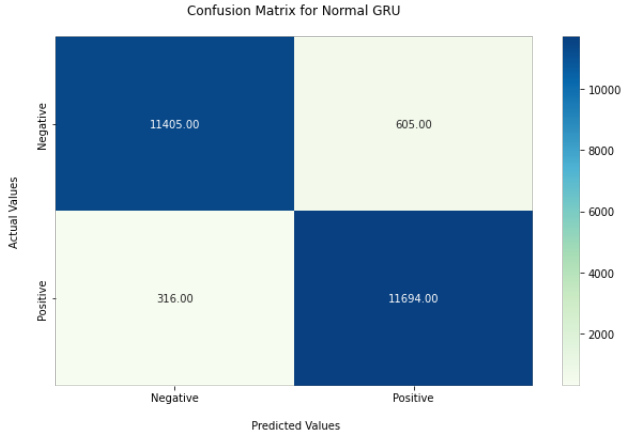


Figure 6.3: Confusion Matrix of Simple GRU

Table 6.3: Classification Report of Simple GRU

	Precision	Recall	F1-score	Support
Negative	0.95	0.97	0.96	11721
Positive	0.97	0.95	0.96	12299
Accuracy			0.96	24020
Macro Avg	0.96	0.96	0.96	24020
Weighted Avg	0.96	0.96	0.96	24020

CNN

In a CNN for binary classification, the input data is passed through a series of convolutional layers that learn to extract features from the input. The accuracy of CNN is 0.90 or 90%. he confusion matrix of the Simple-GRU is given below:

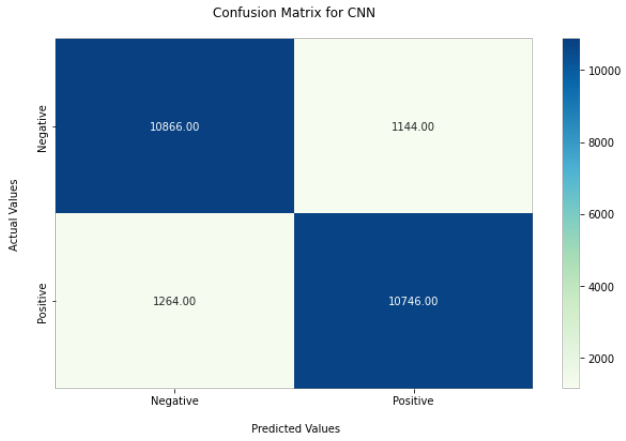


Figure 6.4: Confusion Matrix of CNN

Table 6.4: Classification Report of CNN

	Precision	Recall	F1-score	Support
Negative	0.90	0.91	0.90	12130
Positive	0.89	0.89	0.90	11890
Accuracy			0.90	24020
Macro Avg	0.90	0.90	0.90	24020
Weighted Avg	0.90	0.90	0.90	24020

1dCNN-BiCuDNNLSTM

A 1D CNN-BiCuDNNLSTM (1-dimensional convolutional neural network with bidirectional CuDNN LSTM) is a combination of two different types of neural networks: a 1D CNN and a BiCuDNNLSTM (bidirectional CuDNN LSTM). In this architecture, the input is first passed through a 1D CNN which learns to extract features from the input and then passed through a BiCuDNNLSTM, which is an optimized version of LSTM that uses CUDA library to speed up computation. The accuracy of 1dCNN-BiCuDNNLSTM is 0.95 or 95%. he confusion matrix of the Simple-GRU is given below:

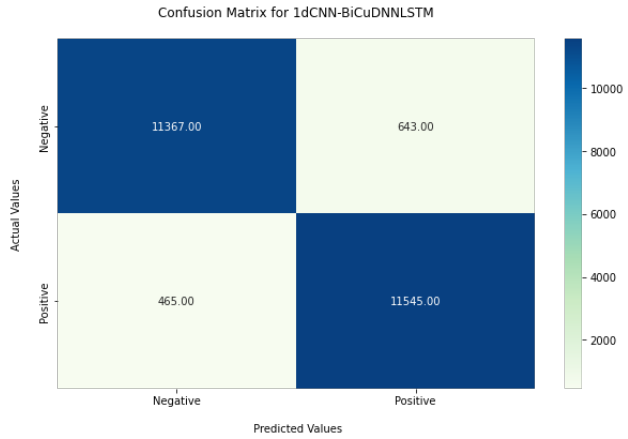


Figure 6.5: Confusion Matrix of 1dCNN-BiCuDNNLSTM

Table 6.5: Classification Report of 1dCNN-BiCuDNNLSTM

	Precision	Recall	F1-score	Support
Negative	0.95	0.96	0.95	11832
Positive	0.96	0.95	0.95	12188
Accuracy			0.95	24020
Macro Avg	0.95	0.95	0.95	24020
Weighted Avg	0.95	0.95	0.95	24020

6.1.2 Machine Learning Approach

The percentage of accuracy of TF-IDF for Multinomial Naive Bayes Model is 92%, for Bernoulli Naive Bayes is 77%, for Logistic Regression is 93%, for Support Vector Machine is 93%, for Random Forest 80%, for K-Nearest Neighbors is 78%. The percentage of accuracy of CountVectorizer (BOW) for Multinomial Naive Bayes Model is 92%, for Bernoulli Naive Bayes is 81%, for Logistic Regression is 93%, for Support Vector Machine is 93%, for Random Forest 76%, for K-Nearest Neighbors is 79% in Binary classification.
For machine learning model feature extraction plays an important role. For this study, CountVectorizer and Term Frequency (TF) Inverse Term Frequency (IDF) is used.

Multinomial Naive Bayes

Multinomial Naive Bayes is frequently used for text classification and natural language processing tasks. In text classification tasks, the input features are typically the words in a document, and the goal is to determine the class label (e.g. positive or negative sentiment) of the document based on the presence and frequency of certain words. The accuracy of multinomial naive bayes is 92%. Confusion matrix and classification report is given below:

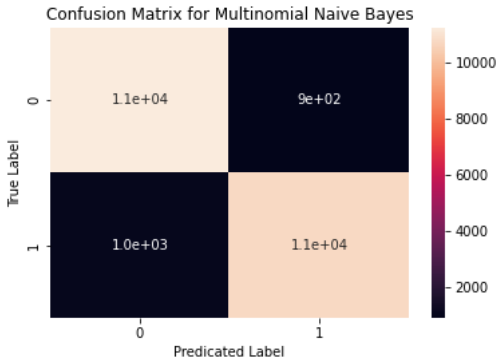


Figure 6.6: Confusion Matrix of Multinomial Naive Bayes

In the figure above, 0 and 1 represent negative and positive sentiments respectively.

