City, University of London

MSc Data Science

Project Report

2022

Sentiment Analysis and Classification of Multicategory News Using Natural Language Processing

Researcher: Vighnesh Mane

Supervised by: Dr Oleksandr Galkin

Date: 21/12/2022

Declaration

By submitting this work, I declare that this work is entirely my own except those parts duly identified and referenced in my submission. It complies with any specified word limits and the requirements and regulations detailed in the assessment instructions and any other relevant programme and module documentation. In submitting this work, I acknowledge that I have read and understood the regulations and code regarding academic misconduct, including that relating to plagiarism, as specified in the Programme Handbook. I also acknowledge that this work will be subject to a variety of checks for academic misconduct.

Signed: *Vighnesh Mane*

Abstract

The internet has made it possible for the vast majority of people to access a wealth of information in various formats, including text, photographs, and videos. While it's straightforward to categorise visual content like photos and videos, the technology behind doing the same for text is still in its infancy. Though it's simple to analyse the emotions conveyed in visual media like photos and videos, text sentiment analysis is still in its babyhood. Sentiment analysis and text classification are two problems that may be amenable to various machine learning and deep learning approaches.

In this research project, we investigate the feasibility of using data analysis, time series analysis, time series forecasting, machine learning techniques, and deep learning approaches to address the issues of sentiment analysis and text classification. Real-world data in the form of text is gathered from a variety of news websites for this project. Methods for analysing the data, including time series analysis, machine learning and deep learning, are described in detail in this report. In this document, the key performance indicators of these methods are examined.

Keywords: Data analysis, time series analysis, time series forecasting, machine learning, deep learning.

Table of contents

Chapter 1. Introduction and Objectives

In the first part of this section, the difficulties associated with the sentiment analysis are discussed. In addition to this, the ways in which these issues are approached using previously developed solutions are discussed. In addition, the issue of classifying news is broken down in great detail, and the many approaches that have been proposed by different organisations are explored. Second, the purpose of the project is broken down in this part of the document, and this section also describes the types of entities that will benefit from making use of the remedies that are presented. Thirdly, the objectives of this part are to work toward the goals that will be helpful in responding to the research questions. In addition to that, this portion is made up of the procedures that are utilised to supply the performance of the solutions in order to solve the issues that are associated with sentiment analysis and news classification. In conclusion, this section's concentration on the work plan is meant to guarantee that the project's progress will be well-organized and fruitful.

1.1 Problem background

One of the main focuses of natural language processing is “text classification” because its undeniable value in linguistic studies is more challenging because of a paucity of labelled “training data” (Abebe Azime et al., 2021, p. 1). “Sentiment Analysis” is a subfield of text categorisation that makes significant advances in the field of study on a regular basis. Sentiment categorisation, or determining whether a viewpoint is “positive or negative”, is the primary goal of “opinion mining” (Agarwal et al., 2016, p. 1). For this project, the categorisation of the neutral opinion is also involved. The initial step in performing “Sentiment analysis” is to detect the “polarity” of opinions present in digital networking and to categorise them as either “positive or negative”, and the majority of the research that has been conducted in the discipline of sentiment classification thus far has concentrated on the development of a “classifier” (Hama Aziz et al., 2021, p. 562). For this project, the problem of the polarity of the neutral category was also identified. These days, the web is home to a staggering quantity of data, and it's crucial that we find ways to mine it for insights; numerous challenges can be tackled with the help of various methods of “data mining” and the extraction of information and evaluation of said content for review purposes have become necessities in the age of information and websites and analysing the opinions people have on a topic might reveal if they are “positive or negative” also if they are neutral and the “polarity” of a sentence can be influenced by a to somewhat innovation called denial, which must be taken into account in “sentiment analysis”, and in order to properly function, text processing tools like analysis of the sentiments require the ability to automatically identify denial from News articles (Shirsat et al., 2017, p. 1).

Many materials are in their raw, unorganised form, so understanding “unstructured” text requires a significant amount of effort from a “human” worker because it is so difficult to find useful information within it. As a result, natural language processing specialists have given a lot of focus to computerised text categorisation recently (Barua et al., 2021, p. 113). The sheer volume of data generated by both traditional and online sources necessitates sophisticated data analysis techniques; however, storing all of this “information” in one place without first classifying it into meaningful groups would yield no useful insights from which to draw any meaningful conclusions (Singh et al., 2021, p. 186). Internet news platforms are proliferating at a rapid rate in today's age of ubiquitous internet access, and by displaying ads for various consumer goods, these “online news portals” generate substantial revenue due to their “lack of intelligence” however, conventional news websites are unable to determine which sorts of news are most popular with their readers, and because of such news websites frequently subject their viewers to ads that have little to do with the content they're reading, causing substantial financial harm to the companies who pay for such ads and news websites can increase their revenue by targeting their advertising to the content their users are interested “contextual” advertising (Ahmed et al., 2021, p. 609). The most popular internet sites are currently the primary way that people find out about current events, and on the other hand, as the online source of media websites grows, so make the difficulties that come along with it and the demands of contemporary society cannot be satisfied by the conventional

“text classification” techniques so, text categorisation algorithm development has been a major focus in “text mining” during the past few years Li (2013, quoted in Liu, 2016, p. 1). The textual news categorisation is rapid in processing all text input and reliable in its classification of label categories which means that automatic categorisation can greatly enhance the accuracy with which the information site's text categorisation task is carried out, as well as assist the organisation in conserving money (Li et al., 2016, p. 1).

1.2 Purpose and beneficiaries of the project

This study aims to apply various data analysis, machine learning, and deep learning approaches to text data to address challenges, including sentiment analysis and news classification. The answers to the following research questions will give you graphical representations of real-world text data and classification that may be compared to established norms.

1. What is the proportion and percentage of negative, neutral, and positive attitudes, and how does the globe feel about news from categories such as business, food, the environment, entertainment, sports, politics, and health?

2. How much time is spent on reading different types of online news on an hourly, weekly, and monthly basis in different parts of the world? Which nation's news is read the most widely around the world? What is the expected rate of the news data going forward?

3. Will the newly created machine learning techniques for sentiment analysis for this project where huge data is used be able to attain a fair degree of performance when measured against the machine learning approaches that have been employed in the history?

4. Will the deep learning algorithms that have been designed specifically for this study with the enormous dataset be able to exceed the deep learning techniques that other researchers have utilised in the past?

The beneficiaries will be benefitted from the results of this project are as follows:

1. Multiple groups of individuals who take in the news, most notably the group of individuals who read the news from a variety of domains on various internet platforms.

2. As a result of the fact that businesses produce enormous amounts of text data, only a few of those businesses are able to organise the text data according to the categories. As a result, such businesses could benefit from this study in terms of classification strategies.

3. The analytics report of the text data is not kept by many organisations, and these same institutions also do not gather feedback from customers. Therefore, the findings of this paper may be utilised by establishments of this kind.

4. Numerous online news websites stand to gain advantages if they are able to publish stories solely in response to requests from their readership. If this constitution places emphasis on the various techniques for classifying things, then the goal can be accomplished.

5. Academics and other individuals who utilise the feedback system stand to benefit from this study if the company is able to make use of the study that's been presented in this project and is able to create a text feedback system.

1.3 Objectives of the project

The project should be approached in a systematic manner so there is a high probability of getting the results as per the expectations, especially for the machine learning and deep learning algorithms which

will be used further. Therefore, the objectives of the project are to provide the solution to the research questions in a structured way. Objectives of this study are as follows:

1. Literature Survey - The literature review is carried out in order to collect information that relates to the topic. The knowledge that has been gathered in relation to a particular area comes from the many studies that have been proposed in the past by other scholars. The literature review also provides a synopsis of the project's background information.

2. Data collection – The information that will be acquired for this section originates from a wide range of different sources and locations. The site scraping will also be carried out in addition.

3. Data cleaning - The reason for the data cleaning is that data gathered from various sources is not cleaned data, and it needs to be cleaned before using the techniques of data visualisation, machine learning, and deep learning in order to receive accurate results.

4. Data visualisation – The objective of this section is to generate some insights linked to the topic of the project so that those insights may be used to assist in the collection of opinions regarding the topic.

5. Machine learning models – Within this research's scope, sentiment analysis and news classification will be carried out. As a consequence of this, machine learning models are utilised so that the implications of the emotion can be classified.

6. Deep learning models – The approaches of deep learning are utilised in order to make classification regarding the outcomes of the text news classification.

7. Evaluation of results – In this particular area, the findings that were gained from this project will be compared with the findings that were obtained by the researchers who came before.

1.4 Work plan

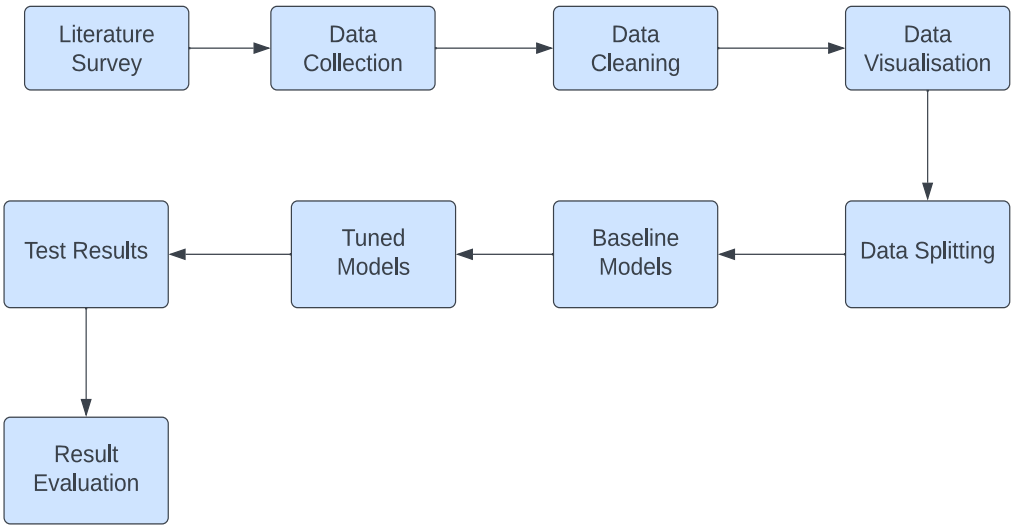

Figure 1.4(a) Work plan of the project

The project's operational strategy is depicted in Figure 1.4(a). The project will begin with a literature survey, during which time all of the information that is pertinent to the issue will be gathered. After then, the data collection process will be carried out, during which numerous online sources' data will be compiled. The newly acquired data will then undergo the process of data cleaning after that data splitting will be performed and the results of the split will be used in machine learning and deep learning

models. Afterwards, the data will be split. The outcomes classified by models will be examined and analysed.