Table 6.6: Classification of Multinomial Naive Bayes

	Precision	Recall	F1-score	Support
Negative	0.91	0.93	0.92	11868
Positive	0.93	0.91	0.92	12152
Accuracy			0.92	24020
Macro Avg	0.92	0.92	0.92	24020
Weighted Avg	0.92	0.92	0.92	24020

Bernoulli Naive Bayes

Bernoulli Naive Bayes is a Naive Bayes variation built primarily for binary classification applications. It is presumptively predicated on the premise that all characteristics are binary (i.e., they can take on only two values, such as 0 or 1, true or false, etc.). Confusion matrix and classification report is given below:

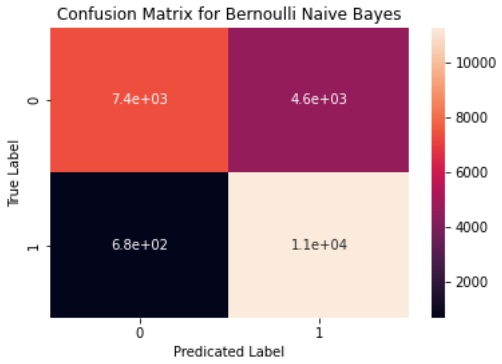


Figure 6.7: Confusion Matrix of Bernoulli Naive Bayes

Table 6.7: Classification of Bernoulli Naive Bayes

	Precision	Recall	F1-score	Support
Negative	0.91	0.61	0.73	12026
Positive	0.70	0.94	0.81	11994
Accuracy			0.77	24020
Macro Avg	0.81	0.77	0.77	24020
Weighted Avg	0.81	0.77	0.77	24020

Logistic Regression

The logistic regression model uses a logistic function to model the probability of the positive class (denoted as P(y=1—x)) given the predictor variables. Once the model has been trained, it can be used to make predictions on new data. The predicted probability of the positive class can be used to classify the new data point as the positive class if the probability is above a certain threshold (usually 0.5) or as the negative class otherwise. The accuracy of logistic regression is 93%. Ihe confusion matrix is given below:

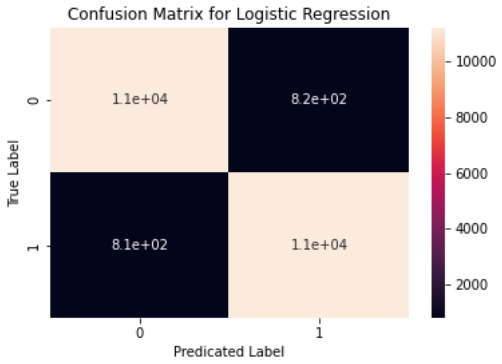


Figure 6.8: Confusion Matrix of Logistic Regression

In the figure above, 0 and 1 represent negative and positive sentiments respectively.

Table 6.8: Classification of Logistic Regression

	Precision	Recall	F1-score	Support
Negative	0.94	0.93	0.93	12087
Positive	0.93	0.94	0.93	11933
Accuracy			0.93	24020
Macro Avg	0.93	0.93	0.93	24020
Weighted Avg	0.93	0.93	0.93	24020

SVM

The goal of a SVM for binary classification is to find the best boundary (or hyperplane) that separates the data points of one class from the other class. The accuracy of SVM model is 93% The confusion matrix is given below:

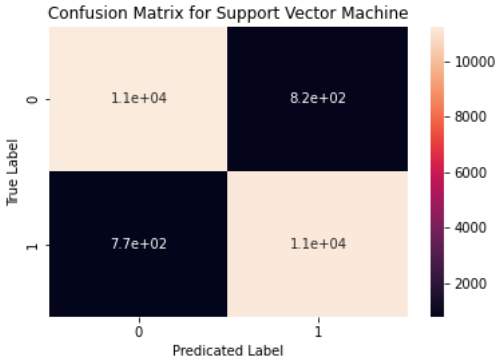


Figure 6.9: Confusion Matrix of Support Vector Machine

In the figure above, 0 and 1 represent negative and positive sentiments respectively.

Table 6.9: Classification of Support Vector Machine

	Precision	Recall	F1-score	Support
Negative	0.94	0.93	0.93	11989
Positive	0.93	0.94	0.93	12031
Accuracy			0.93	24020
Macro Avg	0.93	0.93	0.93	24020
Weighted Avg	0.93	0.93	0.93	24020

Random Forest

In binary classification, the Random Forest algorithm can predict the probability of a given input belonging to each class. Once the probability is obtained, a threshold can be set. For example, if the threshold is 0.5, any probability greater than 0.5 is classified as class A and anything less than 0.5 is classified as class B. The accuracy of random forest is 76%. The confusion matrix is given below:

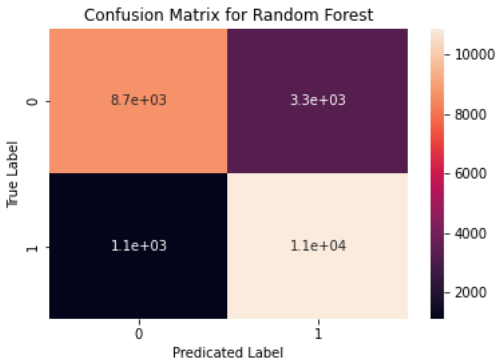


Figure 6.10: Confusion Matrix of Random Forest

In the figure above, 0 and 1 represent negative and positive sentiments respectively.