1.5 Change in goals or methods

For this project, the procedures begin with the literature review, which is where the material pertaining to this subject will be gathered. The enormous dataset will be gathered through web scraping and it will be unique. After that, the text data will be cleaned by reducing the amount of text, applying lemmatising, eliminating punctuations, removing missing data, or at some points, replacing the null values with zero, along with a great deal of other text pre-processing techniques. Following that, the data will be visualised, and the results of that visualisation answered an initial couple of study questions. Now that the features have been selected and the data will be separated, the data will be prepared for modelling. After that, various machine learning models, including multinomial naive bayes, random forest, xgboost, and stochastic gradient descent, will be utilised in order to make forecasts regarding the respondents' feelings. The machine learning models that are going to be adjusted here will be compared to machine learning models that have been utilised in the past by other researchers. After that, deep models like cnn, lstm, gru, and mlp will be utilised in order to make classification regarding the news. These deep learning models will be one of a kind, and new types of hyperparameters, such as hyperband and random search, will be implemented into the deep learning models. After that, an analysis of the results will be carried out. In conclusion of this part, the data set that is used for this research will be enormous, and the models that will be developed for this project are one of a kind when compared to studies that other researchers have done in the past.

1.6 Structure of the report

The following chapters constitute the structure of this project:

Chapter 2 – Context – This part offers the understanding that was put into practice in this study based on what was found in the previous research. Additionally, it discusses the ways in which this project might be enhanced according to the results of the earlier investigations.

Chapter 3 – Methods – The methods of data collecting, data cleaning, data visualisation, data splitting, machine learning models, and deep learning models are briefly described in this chapter.

Chapter 4 – Results – This chapter discusses the findings that were achieved through the use of data visualisation, machine learning models, and deep learning models.

Chapter 5 – Discussion – Discussions will centre on the uniqueness and effectiveness of the obtained results, as well as the algorithms' designs and whether or not they achieve their goals.

Chapter 6 – Evaluations, Reflections, and Conclusions – This chapter includes an evaluation of the outcomes obtained from using machine learning and deep learning models, as well as some thoughts on the accomplishments made and a brief discussion of the directions in which research could go in the future.

Chapter 2. Context

2.1 Overview of sentiment analysis and news classification

This section gives an overview of the many different machine learning approaches that were utilised for the analysis of sentiment. In addition, this part offers a summary of the numerous deep learning strategies that were applied during the text classification process. Both of these subjects will be discussed in greater detail in the following paragraphs. Comparisons will be made between the research studies that are carried out by other researchers, and any methodologies derived from those studies that appear to have the potential to be applicable to this project will be taken from other studies. In addition to this, the manner in which the data set that the other researcher used in order to carry out their study will be the centre of attention as it pertains to the investigation. In addition to this, the models that are used by the other researchers, namely the tuning procedures that are used by the other researchers, will be explored in great detail.

2.2 Current state of the project

2.2.1 Sentiment analysis models

Here research paper regarding machine learning models is studied as a part of the literature review, and the techniques which are suitable for this project will be applied. It is possible for a text to relate to more than one category; therefore, "text classification" involves dividing textual data into several groups (Dien et al., 2019, p. 79). For classifying the text (Dien et al., 2019) have implemented machine learning algorithms such as knn, naïve bayes and svm, but the for the purpose of this study, the learnings from naïve bayes which is implemented by (Dien et al., 2019) is utilised for this project, but the design for the naïve bayes model which is implemented for this project is unique. (Dien et al., 2019) has implemented the naïve bayes model for text classification; after studying this paper, it is decided to try multinomial naïve bayes for the classification of the sentiment, which will be a unique version as compared to (Dien et al., 2019). Also, the multinomial naïve bayes model will be tuned so that performance of the multinomial naïve bayes baseline model will be enhanced. When the research paper of (Rana et al., 2016) was read, it was discovered that the naive bayes model had received decent accuracy; however, the study is limited, and only parameters such as accuracy recall score and precision score are obtained; furthermore, there is neither a roc curve nor a confusion matrix present, nor is there any tuning of the baseline model, and after reviewing the findings of the research conducted by (Rana et al., 2016), it was decided to incorporate performance parameters into the multinomial naive bayes algorithm. These factors include the fbeta score, mean absolute error, mean square error, and root mean square error. So, the performance parameters such as the accuracy, precision score and recall score or any other performance parameters which are obtained by the research of the mentioned papers will be compared with the performance parameters of this project.

The next step is the random forest machine learning model, which is going to be responsible for the classification of feelings. (Karthika et al., 2019) have implemented the random forest for the classification of the sentiments. Although (Karthika et al., 2019) have achieved good accuracy, the roc curve for the random forest baseline model is not appropriate, and the dataset used is existing data. After studying this paper, the plan is to improve the baseline random forest baseline model and try to run the hyperparameter tuning on the random baseline model. Every branch in the sample is randomly chosen from the "training" dataset; the sampling and ranking procedures are consistent across all trees, and after all of the trees in a forest have classified something, the "final classification" is determined by taking the statistically middle ground between all of the different categories (Kalcheva et al., 2020, p. 2). After looking at (Kalcheva et al., 2020), it has been discovered that this research made use of the cross-fold validation method, but it did not end up with a higher performance score than (Karthika et al., 2019). Also, almost all of the performance parameters of the random forest baseline model will be calculated, and hyperparameter tuning will be introduced to the baseline random forest model. Finally,

the random forest baseline model will be implemented, but the cross-fold validation will not be performed; instead, the data will be split in a different way.

Following that, the classification of sentiments is going to be carried out utilising stochastic gradient descent. The research conducted by (Antonio et al., 2022) was used for this article. According to the findings of the research carried out by (Antonio et al., 2022), text classification was performed on the Twitter dataset using the stochastic gradient model. This was done in accordance with the findings of the aforementioned researchers. (Antonio et al., 2022) have received an accuracy score of 81.7 per cent, it has received a precision score of 89.3 per cent and a recall score of 86.2 per cent, which is good; however, the data set used by (Antonio et al., 2022) was not vast in comparison to the dataset which would be used for this project. As can be observed in work by (Antonio et al., 2022), the data splitting that was done by (Antonio et al., 2022) is into 70: 30. This implies that (Antonio et al., 2022) has allotted 70 per cent of the data for the training and the 30 per cent of the data for the testing. If this works for all of the models, which means after separating the data into the 70 per cent training and 30 testing side, then it will be applied; if not, the data will be split into the 80 per cent training and the 20 per cent testing side. In addition, the stochastic gradient descent algorithm performance parameter, other than the accuracy, precision, and recall, will be mentioned; additionally, the stochastic gradient baseline model will be tuned with the hyperparameters that are appropriate for this machine learning algorithm. In addition to this, the roc curve and the confusion matrices for the stochastic gradient descent process will be shown.

Now, the research-related xgboost machine learning model will be analysed using the one research paper that was produced by (Hama Aziz et al., 2021) in the field. The sentiment was analysed with xgboost and then categorised by (Hama Aziz et al., 2021). Also, while reading the article by (Hama Aziz et al., 2021), it was discovered that the text pre-processing is carried out in a specific order, beginning with the most basic steps (Hama Aziz et al., 2021). The authors of the study (Hama Aziz et al., 2021) started by tokenising the text, then they deleted the stop words, and finally, they applied the stemming to the words. Following an investigation into this topic, a number of text pre-processing strategies have been implemented here. However, for the purpose of this research, these strategies have been implemented on the vast data, whereas (Hama Aziz et al., 2021) have utilised the tiny datasets. In addition, (Hama Aziz et al., 2021) have occupied a variety of performance criteria, including accuracy, precision score, recall score, and f1 score. One factor that sets the work of (Hama Aziz et al., 2021) apart from that of other researchers is the f1 score, which those other researchers did not adopt. The accuracy for the xgboost machine learning model obtained by (Hama Aziz et al., 2021) is 90.8, the precision score obtained by (Hama Aziz et al., 2021) is 92.7, and the recall score obtained by (Hama Aziz et al., 2021) is 98.1, and the f1 score obtained by the (Hama Aziz et al., 2021) is 94 per cent which. The same level of performance is going to be sought after with regard to this project. Also, (Hama Aziz et al., 2021) have not utilised the hyperparameter to boost the performance of the model that will be applied to the xgboost model that will be designed for this project, so the performance of the xgboost model will be increased. This is something that will be done in order to increase the performance of the xgboost model.