Table 6.10: Classification of Random Forest

	Precision	Recall	F1-score	Support
Negative	0.89	0.60	0.72	11958
Positive	0.70	0.92	0.80	12062
Accuracy			0.76	24020
Macro Avg	0.79	0.76	0.76	24020
Weighted Avg	0.79	0.76	0.76	24020

KNN

The goal of KNN is to assign a label to a fresh data point depending on the label assignment of its k closest neighbors. The first thing to do is get a sense of how many neighbors there are, denoted by the quantity k. Once k is established, the algorithm looks through the training instances to find the k closest to the target point. When using KNN to a binary classification problem, the test point will be assigned to the group that has the most of its k closest neighbors. A 78% precision is achieved using the KNN model. Here is the confusion matrix:

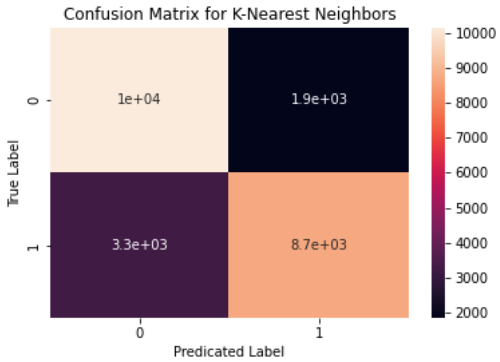


Figure 6.11: Confusion Matrix of K-Nearest Neighbors

In the figure above, 0 and 1 represent negative and positive sentiments respectively.

Table 6.11: Classification of K-Nearest Neighbors

	Precision	Recall	F1-score	Support
0	0.76	0.83	0.79	11956
1	0.82	0.73	0.77	12064
Accuracy			0.78	24020
Macro Avg	0.79	0.78	0.78	24020
Weighted Avg	0.79	0.78	0.78	24020

6.2 Multiclass Classification Model

6.2.1 Deep Learning Approach

The percentage of accuracy of Bi-CuDNNLSTM is 92%, Bi-GRU is 91%, Simple GRU is 91%, CNN is 85%, 1dCNN-BiCuDNNLSTM is 89% and BERT is 95% in Multi-Class Classification.

BERT Model

The proposed method can successfully predict 6238 negative labels, 6278 neutral and 6189 positive labels. The confusion matrix of the Bert is given below:

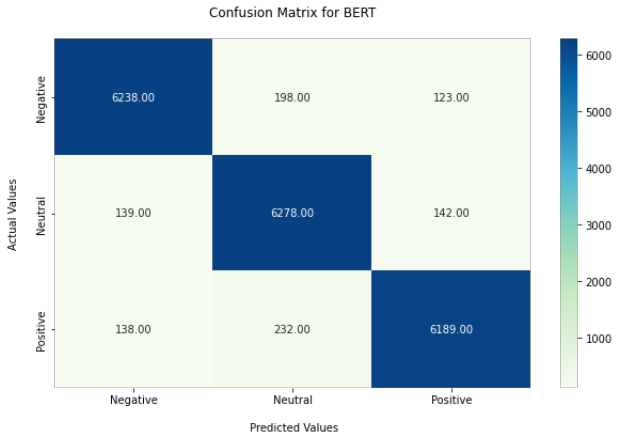


Figure 6.12: Confusion Matrix of BERT

Table 6.12: Classification Report of Bart

	Precision	Recall	F1-score	Support
Negative	0.95	0.96	0.95	6515
Neutral	0.96	0.94	0.95	6708
Positive	0.94	0.96	0.95	6454
Accuracy			0.95	19677
Macro Avg	0.95	0.95	0.95	19677
Weighted Avg	0.95	0.95	0.95	19677

Bi-CuDNNLSTM

The accuracy of is Bi-CuDNNLSTM is 0.92 or 92%. Bi-CuDNNLSTM can successfully predict 6135 negative labels, 5854 neutral and 6155 positive labels. The confusion matrix of the Bi-CuDNNLSTM is given below:

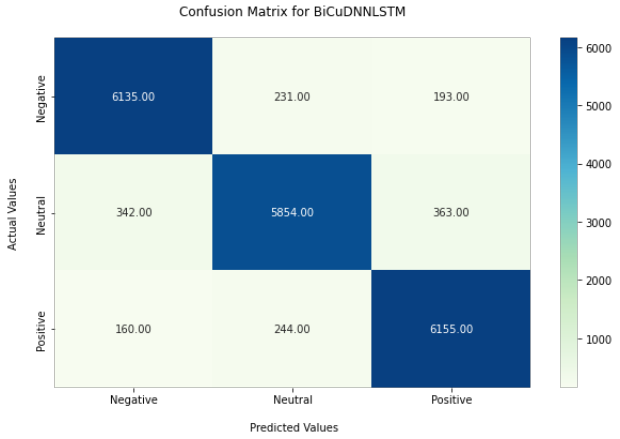


Figure 6.13: Confusion Matrix of Bi-CuDNNLSTM

Table 6.13: Classification Report of Bi-CuDNNLSTM

	Precision	Recall	F1-score	Support
Negative	0.94	0.92	0.93	6637
Neutral	0.89	0.92	0.91	6329
Positive	0.94	0.92	0.93	6711
Accuracy			0.92	19677
Macro Avg	0.92	0.92	0.92	19677
Weighted Avg	0.92	0.92	0.92	19677

Bi-GRU

The accuracy of is Bi-GRU is 0.91 or 91%. Bi-GRU can successfully predict 5885 negative labels, 6083 neutral and 5917 positive labels. The confusion matrix is given below:

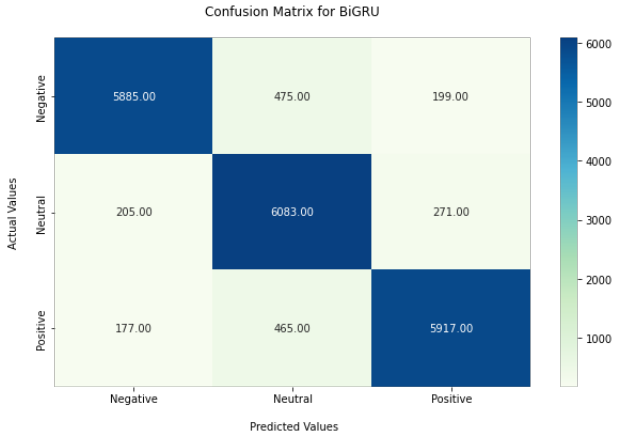


Figure 6.14: Confusion Matrix of Bi-GRU

Table 6.14: Classification Report of Bi-GRU

	Precision	Recall	F1-score	Support
Negative	0.90	0.94	0.92	6267
Neutral	0.93	0.87	0.90	7023
Positive	0.90	0.93	0.91	6387
Accuracy			0.91	19677
Macro Avg	0.91	0.91	0.91	19677
Weighted Avg	0.91	0.91	0.91	19677