2.2.2 News classification analysis models

The initial model that is being investigated for the text news classification is being designed by (Zhao et al., 2022). (Zhao et al., 2022) has completed the implementation of the three convolutional layers, which are then followed by the three enhance pooling layers, and finally a softmax layer is implemented. 100 is the dimension that is involved in the embedding process for each of the three convolution layers that were built by (Zhao et al., 2022). The precision score, the recall score, and the f1 score are the performance parameters that were obtained by (Zhao et al., 2022) for the cnn. The average precision score acquired by (Zhao et al., 2022) for the cnn is 94.61, whilst the average recall score obtained by (Zhao et al., 2022) for the cnn is 93.43, and the f1 score for the cnn created by (Zhao et al., 2022) is 94.01. These performance characteristics for the “SogouCS” dataset were produced by (Zhao et al.,

2022, p. 16). Following the reading of this document, the subsequent step was to refer to the research paper written by (Duan et al., 2020). According to (Duan et al., 2020), three convolution layers have been examined for cnn a, and the bi gru layers that come after each convolution layer have been merged together. After the three convolution layers and the two bigru layers, (Duan et al., 2020) utilises dense layers, and then the softmax layer is taken into consideration as the final step. The data from "SogouC" have a precision that is 91.32 percent accurate (Duan et al., 2020, p. 140). (Deng et al., 2022) was also responsible for the design of the cnn, which is an amalgamation of the gru. (Deng et al., 2022) utilised four gru layers and three convolution layers, which are followed by three max pooling layers and a fully connected layer, and finally the softmax layer was utilised as the final layer. (Ahmed et al., 2021) have created the cnn with a single convolutional layer, which is followed by the "relu" activation layer, and it is subsequently accompanied by the "max pooling layer", and it is the flattened and dense layer that is applied (Ahmed et al., 2021, p. 611). Following the reading of these papers regarding the cnn, a brand-new design for the cnn will be developed for this project. In addition to this, the adjustment of the hyperparameters will be implemented into the cnn baseline model.

The research study that was conducted will now serve as the basis for the investigation of the lstm (Vimali et al., 2021). The bi-lstm model was developed for the research carried out by (Vimali et al., 2021). The structure of this model includes three forward lstm layers as well as three reverse lstm levels. The aggregation receives these forward lstm layers as well as the backward lstm layers, and then the fully linked layers receive them after that. On the enormous data set that will be used for this project, the text news categorisation will be carried out. Because of this, the lstm will need the gpu; hence, the cudnnlstm will be utilised for this purpose. Also, (Vimali et al., 2021) have found that their method has an accuracy of 90.46. During the course of this project, a whole new lstm architecture will be developed, and the baseline model will be enhanced by the addition of hyperparameters.

Research on the gru model is accomplished by referring to research (Duan et al., 2020). (Duan et al., 2020) created the bi-gru, which consists of clusters of gru layers that are divided by forward and backward layers. (Duan et al., 2020) used these clusters to create the bi-gru. The "SogouC" dataset has been given an accuracy rating of 88.62 according to the work of (Duan et al., 2020, p. 140). After completing this study, the next step will be to design the new gru model, after which we will attempt to train and test the model. In addition to that, hyperparameters are going to be applied, and loss is going to be determined. Along the same lines as lstm, cudnngru will be developed to manage the enormous datasets. Regarding the performance parameters of the cudnngru baseline model, some examples of these are accuracy, precision score, f1 score, fbeta score, mean square error, and root mean square error. Additionally, three additional study papers were cited in relation to the gru; however, researchers have already developed the gru, which is a combination of the cnn and the gru, and the gru model is designed independently in this project.

MLP will serve as the final deep learning model that is applied. A couple of publications are looked at in order to research the mlp approaches. (Shiguihara et al., 2022) began their research by employing an automated mlp classifier. This was done so that they could conduct their study. The f1 score and the amount of time were determined to be the performance characteristics established by (Shiguihara et al., 2022). (Shiguihara et al., 2022) have accomplished a score of 58.62 percent on the f1 test. The mlp classifier (Shiguihara et al., 2022) requires 48.48 seconds of processing time to complete its analysis. Secondly, (Tang et al., 2022) likewise built an mlp model, and the performance parameters that were utilised by (Tang et al., 2022) are also the f1 score and the amount of time it took for the model to train. (Tang et al., 2022) have obtained the maximum accuracy of an mlp deep learning model, which is 86.35 per cent, and the amount of time that model requires to train is 36 minutes and 8 seconds. Following this analysis of mlp, it was decided to start from scratch when designing the mlp model and to employ various tuning strategies when working on the model in order to improve its overall performance.

2.3 Conclusion

Concluding the entire section of context, the important findings found are that machine learning models such as knn svm, naive bayes, linear regression, xgboost, stochastic gradient descent, random forest, and decision tree have been utilised by a significant number of researchers for the purpose of doing sentiment analysis. As part of the process of conducting sentiment analysis, a large number of research papers are evaluated for their inclusion in the literature review. However, it was discovered that the research studies had based their work on an already existing dataset, which meant that the dataset did not have any distinguishing characteristics. In addition, no researchers utilised hyperparameters to improve the efficiency of machine learning models. For the purpose of conducting sentiment analysis for this project, different machine learning models, including multinomial naive bayes, random forest, xgboost, and stochastic gradient descent, will be constructed. For the purpose of this project, the data will be distinctive and substantial, and in order to improve the overall performance of the machine learning models, each of those models will be trained utilising the hyperparameters. In a similar vein, while doing a literature analysis for the purpose of classifying text news, it was discovered that many researchers had utilised models such as cnn, lstm, bi-gru, bi-lstm, mlp, and gru. In addition, the research articles that have been analysed for text news categorisation have made use of the dataset that was already in existence, and they have not applied the hyperparameter tuning that is necessary to improve the performance of the deep learning models. After reading the research papers, the next step will be to collect a massive dataset from a variety of locations, then create one-of-a-kind machine learning and deep learning models, and finally, use hyperparameter tuning to improve the models' overall effectiveness.

Chapter 3. Methods

3.1 Data collection

In order to obtain the necessary data for this project, many different sources were researched in order to find the relevant data. In the end, it was concluded that web scraping would be the best way to get the data. This project's data collection was obtained with the use of an application programming interface (api) from a piece of software known as octoparse 8 and from the online app News data io. The data scraped from online news websites from various countries such as United Kingdom, India, Japan, United States of America, Australia, Philippines, Singapore, Canada Pakistan, Malaysia, Ireland, China, Nigeria, France, South Africa, Saudi Arabia, United Arab Emirates, Netherland, Thailand, Cuba, South Korea, Germany, Slovakia, New Zealand, Israel, Italy, Russia, Hong Kong and Vietnam. The name of the companies from which the data was acquired are Mail Online, Buzzfeed, Glasgow Times, The Hindu, The Japan Times, Huffpost, Cairns Post, Salon, Perthnow, Indaily, Scottish Field, Evening Standard, Ktla5, Laist, Journal Online, The Guardian, Yo Vizag, Google News, The Sportsrush, The Washington Times, Americas Newspaper, New York Post, Profootballtalk, Nine Newyork Cbs Local, Malay Mail, Inforum, Fox Sports, Detroit Free Press, Ottawa Sun, Pff, 411mania, Belfastlive, Lasvegas Review Journal, Kron4, Montreal Gazette, Teamtalk, Manchester Evening News, The News International , La Times, The Star, Dodgers Digest, The Province, Chicago Tribune, Zee News, Cnn, Independent, Toronto Sun, Business Standard, Si, India Today, Orissa Post, Toronto Star, Talksport, South China Morning Post, Kfor, The Times Of India, Ndtv, The West, Sports Star, The Sun, The News, The Irish Sun, The Bolton News, Dorset Echo, Express, The New York Times, Prince George Citizen, Hitc, Espn, Daily Post, Daily Record, Sports Mole, News 18, Leadership, Afloat, Telangana Today, Kgw, Herald Sun, Mirror, The Scotsman, Sky Sports, Planet F1, France 24, The Sentinel, Uf, The42, Hindustan Times, Telegraph Argus, Iol, Birmingham Live, Kildare Nationalist, Breakingnews, Premium Times, Brisbane Times, Wales Online, Metro, Football365, Autosport, Grimsby Live, The Argus, The Tribune India, Al Bawaba, Citynews Everywhere, India Tv, Mint, The Washington Post, Rt, The Scottish Sun, The Paris News, Cbc, The Daily Telegraph, Sportsnet, Realgm, Pr Newswire, Thunder Bay News, The Press, Oakville News, Sanfrancisco Cbs Local, Minnesota Cbs Local, Ctv News, Op India, The Economic Times, Money Control, Racefans, Chek News, The Northern Echo, The Manila Times, Highland Radio, Golf, Richmond News, Punch, The Herald, Kentonline, Dutchnews, Abc7, The Hans India, Kashmir Reader, Northern Ireland World, The Standard, Cbs Boston, Wfaa, Belfast Telegraph, Halifax Today, Bangkok Post, Startribune, Wcpo, Wfla, Abc 30, Global News, La Prensa Latina, Dna, Abc 15, Sootoday, Ewn, The Insidexpress A Central Magazine, Gud Story, Firstpost, Liverpool Fc, 6abc, The Korea Herald, Tng News, Umhoops, Citynews, Information Nigeria, Edmonton Journal, Abc 13, Euronews, Dfw Cbs Local, Globe Newswire, Vc Star, The Scarborough News, 247 News Around The World, Krook News, Watoday, Star Of Mysore, The Week, Regina Leader Post, Kitchener News, Southern Daily Echo, The Star Phoenix, Castanet, The Record, Can, India Observers, The Sydney Morning Herald, Hypebeast, The Hamilton Spectator, The Northlines, Ns News, Deadline, The Hindu Business Line, Informal News, Narcity, Tech Crunch, Calgary Herald, Itb Berlin News, National Post, Whatculture, Deccan Chronicle, Spectator, The Headlines Of Today, News Blare, Burnaby Now, Quint Daily, The Age, Vancouver Sun, One India, The Poke Time Well Wasted, Mandatory, Times Chronicle, Newstalkzb, Voxy, Nzherald, The Times Of Bengal, Mercury, The Courier Mail, The Asian Age, Saturday Down South, Dkoding, Articleify Enlight Your Creative World, Curry Coastal Pilot, Ars Technica, Medical News, Lifehacker, The Bbc, Npr, Globes, Cbs News, Ottawa Citizen, Medical Xpress, Financial Express, Read To Lead, Nbc News, Science Daily, Tech Explorist, Necn, Fox News, Politico, Global Press Journal, Cnbc, 9news, The News Minute, Mail Guardian, The Eastern Herald, The French Daily News, Tentaran, Rand Objective Analysis Effective Solutions, Brookings, News 4 Masses, Startup Reporter, Moneycontrol, Forbes, Marketwatch, Fool, Bq Prime, Bitcoin News, Gaadi Waadi, Clean Technica, Financial Times, Carandbike, Thejournal, Businesstech, The Business Times, The Irish Times, Wsj, Ynet News, Realestate, Oil Price, Chandigarh Metro, Techgenyz, Wired, The Hollywood Reporter, Phys.org,