Simple GRU

The accuracy of Simple GRU is 0.92 or 92%. Simple GRU can successfully predict 6052 negative labels, 6045 neutral and 5871 positive labels. The confusion matrix of the Simple GRU is given below:

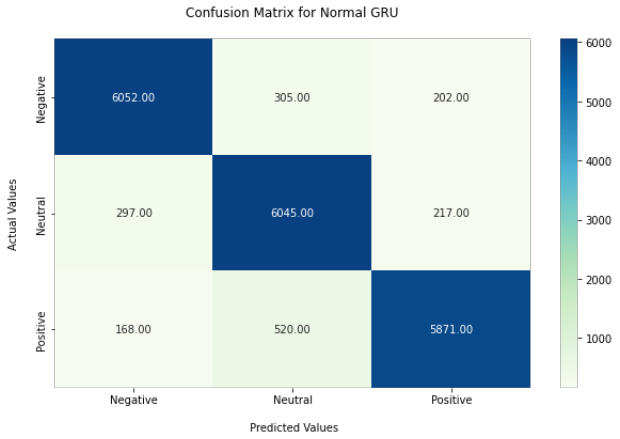


Figure 6.15: Confusion Matrix of Simple GRU

Table 6.15: Classification Report of Simple GRU

	Precision	Recall	F1-score	Support
Negative	0.92	0.93	0.93	4756
Neutral	0.92	0.88	0.90	8130
Positive	0.90	0.93	0.91	6791
Accuracy			0.91	19677
Macro Avg	0.91	0.91	0.91	19677
Weighted Avg	0.91	0.91	0.91	19677

CNN

The accuracy of is Bi-CuDNNLSTM is 0.85 or 85%. CNN can successfully predict 5835 negative labels, 6224 neutral and 5757 positive labels. The confusion matrix of the CNN is given below:

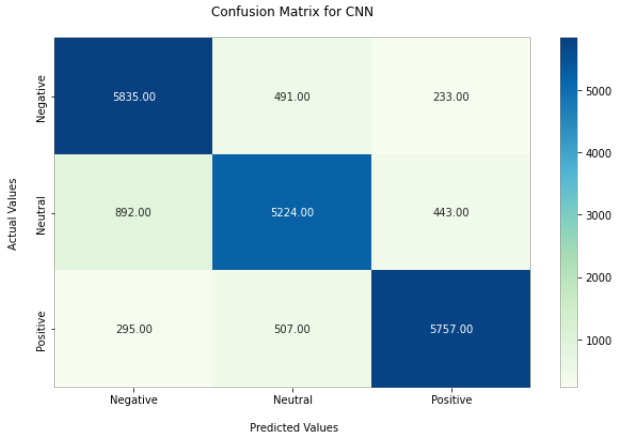


Figure 6.16: Confusion Matrix of CNN

In the figure below, the precision value, recall value, and f1-score varies for both sentiments.

Table 6.16: Classification Report of CNN

	Precision	Recall	F1-score	Support
Negative	0.89	0.83	0.86	7022
Neutral	0.80	0.84	0.82	6222
Positive	0.88	0.89	0.89	6433
Accuracy			0.85	19677
Macro Avg	0.85	0.86	0.85	19677
Weighted Avg	0.86	0.85	0.85	19677

1dCNN-BiCuDNNLSTM

The accuracy of 1dCNN-BiCuDNNLSTM is 0.89 or 89%. 1dCNN-BiCuDNNLSTM can successfully predict 6124 negative labels, 5736 neutral and 5746 positive labels. The confusion matrix of the 1dCNN-BiCuDNNLSTM is given below:

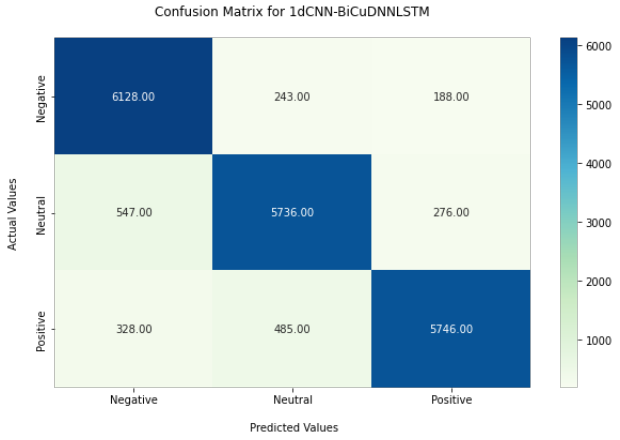


Figure 6.17: Confusion Matrix of 1dCNN-BiCuDNNLSTM

In the figure below, the precision value, recall value, and f1-score varies for both sentiments.

Table 6.17: Classification Report of 1dCNN-BiCuDNNLSTM

	Precision	Recall	F1-score	Support
Negative	0.93	0.88	0.90	7003
Neutral	0.87	0.89	0.88	6464
Positive	0.88	0.93	0.90	6210
Accuracy			0.89	19677
Macro Avg	0.89	0.90	0.89	19677
Weighted Avg	0.90	0.89	0.90	19677

6.2.2 Machine Learning Approach

The percentage of accuracy of TF-IDF for Multinomial Naive Bayes Model is 77%, for Bernoulli Naive Bayes is 63%, for Logistic Regression is 80%, for Support Vector Machine is 81%, for Random Forest 58%, for K-Nearest Neighbors is 54%.

Multinomial Naive Bayes

The accuracy of Multinomial Naive Bayesis 0.77 or 77%. can The confusion matrix of the Multinomial Naive Bayes is given below:

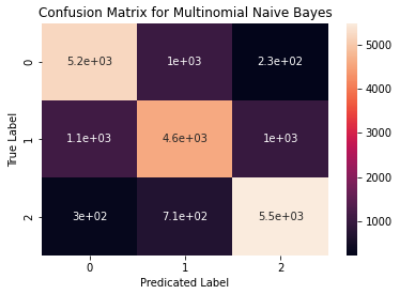


Figure 6.18: Confusion Matrix of Multinational Naive Bayes

In the figure above, 0, 1 and 2 represent negative, neutral and positive sentiments respectively.