Ispreview, The Economist, Itweb, Thenewstrace, Ips News, News Blaze, Positive News, Dexerto.es, Crikey, Uj City News, Pravda, 4 News, Right Wire Report The Truth Be Told, Tech News, Cointelegraph, Finextra, Brampton News, Kraken Blog, The Shovel, Blockchain News, Ripple Coin News, Let Me Think, Nbr, Consett Magazine, Emit Post, Reuters, Bitcoinworldcoin, Benzinga, Thenewscrypto, Comic Book, Showbiz Cheatsheet, Etcanada, Soompi, Kotaku, Billboard, Gma, Variety, Dark Horizons, Nintendo Life, Interaksyon, Et, Coming Soon, Tmz, Vulture, Av Club, Bollywood Hungama, Nme, Daily Research Plot, Vg247, Push Square, Rolling Stone, Cinemablend, Gamespot, Hola, Gizmodo Australia, Geek Culture, Charlie Intel, Buzz Feed News Reporting To You, The Christian Science Monitor, Media Mole, Pink News, Cambridgeshire Live, Motor1, See Latest, The Canadian Jewish News, Politicalite, Mashable, Ign Southeast Asia, Surgezirc United Kingdom, Rabble Ca, Usa Today, The Maritime Anchor, The Internet Protocol, Hypertext, Joe, Docbao, Rfi, National Herald, Macleans, The News Himachal, Politics Co Uk. The information is collected in the following seven categories: business, sports, entertainment, environment, health, and food. Politics is the seventh category. Due to the fact that the dataset was scraped with the assistance of software, the format of the dataset is in the csv format. The total number of rows that result from integrating all the datasets is 823105, and the total number of columns that result from doing so is 13. The news in the all the datasets is between 1st May 2022 to 30th September 2022.

3.2 Data cleaning

After collecting the data from a wide variety of news websites, all of the datasets were imported, and cleaning was carried out on each of those datasets individually. To begin, the food dataset was given a thorough cleaning. To begin the process of cleaning up the food dataset, the information contained inside the food dataset was first gathered. Following that, the published date of the food dataset was converted into the datetime format, and a column titled datetime was added to the food dataset. Following that, the date and time were removed from the dataset, including the food. The datetime column was removed after it was determined that its presence was no longer necessary. After then, a check was performed on the values that were missing. It was discovered that the columns for an author, image url, and video url were missing some of their respective data. These missing values were replaced with the unknown rather than the zero that would have been expected to stand in their place. Because the food dataset is scraped, in addition to the food class, there were additional classes that were also present; however, these were disregarded, and only the food class was taken into account. After this, it was discovered that there were certain titles that were repeated, and these were required to be removed. Because of this, we didn't include the titles that appeared more than once in the dataset. Additionally, it was discovered that the entire dataset had a significant quantity of emoticons; hence, the emoticons were eliminated from the dataset. Now the text-cleaning process was carried out on the food dataset. Within this section, the text-cleaning process is carried out on the description column of the food dataset. To begin, the text that is included in the description has been changed so that it is all lowercase. Following that, the punctuations that were included in the food dataset's description were taken out. Following that, the stop words that had been included in the description column of the food dataset were eliminated. Following the removal of the stop words, the numbers that were present in the description column of the food dataset were eliminated. After that, the text underwent tokenization, and after that, parts of speech tagging were carried out on the text. After that, the lemmatization technique was used on the verbs that had been obtained from the tagging of the various parts of speech. A data cleaning method that is analogous to the one described above is applied to the datasets pertaining to sports, business, politics, the environment, entertainment, and health, respectively. After the completion of all of the datasets respective cleaning processes, these datasets were finally joined.

3.3 Data visualisation

This section is subdivided in to the two sections which is data visualisation of the sentiments and the data visualisation of the news classification. For the data visualisation python plotly and tableau was utilised.

3.3.1 Data visualisation of sentiments

In order to begin the process of data visualisation of the opinions, the aggregated dataset was initially imported into a separate file. First, a check was done to see whether any values were missing. After combining the datasets, it was discovered that this led to the disappearance of some of the values. Because of this, the values that were absent from the author and description columns were eliminated. After that, a decision was made to acquire a visualisation of the total count as well as the percentage of the sentiments. In order to obtain the feelings, the polarity of the description column text was extracted, and on the basis of that, the positive, the negative, and the neutral sentiments were obtained. After then, the columns for the month, the week, and the hour were taken from the date column. After that, the counts of the sentiments are taken and plotted on the bar plot using plotly. This may be seen in the results area of the website. On the doughnut chart, the percentage of the sentiments has also been plotted. The next step that was decided upon was to compile a count of the opinions expressed within each subcategory. Because of this, the count of each sentiment with regard to the feelings was taken into consideration. In addition, it was determined to determine the average feelings held toward each nation. In order to do this, the pivot table of the sentiments, in which the feelings are the rows and the country columns are the columns, was established. This information was then plotted on the line chart with plotly.

3.3.2 Data visualisation of news

After carrying out the procedure of data visualisation of the opinions, it was agreed upon that the insights for the news should be obtained. To begin, it was decided that a count and percentage breakdown of the news categories should be obtained. As a result, the count value of the different news categories was taken into consideration, and the proportion of the news that is displayed on the pie chart, while the count is plotted on the bar plot. After then, it was agreed to get the standard number of news stories for each country. As a result, the pivot table was constructed for the nations and categories, and then it was plotted on the line chart with plotly. After that, it was determined to collect the news on a daily basis during the week for each news category. On the basis of the stacked bar plot, it was obtained. After that, it was determined to find out how many average news stories are published on weekdays, and this would be done by grouping them according to the category and the day of the week. Then a decision was made to get the average number of news items published each month, and this information was plotted on a bar plot with plotly, and both the month column and the category column were taken into consideration for this plot. The decision was made to get the news once every hour for each of the categories. Plotly is used to plot the results of the comparison between the hour column and each category column on a line chart. The decision was made to obtain the news once a month for each of the different categories of news. Python Plotly was utilised in order to plot this chart, which involved taking into consideration the category column as well as the month column. The chart was then plotted on a heatmap. The next step was to come to a decision regarding how to get charts that would show which nations generated the most amount of news between May 1 and September 20, 2022. This plot was created with the help of tableau software. Also, the future prediction has done of the of the 150 days and it is plotted on the line chart.

3.4 Data splitting

For the models of sentiment analysis, the sentiment column is chosen to be the output, while the description column is chosen to be the input. Eighty percent of the data for the machine learning models that will be used for the sentiment analysis has been set aside for the training phase, while the remaining twenty percent has been set aside for the testing phase. This distribution of the data is referred to as the 80:20 split. The method of label encoding is utilised so that the sentiment column can be transformed into the labels. In a similar fashion, for the deep learning models, the category column is regarded as the output, while the description column is chosen to serve as the input. Additionally, the data is split in a ratio of 80:20, with 80% of the data being utilised for the training side and the remaining 20% being

utilised for the testing side. In addition, the output, which is the category column, is then encoded using the label encoding method before being transformed into the labels.

3.5 Sentiment analysis model design

For sentiment analysis the machine models such as the multinomial naïve bayes, random forest, stochastic gradient descent and xgboost model is designed by the taking the motivation from the various research papers. Also, the design of the models is changed and also the model is applied on the huge dataset. Also, the hyperparameters for each machine learning model is applied to boost the performance of the machine learning models and performance parameters more than accuracy, precision and the recall is chosen and it is compared with the previously designed models by other researchers.

3.5.1 Design of naïve bayes model

Firstly, the multinomial naïve bayes model is designed for this project which is different from the (Dien et al., 2019). The basic principle of this method is to forecast the topic likelihood of a classed text by utilising the "conditional probability between words or phrases" and subjects (Dien et al., 2019, p. 79). (Dien et al., 2019) have applied the naïve bayes model on the software. Also, the text conversion converts the text data in to vector form is done in the "tf*idf" by (Dien et al., 2019, p. 80). In this project the text is converted in vectors by using the count vectorizer. Also (Dien et al., 2019) have used the naïve bayes model on the classifier software. For this project the multinomial naïve bayes is made using the sklearn library. Multinomial naïve bayes model is trained on the huge data set and the performance parameters such as accuracy, precision score, recall score, mean absolute error, mean square error, root mean square error. Also, the roc curve for the multinomial naïve bayes is plotted. The hyperparameters for the multinomial naïve bayes used are alpha, class prior and fit prior is applied to the baseline model and the multinomial naïve bayes tuned model is created.

3.5.2 Design of random forest model

Secondly, the random forest model is designed. The design of the random forest is motivated from the from various research paper. The "supervised learning algorithm Random Forest" is commonly employed for a variety of "classification and regression" tasks and the term "Random Forest" is used to describe a set of "trees" in which every "tree" is unique. When "training" and "outputting a class", it builds many "decision trees" and then "merges" them into a single, more accurate one (Karthika et al., 2019, p. 2). Following the completion of each single categorization by each tree, the overall result of the forest is determined by determining the "statistical" characteristic of all of the single classifications (Kalcheva et al., 2020, p. 2). Also, (Kalcheva et al., 2020) have used the automatic classifier for the random forest model. The change which is made in the model is that the random forest model is trained on the huge data and the model is designed on the sklearn library and the hyper parameters such as n estimators and the max features are applied. For tuning the baseline models grid search is used. The correct roc curve and the confusion matrix for both the baseline model and the tuned model is applied.