Table 6.18: Classification Report of Multinomial Naive Bayes

	Precision	Recall	F1-score	Support
Negative	0.78	0.80	0.79	6579
Neutral	0.72	0.68	0.70	6629
Positive	0.82	0.85	0.83	6469
Accuracy			0.77	19677
Macro Avg	0.77	0.78	0.77	19677
Weighted Avg	0.77	0.77	0.77	19677

Bernoulli Naive Bayes

The accuracy of Bernoulli Naive Bayes Is 0.63 or 63%. can The confusion matrix of the Bernoulli Naive Bayes is given below:

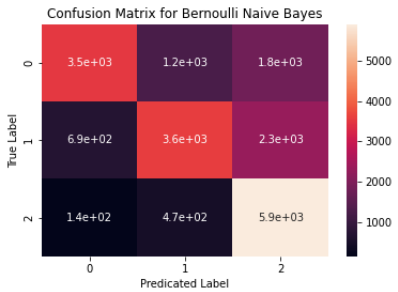


Figure 6.19: Confusion Matrix of Bernoulli Naive Bayes

In the figure above, 0, 1 and 2 represent negative, neutral and positive sentiments respectivel.

Table 6.19: Classification Report of Bernoulli Naive Bayes

	Precision	Recall	F1-score	Support
Negative	0.76	0.48	0.59	6650
Neutral	0.63	0.51	0.57	6529
Positive	0.57	0.89	0.70	6498
Accuracy			0.63	19677
Macro Avg	0.65	0.63	0.62	19677
Weighted Avg	0.65	0.63	0.62	19677

Logistic Regression

The accuracy of Logistic Regression Is 0.81 or 81%. The confusion matrix is given below:

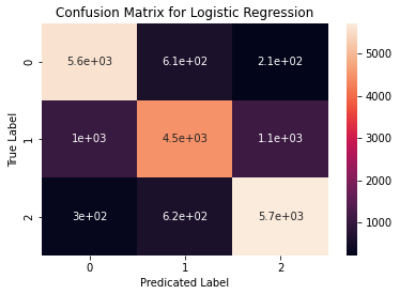


Figure 6.20: Confusion Matrix of Logistic Regression

In the figure above, 0, 1 and 2 represent negative, neutral and positive sentiments respectively.

Table 6.20: Classification Report of Logistic Regression

	Precision	Recall	F1-score	Support
Negative	0.84	0.85	0.84	6580
Neutral	0.78	0.71	0.74	6562
Positive	0.80	0.86	0.83	6535
Accuracy			0.81	19677
Macro Avg	0.81	0.81	0.81	19677
Weighted Avg	0.81	0.81	0.81	19677

SVM

The accuracy is SVM 0.81 or 81%. The confusion matrix is given below:

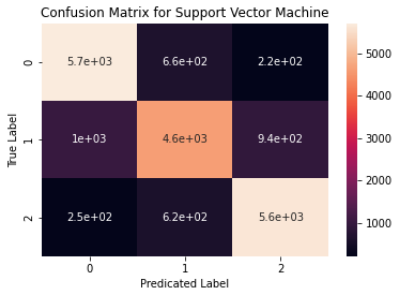


Figure 6.21: Confusion Matrix of Support Vector Machine

In the figure above, 0, 1 and 2 represent negative, neutral and positive sentiments respectively.

Table 6.21: Classification Report of Support Vector Machine

	Precision	Recall	F1-score	Support
Negative	0.81	0.86	0.84	6595
Neutral	0.78	0.70	0.73	6569
Positive	0.83	0.86	0.84	6513
Accuracy			0.81	19677
Macro Avg	0.81	0.81	0.80	19677
Weighted Avg	0.81	0.81	0.80	19677

Random Forest

The accuracy is Random Forest 0.59 or 59%. can The confusion matrix of the Random Forest is given below:

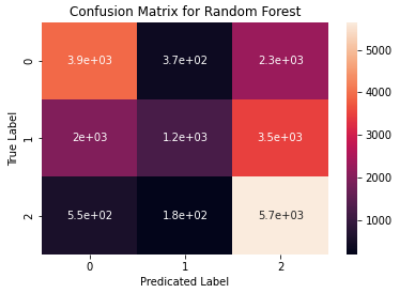


Figure 6.22: Confusion Matrix of Random Forest

In the figure above, 0, 1 and 2 represent negative, neutral and positive sentiments respectively.

Table 6.22: Classification Report of Random Forest

	Precision	Recall	F1-score	Support
Negative	0.68	0.57	0.62	6627
Neutral	0.55	0.42	0.48	6442
Positive	0.55	0.78	0.64	6608
Accuracy			0.59	19677
Macro Avg	0.60	0.59	0.58	19677
Weighted Avg	0.60	0.59	0.58	19677

KNN

The accuracy is KNN 0.54 or 54%. can The confusion matrix of the KNN is given below:

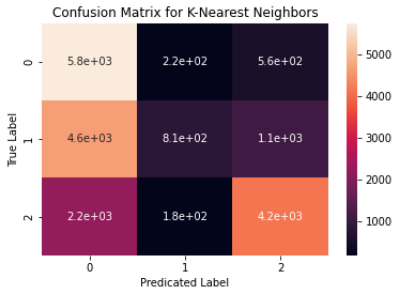


Figure 6.23: Confusion Matrix of K-Nearest Neighbors

In the figure above, 0, 1 and 2 represent negative, neutral and positive sentiments respectively.

Table 6.23: Classification Report of K-Nearest Neighbors

	Precision	Recall	F1-score	Support
-1	0.46	0.86	0.60	6598
0	0.63	0.13	0.22	6521
1	0.70	0.63	0.66	6558
Accuracy			0.54	19677
Macro Avg	0.59	0.54	0.49	19677
Weighted Avg	0.59	0.54	0.49	19677

6.3 Combined Analysis

The following table displays a summary categorization report for all active models. Precision, recall, f1-score, support values, accuracy, macro average, and weighted average are listed below for each category. The suggested model, Bi-GRU, outperforms the competition when it comes to binary categorization.