3.5.3 Design of stochastic gradient descent model

After the random forest model, the stochastic gradient descent model is designed. To determine the bare bones of usability, the SGD technique employs an "iterative optimization" procedure and at the outset of the procedure, the algorithm does the "tapping" and afterward, the "pushing error" is fixed by the guessing loop's application of the "gradient rule" to the target function (Antonio et al., 2022, p. 1368). The stochastic gradient descent model designed by (Antonio et al., 2022) is trained on the minimal data which is only on 830 tweets. As compare to the stochastic gradient descent designed by (Antonio et al., 2022) the model is trained on the 658484 news entries.

3.5.4 Design of xgboost model

It is also possible to refer to xgboost as the "Regularized Boosting technique" and this is due to the fact that it incorporates a number of regularizations, which work to reduce instances of "over-fitting" and to enhance the accuracy of categorisation (Hama Aziz et al., 2021, p. 564). For this project the xgboost

model is also trained on the 658484 news data and also the baseline model is tuned with the stochastic gradient descent hyper parameters such as gamma and max depth. (Hama Aziz et al., 2021) have used the meta classifier for the xgboost whereas for this the xgboost machine learning model is design using the sklearn.

3.6 News classification analysis model design

3.6.1 Design of cnn model

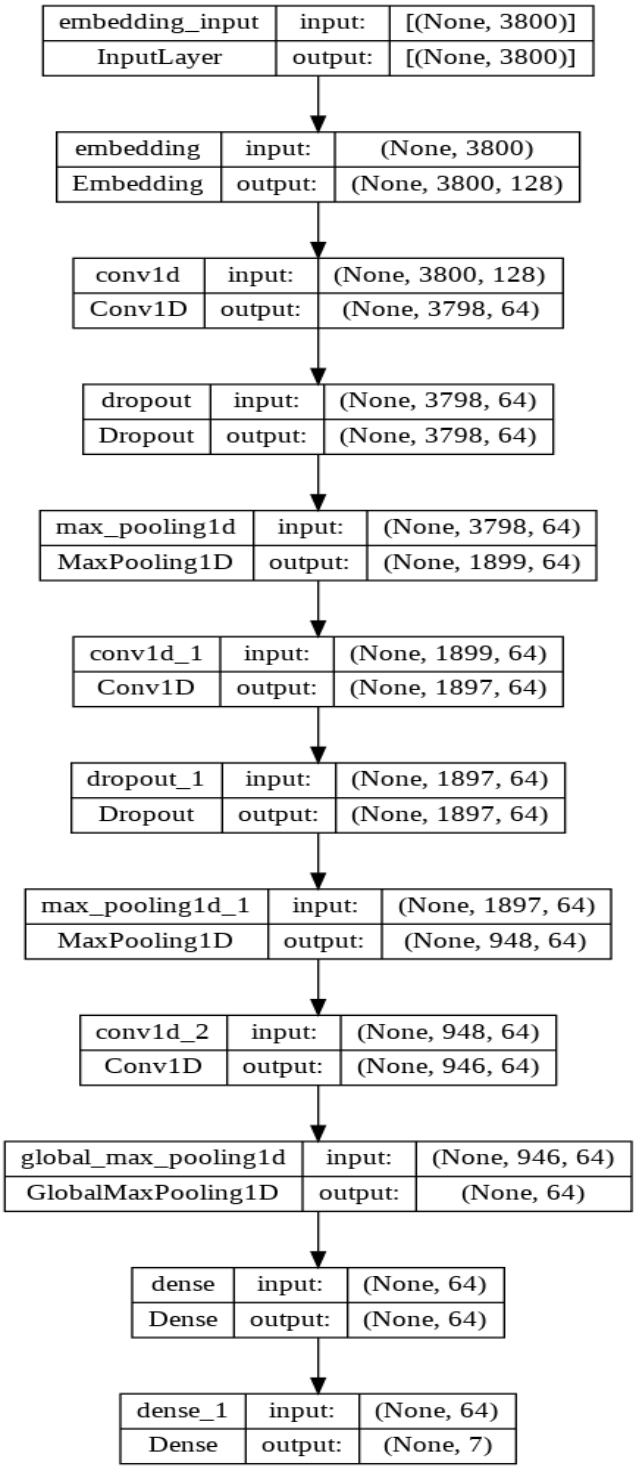

Figure 3.6.1(a): CNN model design

Convolution layer - There is no benefit to be gained from performing “natural language processing” using a “two-dimensional convolution of text data” because of this, the “convolution kernel” is only capable of performing sliding in a single dimension which means that the width of the "convolution kernel" is equivalent to the “dimension of the word vector” and when a convolution operation is carried out in a “window of height”, a newly discovered neighbourhood feature is retrieved after each iteration(Zhao et al., 2022, p. 2300). Equation 1 is given by (Zhao et al., 2022).

$$pi=f(w\cdot xi:i-h+1+\_b) \quad \text{.....(1)}$$

“pi represents the feature extracted by one convolution kernel in one con-volution, x represents the word vector with window h, w represents the convolution kernel, b represents the offset, and f represents the relu activation function” (Zhao et al., 2022, p. 2300). Following the separation of the data into training and testing sets, the data is tokenised, and then padding is added to the data set. Both the embedding dimension, which is set at 128 and the feature size, which is set at 1500, are taken into consideration for the CNN. The keras framework was utilised in the construction of this CNN model. To begin, an embedding layer was added to the model, and its feature size served as its constituent part. In addition, the embedding layer was composed of the embedding dimension, and the input length that was given to the CNN was 3800. The application of the feature size was made in an incremental manner. Now, for the first convolution layer, 64 filters are taken into consideration, and there are three kernel sizes taken into account. The three axes are represented by the kernel size. After the first convolution layer, the dropout layer was applied to the CNN model, which already had the output shape of 3798. This completed the process of creating the output shape of 3798. Max pooling layer – When you execute max-pooling, you will take the value that is the highest among all of the “feature values”, and then you will input that value into the layer that is “completely linked” (Zhao et al., 2022, p. 2300). Following this step, the first maximum pooling layer is used in the CNN model, which results in the output shape being reduced to 1899. Then the second layer of convolution was applied, this time using 64 filters and a kernel size of 3. The value 1897 is the shape that is output by the second convolution layer. Following this step, the application of the second dropout layer takes place, which does not alter the second layer output shape. After that, the 948 is used as the output shape for the cnn, which is accomplished by applying the 2nd max pooling layer, which reduces the output shape of the 2nd convolutional layer. After that, the third layer of convolution is applied to the CNN, it has 64 filters and a kernel size of 3. The number 946 is the form that is produced by the third convolution layer. The cnn model then has the global max pooling layer applied to it when this step is complete. In the final step, a couple of thick layers are added to the CNN model. The second dense layer has seven different categories. Following the construction of the baseline model, this model was trained, and the training loss in addition to the training accuracy were collected. Following this, the cnn baseline model was modified by utilising the hyper band technique, with the hyperparameters max epochs, hyperband iterations, objective, and the factor being utilised. Also, the cnn baseline model is tuned using the random search technique. Then, using these hyperparameters, four models are constructed, and each of these models is trained and tested using a variety of epochs and batch sizes.

3.6.2 Design of lstm model

The construction of the lstm model is depicted in Figure 3.6.2(a). The first layer of the lstm model is the embedding layer, and it has the feature size, the embedding dimension, and the input length. This can be observed from the design of the lstm model. 128 is the value for the embedding dimension. The lstm model is then given a first dropout layer to be applied to it. 3800 is the form that is produced as the output of the first dropout layer. The next step is to apply the cudnnlstm layer with 128 filters. This model made use of the GPU because the dataset it was working with was so large. After that, the dense layer is added to the lstm model while the relu activation is taking place. After that, the second dropout layer has a value of 0.1 applied to it, and then, as a final step, the dense layer containing all seven categories is added. After this the baseline model is tuned using the hyperband and random search

techniques and based on that the four models where created based on the different epochs and batch size.

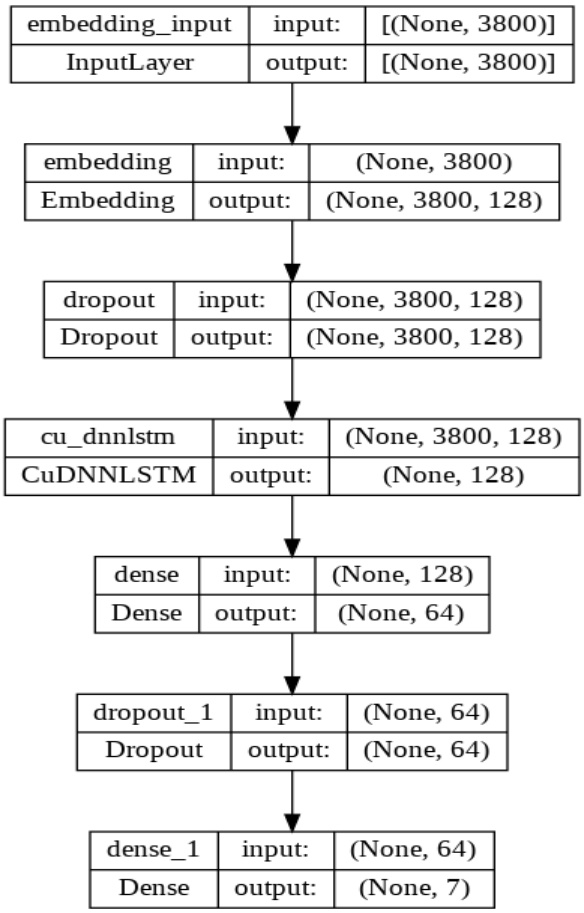

Figure 3.6.2(a): LSTM model design

3.6.3 Design of gru model

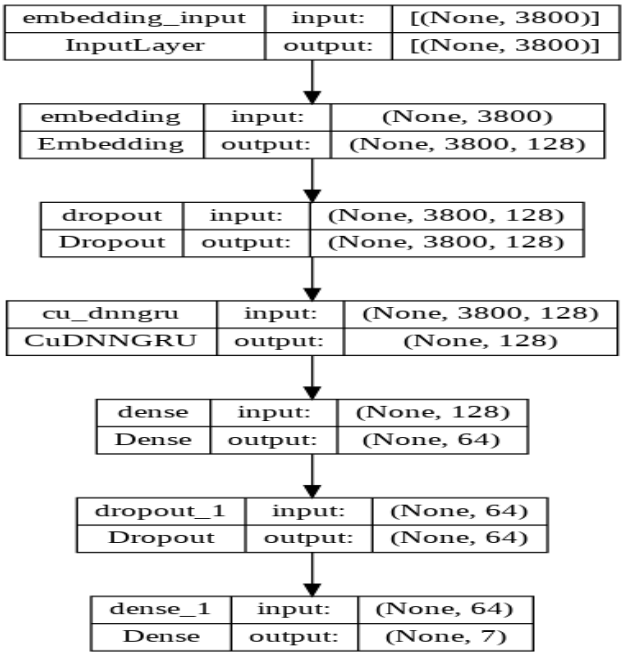

Figure 3.6.3(a): GRU model design

The blueprint for the gru model is displayed in Figure 3.6.3(a). The embedding dimension that was taken into consideration for the embedding layer, as can be observed, is exactly the same as that of the lstm model. Additionally, the feature size that is taken into consideration for the gru is comparable to that of the lstm. After that, the new model gets its very first dropout layer, which has a value of 0.1 added to it. The cudnngru layer is then applied to the gru model, along with 128 different filters. After that, the gru model receives its first dense layer, which is comprised of 64 filters and sigmoid activation. Following that, the 0.1 value for the second dropout layer is applied to the image. The final dense layer, the second one, contains seven classes and uses the softmax activation function. After that, the basic model of the organism is modified by adding the hyper parameters. The hyperband and random search tuning strategies are both employed in this process.

3.6.4 Design of mlp model

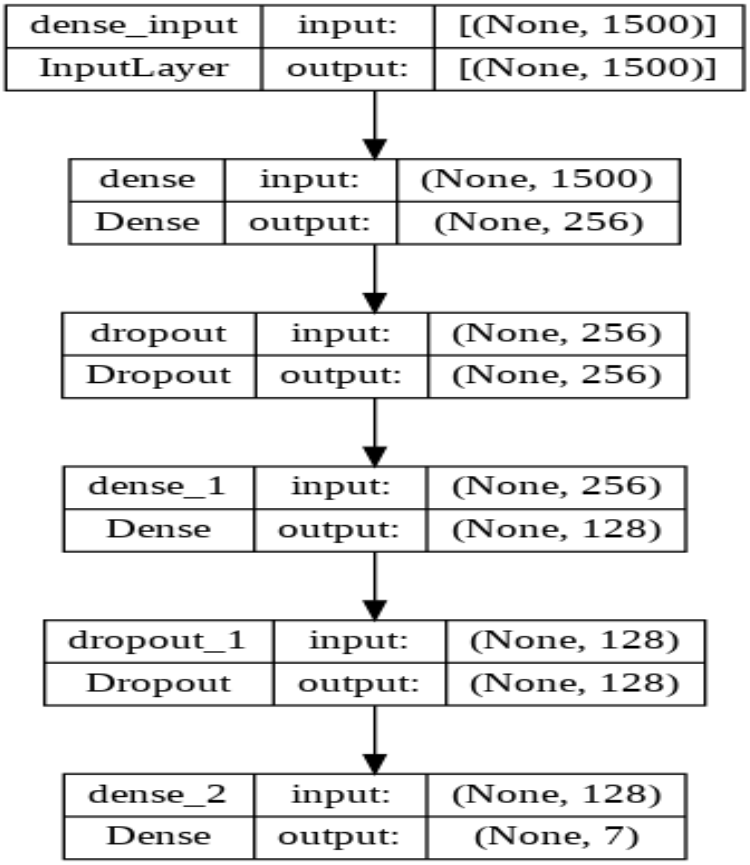

Figure 3.6.4(a): MLP model design

The mlp model design is shown in figure 3.6.4(a). As can be seen, the dense layer of the MLP model is initially applied with 256 filters and has an input shape of 1500. This is the first step in the MLP modelling process. After that, the 0.1 value for the first dropout layer is applied to the image. Following the application of that second thick layer to the mlp model, the relu activation and 128 filters are included in it. After that, the 0.1 value is utilised for the application of the second dropout layer. After that, the dense layer is applied using all seven classes and sigmoid activation as the finishing touch. Following this step, the baseline model of mlp is tuned utilising the hyperband and random search strategies of keras, and an attempt is made to improve the model's overall performance.

3.7 Specifications of tools

Name	Specifications
Tableau	2022.3
Keras	2.9.0
Sklearn	1.0.2
IDE	Jupyter, Google colab pro plus
CPU	12.68gb google colab pro plus, system cpu 2.10 GHz utilised.
GPU	83.48gb google colab pro plus
Nltk	3.7
Python	3.8.12
Pandas	1.3.5
Matplotlib	3.2.2
Plotly	5.5.0

Table 3.7(I): Specifications of the tools used for this project

Chapter 4. Results

4.1 Results of data visualisation of sentiments

This section's goal is to offer a descriptive response to the first research question.

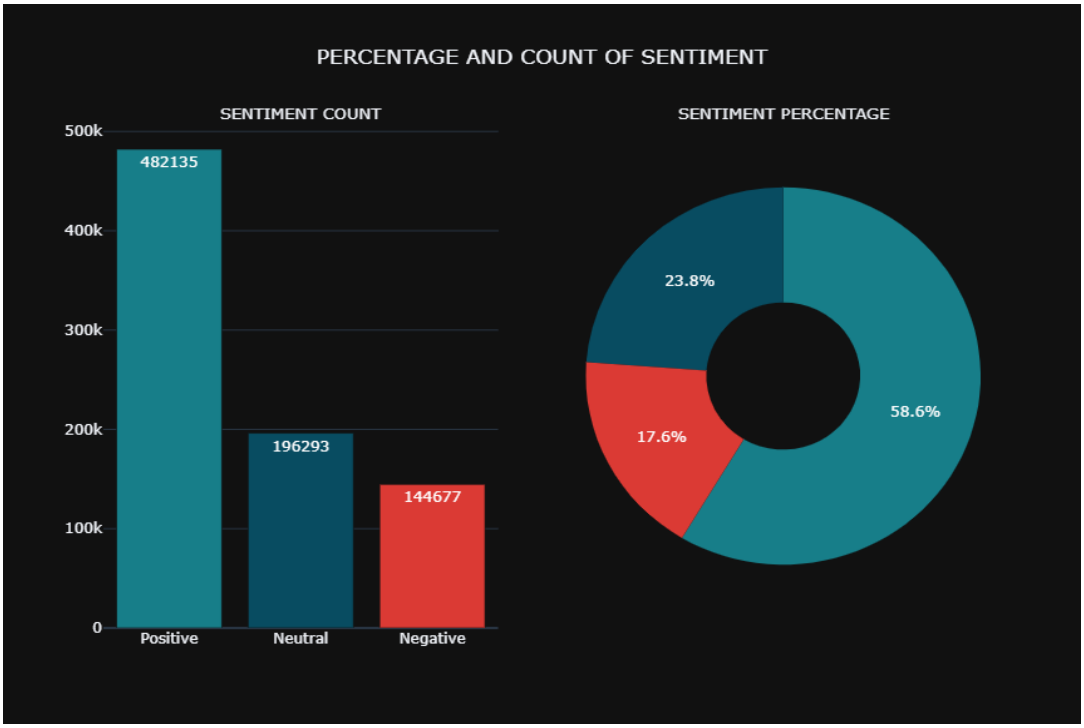

Figure 4.1(a): Percentage and count of sentiment

The polarity score provides insight into the respondent's feelings. It was carried out on the seven different types of news, including sports, entertainment, environment, business, politics, food, and health. Figure 4.1(a) illustrates the percentage of responses that fall into each of the three categories of positive, neutral, and negative feelings, as well as the total number of responses. Between May and September in the year 2022, people all throughout the world expressed a total of 482,135 favourable sentiments. The number of people who have a neutral opinion across all seven news categories is 196,293, while the number of people whose opinion is negative is 144,677. In addition, the percentage of sentiments of the people are also represented in figure 4.1(a). The positive sentiment percentage is the highest out of the three sentiments, which are positive, neutral, and negative. The positive sentiment percentage is 58.6 per cent, which is followed by the neutral sentiment percentage, which is responsible for 23.8 per cent, and finally, the negative sentiment percentage, which is 17.6 per cent.