Table 6.24: Classification Report of proposed model, other machine learning and deep learning model for Binary Classification

Models	**Class**	**Negative**	**Positive**	**Accuracy**	**Macro Avg**	**Weighted Avg**
Bi-GRU **(Proposed Model)**	Precision	0.97	0.97	-	0.97	0.97
	Recall	0.97	0.97	-	0.97	0.97
	F1 Score	0.97	0.97	0.97	0.97	0.97
	Support	12016	12004	24020	24020	24020
Bi-CuDNNLSTM	Precision	0.96	0.97	-	0.97	0.97
	Recall	0.97	0.96	-	0.97	0.97
	F1 Score	0.97	0.97	0.97	0.97	0.97
	Support	11986	12034	24020	24020	24020
Normal GRU	Precision	0.95	0.97	-	0.96	0.96
	Recall	0.97	0.95	-	0.96	0.96
	F1 Score	0.96	0.96	0.96	0.96	0.96
	Support	11721	12299	24020	24020	24020
CNN	Precision	0.90	0.89	-	0.90	0.90
	Recall	0.90	0.90	-	0.90	0.90
	F1 Score	0.90	0.90	0.90	0.90	0.90
	Support	12130	11890	24020	24020	24020
1dCNN-BiCuDNNLSTM	Precision	0.95	0.96	-	0.95	0.95
	Recall	0.96	0.95	-	0.95	0.95
	F1 Score	0.95	0.95	0.95	0.95	0.95
	Support	11832	12188	24020	24020	24020
Multinomial Naive Bayes	Precision	0.91	0.93	-	0.92	0.92
	Recall	0.93	0.91	-	0.92	0.92
	F1 Score	0.92	0.92	0.92	0.92	0.92
	Support	11868	12152	24020	24020	24020
Bernoulli Naive Bayes	Precision	0.94	0.74	-	0.84	0.84
	Recall	0.66	0.96	-	0.81	0.81
	F1 Score	0.78	0.83	0.81	0.81	0.81
	Support	11999	12021	24020	24020	24020

Models	**Class**	**Negative**	**Positive**	**Accuracy**	**Macro Avg**	**Weighted Avg**
Logistic Regression	Precision	0.93	0.93	-	0.93	0.93
	Recall	0.93	0.93	-	0.93	0.93
	F1 Score	0.93	0.93	0.93	0.93	0.93
	Support	11993	12027	24020	24020	24020
SVM	Precision	0.94	0.93	-	0.93	0.93
	Recall	0.93	0.94	-	0.93	0.93
	F1 Score	0.94	0.93	0.93	0.93	0.93
	Support	11969	12031	24020	24020	24020
Random Forest	Precision	0.91	0.74	-	0.82	0.82
	Recall	0.68	0.93	-	0.80	0.80
	F1 Score	0.78	0.82	0.80	0.80	0.80
	Support	12113	11907	24020	24020	24020
KNN	Precision	0.76	0.82	-	0.79	0.79
	Recall	0.83	0.73	-	0.78	0.78
	F1 Score	0.79	0.77	0.78	0.78	0.78
	Support	11956	12064	24020	24020	24020

For multiclass classification, the proposed Bert model generates the highest scores compared to the rest.

Table 6.25: Classification Report of proposed model, other machine learning and deep learning model for Multiclass Classification

Models	Class	Negative	Neutral	Positive	Accuracy	Macro Avg	Weighted Avg
BERT **(Proposed model)**	Precision	0.95	0.96	0.94	-	0.95	0.95
	Recall	0.96	0.94	0.96	-	0.95	0.95
	F1 Score	0.95	0.95	0.95	0.95	0.95	0.95
	Support	6515	6798	6454	19677	19677	19677
Bi-GRU	Precision	0.90	0.93	0.90	-	0.91	0.91
	Recall	0.94	0.87	0.93	-	0.91	0.91
	F1 Score	0.92	0.90	0.91	0.91	0.91	0.91
	Support	6267	7023	6387	19677	19677	19677
Bi-CuDNN LSTM	Precision	0.94	0.89	0.94	-	0.92	0.92
	Recall	0.92	0.92	0.92	-	0.92	0.92
	F1 Score	0.93	0.91	0.93	0.92	0.92	0.92
	Support	6637	6329	6711	19677	19677	19677
Normal GRU	Precision	0.92	0.92	0.90	-	0.91	0.91
	Recall	0.93	0.88	0.93	-	0.91	0.91
	F1 Score	0.93	0.90	0.91	0.91	0.91	0.91
	Support	4756	8130	6791	19677	19677	19677
CNN	Precision	0.89	0.80	0.88	-	0.85	0.86
	Recall	0.83	0.84	0.89	-	0.86	0.85
	F1 Score	0.86	0.82	0.89	0.85	0.85	0.85
	Support	7022	6222	6433	19677	19677	19677
1dCNN-BiCu DNNLSTM	Precision	0.93	0.87	0.88	-	0.89	0.90
	Recall	0.88	0.89	0.93	-	0.90	0.89
	F1 Score	0.90	0.88	0.90	0.89	0.89	0.90
	Support	7003	6464	6210	19677	19677	19677
Multinomial Naive Bayes	Precision	0.78	0.72	0.82	-	0.77	0.77
	Recall	0.80	0.68	0.85	-	0.78	0.77
	F1 Score	0.79	0.70	0.83	0.77	0.77	0.77
	Support	6579	6629.	6469	19677	19677	19677
Bernoulli Naive Bayes	Precision	0.76	0.63	0.57	-	0.65	0.65
	Recall	0.48	0.51	0.89	-	0.63	0.63
	F1 Score	0.59	0.57	0.70	0.63	0.62	0.62
	Support	6650	6529	6498	19677	19677	19677
Logistic Regression	Precision	0.84	0.78	0.80	-	0.81	0.81
	Recall	0.85	0.71	0.86	-	0.81	0.81
	F1 Score	0.84	0.74	0.83	0.81	0.81	0.81
	Support	6580	6562	6535	19677	19677	19677
SVM	Precision	0.81	0.78	0.83	-	0.81	0.81
	Recall	0.86	0.70	0.86	-	0.81	0.81
	F1 Score	0.84	0.73	0.84	0.81	0.80	0.80
	Support	6595	6569	6513	19677	19677	19677
Random Forest	Precision	0.68	0.55	0.55	-	0.60	0.60
	Recall	0.57	0.42	0.78	-	0.59	0.59
	F1 Score	0.62	0.48	0.64	0.59	0.58	0.58
	Support	6627	6442	6608	19677	19677	19677
KNN	Precision	0.46	0.63	0.70	-	0.59	0.59
	Recall	0.86	0.19	0.63	-	0.54	0.54
	F1 Score	0.60	0.22	0.66	0.54	0.49	0.49
	Support	6598	6521	6558	19677	19677	19677