The quantity of feelings associated with each news category is displayed in figure 4.1(b). Because there are seven different types of content, namely sports, entertainment, food, health, business, politics, and the environment, and three different feelings, namely positive, neutral, and negative, there are also seven different types of categories. 172.46k persons have a favourable attitude toward the sports news category as a topic of conversation. The number of people who have a negative emotion toward the sports news category is 50.45 thousand, while the number of people who have a neutral sentiment toward the sports category is 62.18 thousand. The category of sports news has the biggest amount of good emotion, which is greater than the value of 160k. Seven thousand three hundred ten people have a negative opinion of politics, while there are 17600 people who have a positive opinion of politics. The neutral sentiment for the politics news category is responsible for the 12.81k neutral sentiments, while the negative sentiment for the politics category is responsible for the negative sentiments. The number of people who have a neutral or positive attitude toward the health news industry is 7821, while the number of people who have a negative attitude is only 7336. The 25.05k represents the number of

people who feel positively about the health news category. Since the food category had the least amount of news and the least amount of news was available, the positive sentiment for the food category of news is 2089, while the number of negative sentiments for the food category is only 657, and the neutral sentiment for the food category of news is 1010. The number of people who feel positively about the environment news category is 4083, while the number of people who feel negatively about the environment news category is 1439. The number of people who feel neutral about the environment news category is 1759. Since the entertainment news category is the second highest category, the number of people who have a positive sentiment toward the entertainment news category is 162.45k, making it the second highest in the positive sentiment stage. On the other hand, the number of people who have a negative sentiment toward entertainment is 43.71k, and the number of people who have a

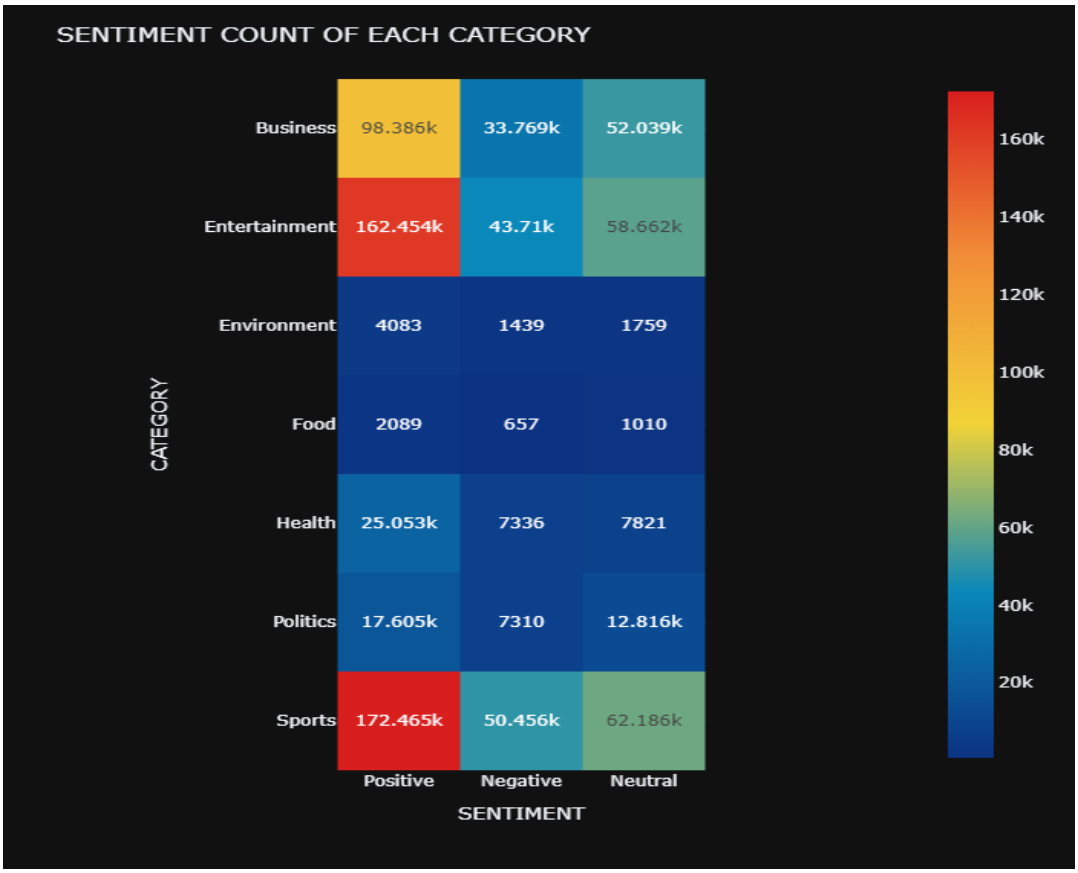

Figure 4.1(b): Sentiment count of each category

neutral sentiment toward entertainment is 58.66k. At long last, the business sector has reached the third highest position in the stage reserved for the news category. Because of this, the number of people who feel positively about the business news category is 98.38 thousand. The number of individuals who have a negative opinion regarding business news is 33.76 thousand, while the number of individuals who have an impartial opinion is 52.03 thousand.

The average number of sentiments collected from international news sources is seen in Figure 4.1(c). There are citizens here from the following countries: Australia, Canada, Chile, Cuba, India, Ireland, Israel, France, Germany, Hong Kong Italy, New Zealand, Nigeria, Pakistan, the Philippines, Russia, Saudi Arabia, Singapore, Slovakia, South Africa, South Korea, Thailand, Japan, Malaysia, the Netherlands, the United Arab Emirates, the United Kingdom, the USA, and Vietnam. In comparison to Australia, where the average is roughly 20000, the number of voices heard in Canada is slightly higher.

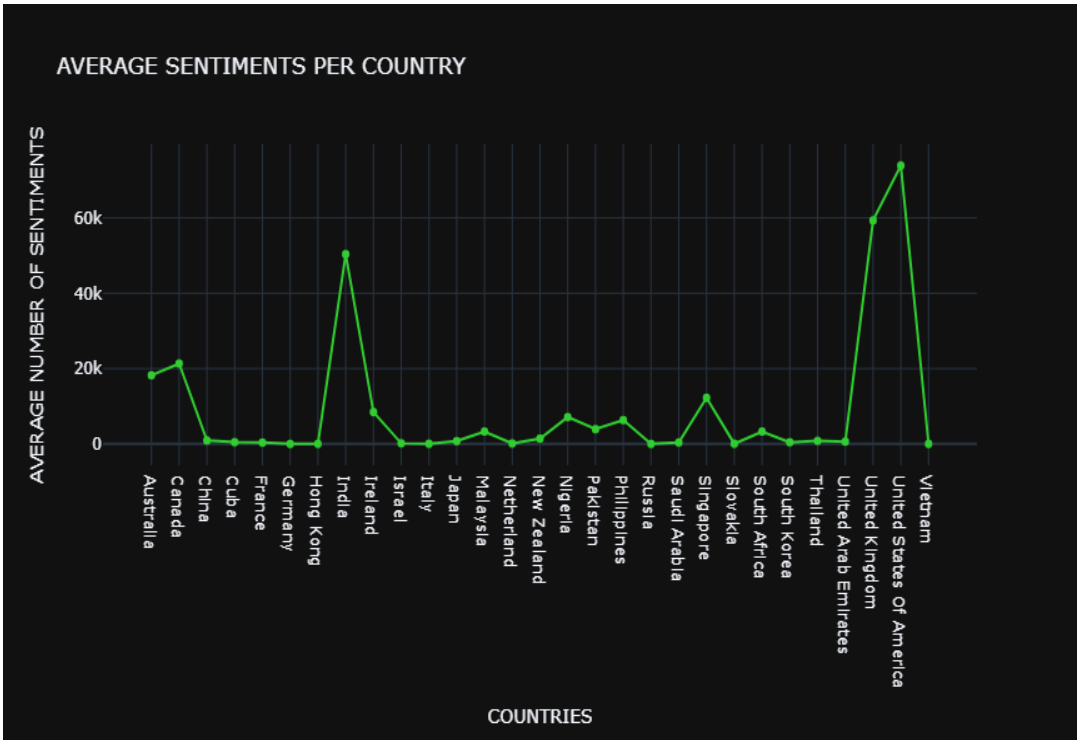

Figure 4.1(c): Average number of sentiments of each country

It has been stated that in Chile, Cuba, France, Germany, and Hong Kong, the number of recorded emotions between zero to one thousand. After that time period, the number of opinions shared towards India increased dramatically, going from the low hundreds to around 50000. Slightly over 8000 expressions of sentiment toward Ireland remain, which is a dramatic decrease. For the subsequent three nations, the range of possible responses is once again between zero to one thousand. For countries like Malaysia, the Netherlands, New Zealand, Nigeria, Pakistan, the Philippines, and Russia, the number of opinions can be anywhere from zero to 10 thousand. For Singapore, which now has over 12000 average sentiments, there appears to have been a substantial increase in the average number of sentiments. For countries like Slovakia, South Africa, South Korea, Thailand, and the United Arab Emirates, the range is zero to 5 thousand. The average number of opinions shared by Americans increased somewhat to roughly 74000. Whereas in the UK the number of expressions of emotion remained relatively stable at around 60k. The average number of people's views on Vietnam has finally dropped below 1000 roughly 74000. Whereas in the UK the number of expressions of emotion remained relatively stable at around 60k. The average number of people's views on Vietnam has finally dropped below 1000.

4.2 Results of data visualisation of news

This section's goal is to offer a descriptive response to the second research question.

As can be seen in figure 4.2(a), it provides both the percentage and the count for each of the seven different types of news stories, which are as follows: sports, entertainment, business, health, politics, and environment and food. As can be seen, the contribution of sports news to the overall news industry is the highest, coming in at 34.6 per cent. This is followed by the entertainment sector, which is the second highest contributor, coming in at 32.2 per cent, which is approximately 2 per cent less than the contribution of sports news. The percentage of the business sector news that is the third highest is 22.4 per cent. The news pertaining to health accounts for 4.89 per cent of the overall news percentage. Following that, the second-to-last and final sector is the environment, which contributes 0.885 per cent, and the food sector contributes 0.456 per cent of the total news percentage, respectively, given that this is a news dataset consisting of data from the past five months. Therefore, the total number of articles in the news connected to sports was 285,107 during the course of the five months. This makes the sports