6.4 Discussion

Binary classification and multiclass classification were the two forms of data categorization employed by the researchers in this work. The most accurate models for binary classification are Bi-GRU (97%), Bi-CuDNNLSTM (97%), Normal GRU (96%), CNN (90%), and 1dCNN-BiCuDNNLSTM (95%), while the most accurate models for machine learning are Multinomial Naive Bayes (92%), Bernoulli Naive Bayes (92%), Logistic Regression (93%), SVM (93%), Random Forest (80%), and KNN (78%), with SVM having the highest. The optimum model for Binary Classification in this case, according to the researchers, is a Bi-GRU model. Bi-GRU is superior to other models due to its bidirectional nature, which enables it to take into account both past and prospective contexts while generating a prediction. When performing NLP jobs where context is crucial, this can be quite helpful. In the case of Bi-GRU, the model's bidirectional nature enables it to consider both the past and the future environment, which may be valuable in applications like named entity identification, language processing, and text categorization. but also It may be more challenging to train and comprehend a bidirectional model since it is often more complicated than a unidirectional model. The bidirectional nature of the model enables it to take into consideration both past and prospective context, which can be useful in applications like classification tasks, language translation, and named entity recognition in the case of Bi-CuDNNLSTM, which is parallel to Bi-GRU (which shows the very same accuracy as Bi-GRU). Contrarily, CuDNN is only supported by NVIDIA GPUs, which restricts the model's ability to function in a variety of hardware setups. GRU can manage brief dependencies in sequences in the instance of Normal GRU, which is crucial for NLP applications like language modeling. Normal GRU could work well for some NLP jobs, but it might not be the greatest solution for other kinds of issues. Because CNNs can recognize spatial hierarchies of features from pictures, they are especially well-suited for image identification and classification applications. CNN's, on the other hand, are only able to analyze data that can be organized into grids, like photographs, which restricts their usefulness to other kinds of data. In the case of 1dCNN-BiCuDNNLSTM, this model can be well-suited for a variety of NLP tasks, including text classification, language translation, and named entity recognition. By incorporating the extracting feature's functionality of the 1D CNN and the sequential process technology of the Bi-CuDNNLSTM. But to train efficiently, the 1D-CNN-BiCuDNNLSTM model needs additional information and processing power. These are the justifications for selecting Bi-GRU as the model that fits Binary classification the best out of all the benefits and drawbacks of other models. The models used for binary classification are all the same for multiclass classification, although the BERT model is also utilized, and the accuracy for all deep learning models is as follows: Bert: 95% Normal GRU: 91%, CNN: 85%, 1dCNN-BiCuDNNLSTM: 89%, Bi-GRU: 91%, Bi-CuDNNLSTM: 92% and machine learning models are: SVM: 81%, Logistic Regression: 81%, Multinomial Naive Bayes: 77%, Bernoulli Naive Bayes: 63% KNN scored 54% and Random Forest 59%. Because Bert has the highest accuracy and BERT has been demonstrated to attain state-of-the-art performance on a wide variety of NLP tasks thanks to its capacity to comprehend text context, the authors

chose Bert as the proposed model in this instance. Furthermore, BERT can manage long-term dependencies in sequences, which is critical in NLP tasks like language modeling and can be fine-tuned on specific tasks with little training data, saving a lot of time and computing resources. BERT is the best option for this research although it might be the greatest option for other sorts of issues, even though it may be well suited for other NLP jobs. Additionally, the reason why the machine learning models in this study did not perform well is because they are not fine-tuned.

Chapter 7

Conclusion and Future Work

7.1 Conclusion

Nowadays, it is crucial for job searchers to locate suitable positions, and businesses must monitor the advancement of their personnel. A website like Glassdoor has become highly helpful for finding, moving, and assessing professional objectives as a result of the current increase in job losses. The gathered dataset will include information on wages, working conditions, benefits, bonuses, time off, pensions, and other factors that may indicate how satisfied employees are. Models from Deep Learning and Machine Learning were used in the investigation. By analyzing the requirements and desires of the employees and the firms, this research intends to assist the employees in finding acceptable employment opportunities and to help the businesses enhance their employee-friendly amenities.

7.2 Future Work

We can use fine tuning techniques to reach flawless accuracy and precision if we are able to gather additional datasets in the future. Additionally, we wish to broaden the scope of this research by incorporating hybrid models such as ensemble models, multi-modal models, transfer learning models, hybrid deep learning models, semantic-based models, etc. These models may be used to include many forms of data, such as text, photos, or audio, and can take use of various learning methodologies, including deep learning or conventional machine learning. One more approach to enhance our corpus. The task's objective is to train utilizing Word2vec, doc2vec, or paragraph2vec vectorization models rather than TF-IDF, with the choice of a hybrid model depending on the particular requirements and peculiarities of the dataset. In contrast to TF-IDF, these models are taken into account. We can also fine tune the models accuracy for future works.

Bibliography

[1] A. M. Kibriya, E. Frank, B. Pfahringer, and G. Holmes, "Multinomial naive bayes for text categorization revisited," in *Australasian Joint Conference on Artificial Intelligence*, Springer, 2004, pp. 488–499.

[2] J. Su, J. S. Shirab, and S. Matwin, "Large scale text classification using semisupervised multinomial naive bayes," in *ICML*, 2011.

[3] K. Irie, Z. Tüske, T. Alkhouli, R. Schlüter, H. Ney, *et al.*, "Lstm, gru, highway and a bit of attention: An empirical overview for language modeling in speech recognition," in *Interspeech*, 2016, pp. 3519–3523.

[4] W. Liu, Y. Wen, Z. Yu, and M. Yang, "Large-margin softmax loss for convolutional neural networks," *arXiv preprint arXiv:1612.02295*, 2016.

[5] N. Luo, Y. Zhou, and J. Shon, "Employee satisfaction and corporate performance: Mining employee reviews on glassdoor. com," 2016.

[6] V. Leah-Martin, "Relative compensation and employee satisfaction," *Available at SSRN 2896268*, 2017.

[7] A. F. Agarap, "Deep learning using rectified linear units (relu)," *arXiv preprint arXiv:1803.08375*, 2018.

[8] Q. Wang, C. Xu, Y. Zhou, T. Ruan, D. Gao, and P. He, "An attention-based bi-gru-capsnet model for hypernymy detection between compound entities," in *2018 IEEE International Conference on Bioinformatics and Biomedicine (BIBM)*, IEEE, 2018, pp. 1031–1035.

[9] L. Zhang, Y. Zhou, X. Duan, and R. Chen, "A hierarchical multi-input and output bi-gru model for sentiment analysis on customer reviews," in *IOP conference series: materials science and engineering*, IOP Publishing, vol. 322, 2018, p. 062 007.

[10] H. A. Abu Alfeilat, A. B. Hassanat, O. Lasassmeh, *et al.*, "Effects of distance measure choice on k-nearest neighbor classifier performance: A review," *Big data*, vol. 7, no. 4, pp. 221–248, 2019.

[11] J. L. D. Alves, "Redes neurais recorrentes aplicadas à classificação de fake news em lıngua portuguesa," 2019.

[12] R. Bajpai, D. Hazarika, K. Singh, S. Gorantla, E. Cambria, and R. Zimmerman, "Aspect-sentiment embeddings for company profiling and employee opinion mining," *arXiv preprint arXiv:1902.08342*, 2019.

[13] Y. Jung and Y. Suh, "Mining the voice of employees: A text mining approach to identifying and analyzing job satisfaction factors from online employee reviews," *Decision Support Systems*, vol. 123, p. 113 074, 2019.

[14] G. Singh, B. Kumar, L. Gaur, and A. Tyagi, "Comparison between multinomial and bernoulli naıve bayes for text classification," in *2019 International Conference on Automation, Computational and Technology Management (ICACTM)*, IEEE, 2019, pp. 593–596.

[15] W. de Vries, A. van Cranenburgh, A. Bisazza, T. Caselli, G. van Noord, and M. Nissim, "Bertje: A dutch bert model," *arXiv preprint arXiv:1912.09582*, 2019.

[16] V. Das Swain, K. Saha, M. D. Reddy, H. Rajvanshy, G. D. Abowd, and M. De Choudhury, "Modeling organizational culture with workplace experiences shared on glassdoor," in *Proceedings of the 2020 CHI conference on human factors in computing systems*, 2020, pp. 1–15.

[17] N. Z. Dina, N. Juniarta, *et al.*, "Aspect based sentiment analysis of employee's review experience," *Journal of Information Systems Engineering and Business Intelligence*, vol. 6, no. 1, pp. 79–88, 2020.

[18] S. F. Feng, "Job satisfaction, management sentiment, and financial performance: Text analysis with job reviews from indeed. com," *Authorea Preprints*, 2020.

[19] Y. He, R. Chen, X. Li, *et al.*, "Online at-risk student identification using rnn-gru joint neural networks," *Information*, vol. 11, no. 10, p. 474, 2020.

[20] N. Kashive, V. T. Khanna, and M. N. Bharthi, "Employer branding through crowdsourcing: Understanding the sentiments of employees," *Journal of Indian Business Research*, 2020.

[21] S. N. Lappan, A. W. Brown, and P. S. Hendricks, "Dropout rates of in-person psychosocial substance use disorder treatments: A systematic review and meta-analysis," *Addiction*, vol. 115, no. 2, pp. 201–217, 2020.

[22] A. R. Lubis, M. Lubis, *et al.*, "Optimization of distance formula in k-nearest neighbor method," *Bulletin of Electrical Engineering and Informatics*, vol. 9, no. 1, pp. 326–338, 2020.

[23] S. Mishra, P. K. Mallick, H. K. Tripathy, A. K. Bhoi, and A. González-Briones, "Performance evaluation of a proposed machine learning model for chronic disease datasets using an integrated attribute evaluator and an improved decision tree classifier," *Applied Sciences*, vol. 10, no. 22, p. 8137, 2020.

[24] M. Singh, M. W. Bhatt, H. S. Bedi, and U. Mishra, "Performance of bernoulli's naive bayes classifier in the detection of fake news," *Materials Today: Proceedings*, 2020.

[25] S. Tangirala, "Evaluating the impact of gini index and information gain on classification using decision tree classifier algorithm," *International Journal of Advanced Computer Science and Applications*, vol. 11, no. 2, pp. 612–619, 2020.

[26] I. Wijngaards, M. Burger, and J. van Exel, "Unpacking the quantifying and qualifying potential of semi-open job satisfaction questions through computer-aided sentiment analysis," *Journal of Well-Being Assessment*, vol. 4, no. 3, pp. 391–417, 2020.

[27] S. H. Yoo, H. Geng, T. L. Chiu, *et al.*, "Deep learning-based decision-tree classifier for covid-19 diagnosis from chest x-ray imaging," *Frontiers in medicine*, vol. 7, p. 427, 2020.

[28] L. Alzubaidi, J. Zhang, A. J. Humaidi, *et al.*, "Review of deep learning: Concepts, cnn architectures, challenges, applications, future directions," *Journal of big Data*, vol. 8, no. 1, pp. 1–74, 2021.

[29] G. Chinazzo, "Investigating the indoor environmental quality of different workplaces through web-scraping and text-mining of glassdoor reviews," *Building Research & Information*, vol. 49, no. 6, pp. 695–713, 2021.

[30] S. Dube and C. Zhu, "The disciplinary effect of social media: Evidence from firms' responses to glassdoor reviews," *Journal of Accounting Research*, vol. 59, no. 5, pp. 1783–1825, 2021.

[31] T. Kattenborn, J. Leitloff, F. Schiefer, and S. Hinz, "Review on convolutional neural networks (cnn) in vegetation remote sensing," *ISPRS Journal of Photogrammetry and Remote Sensing*, vol. 173, pp. 24–49, 2021.

[32] F. J. M. Shamrat, S. Chakraborty, M. M. Billah, P. Das, J. N. Muna, and R. Ranjan, "A comprehensive study on pre-pruning and post-pruning methods of decision tree classification algorithm," in *2021 5th International Conference on Trends in Electronics and Informatics (ICOEI)*, IEEE, 2021, pp. 1339–1345.

[33] F. Shamrat, S. Chakraborty, M. Imran, *et al.*, "Sentiment analysis on twitter tweets about covid-19 vaccines using nlp and supervised knn classification algorithm," *Indonesian Journal of Electrical Engineering and Computer Science*, vol. 23, no. 1, pp. 463–470, 2021.

[34] E. Uchida and Y. Kino, "Study on the relationship between employee satisfaction and corporate performance in japan via text mining," *Procedia Computer Science*, vol. 192, pp. 1730–1739, 2021.