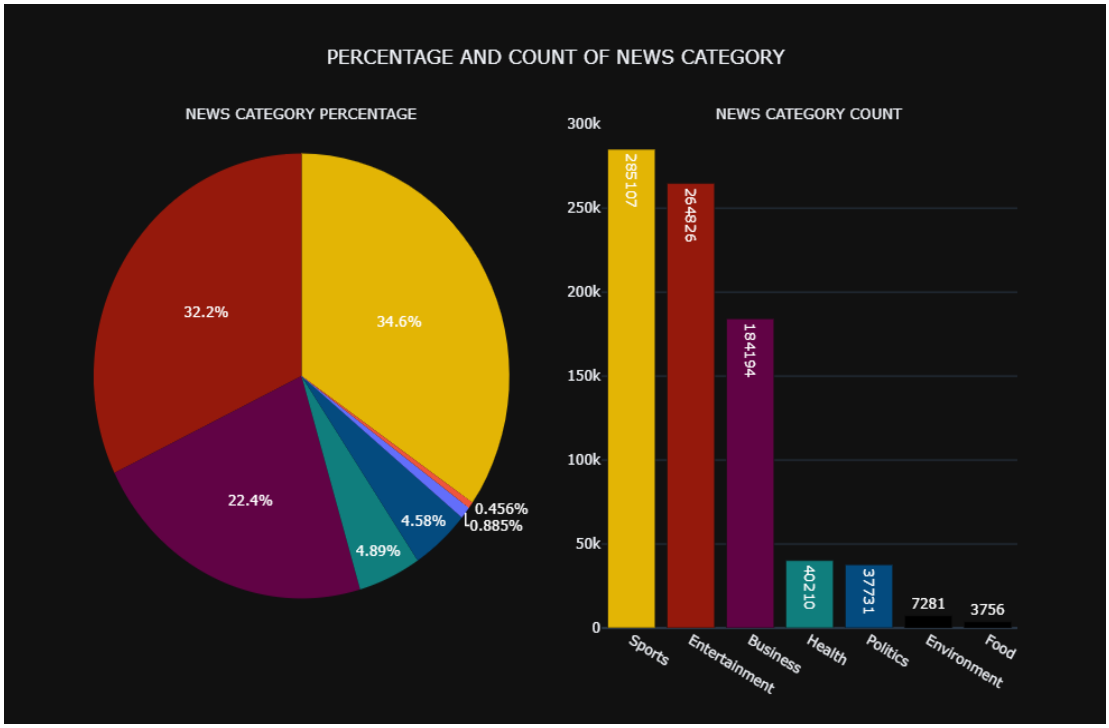

Figure 4.2(a): Percentage and count of news category

news category the most popular news category. The next category, business, is responsible for 184,194 business-related electronic items, which is followed by the entertainment sector, which provided 264,826 pieces of news. The news items pertaining to health are on page 402, which comes after the business sector. The current number of political news is 37731, which places it third from the bottom. The information regarding the environment and the food sector are, respectively, 7281 and 3756.

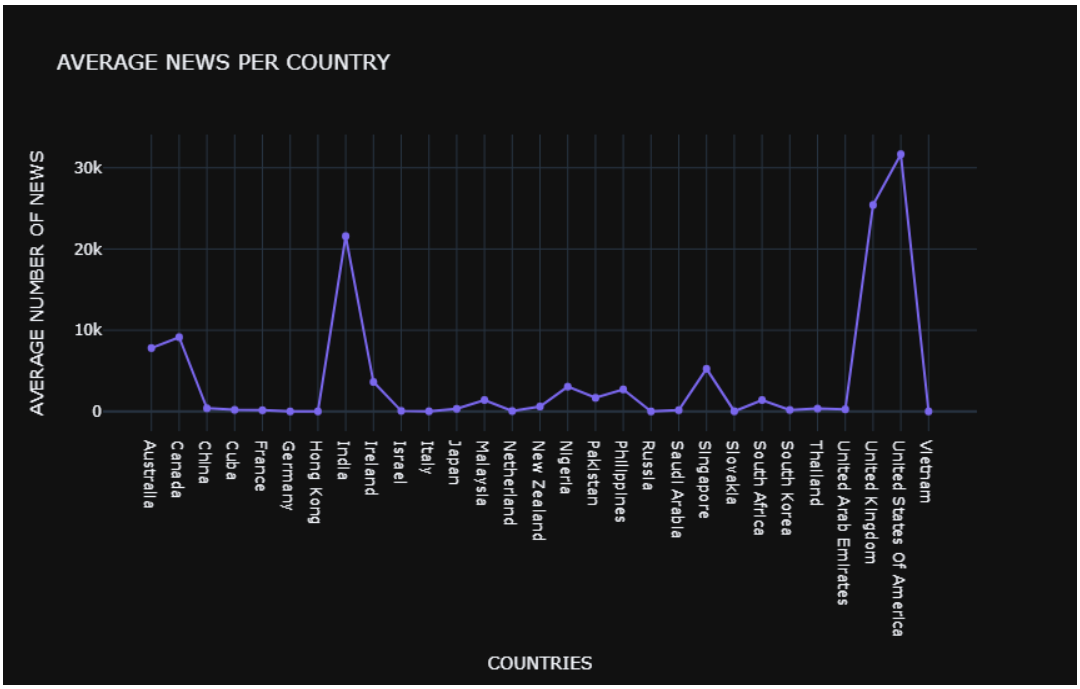

Figure 4.2(b): News average number of each category

Figure 4.2(b) displays the median amount of news across seven categories culled from worldwide news sources. Australia, Canada, Chile, Cuba, India, Ireland, Israel, France, Germany, Hong Kong, and the

United Kingdom are among them. Italy, New Zealand, Nigeria, Pakistan, the Philippines, Russia, Saudi Arabia, Singapore, Slovakia, South Africa, South Korea, Thailand, Japan, Malaysia, the Netherlands, the UAE, the UK, the United States of America, and Vietnam. Compared to Australia's average (about 8000), Canadians are exposed to significantly more media. According to reports, there has been a decrease in the total number of recorded news stories from 0 to 1000 in the countries of Chile, Cuba, France, Germany, and Hong Kong. In the years following, the number of stories in India rose from the hundreds to well over 20000. There are barely over 7500 stories about Ireland, which is a significant drop. The range of reported events is once again between zero to one thousand for the next three countries. There can be zero to five thousand news items for countries including Malaysia, the Netherlands, New Zealand, Nigeria, Pakistan, the Philippines, and Russia. In Singapore, where the average number of news items is now over 5000, there appears to have been a considerable increase in news production. Slovakia, South Africa, South Korea, Thailand, and the United Arab Emirates are among the countries where the numbers range from 0 to 5000. While the number of news stories in the United Kingdom stayed at 26,000, over 30k were published in the United States. There are now less than a thousand stories about Vietnam in the news.

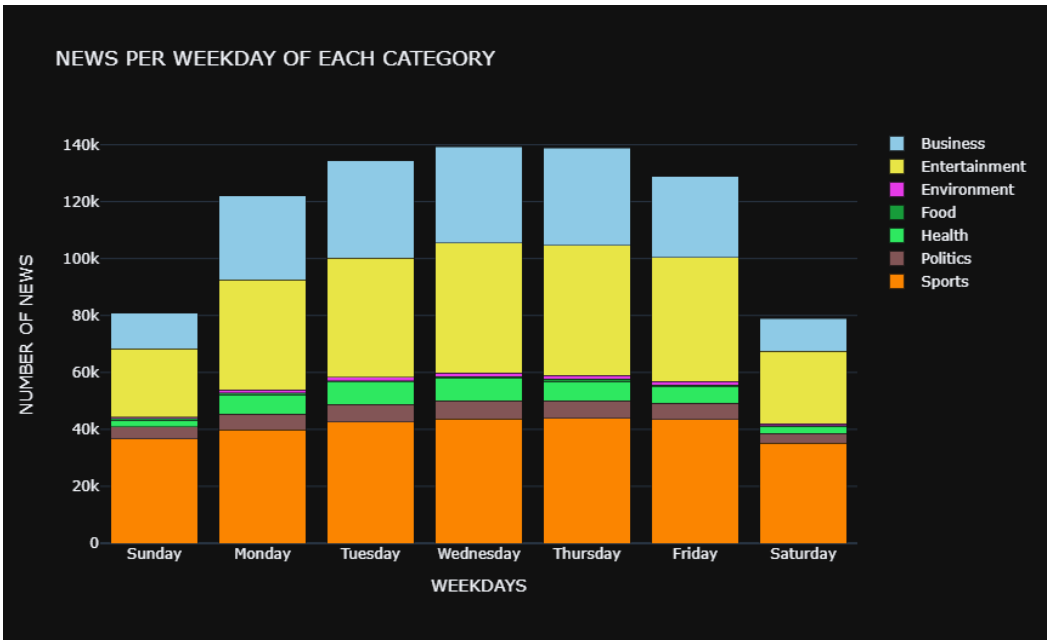

Figure 4.2(c): News of each category per weekday

Figure 4.2(c) displays seven categories of weekly news events. As indicated in Figure 4.2(c), the amount of sports news on Sunday is approximately 36000, but the amount of political news on Sunday is about 4000. On Sunday, the number of health-related articles was around 2500, whereas the number of environment-related articles was about 500. The Sunday food news is roughly 500. Sunday's entertainment news quantity is approximately 24k, while the business news amount is about 12k. The amount of sports news on Monday is about 40k, whereas the amount of political news is about 5500. The number of health-related articles is approximately 7000, while the number of environment-related articles on Monday is about 1200. Monday's food news is roughly 500. On Monday, the quantity of news in the entertainment category is about 39k, and the amount of news in the business category is approximately 30k. The amount of sports news on Tuesday is about 43k, whereas the amount of political news is around 6200. The number of health-related articles is around 8000, while the number of environment-related articles on Tuesday is about 1200. The food news for Tuesday is approximately 500. The amount of entertainment news on Tuesday is about 42k, while the amount of business news is approximately 34k. The amount of sports news on Wednesday is approximately 44k, whereas the amount of political news is approximately 6500. The number of health-related articles is around 8000,

while the number of environment-related articles on Wednesday is approximately 1300. Wednesday's food news runs roughly 500. The amount of entertainment news on Wednesday is approximately 45k, while the amount of business news is approximately 34k. The amount of sports news on Thursday is approximately 44k, whereas the amount of political news is about 6200. The number of health-related articles is approximately 7000, whereas the number of environment-related articles on Thursday is approximately 1300. Thursday's food news is approximately 700. The amount of entertainment news on Thursday is approximately 45k, while the amount of business news is approximately 34k. The amount of sports news on Friday is approximately 43k, whereas the amount of political news is approximately 5500. The number of health-related articles is approximately 6000, while the number of environment-related articles on Friday is approximately 1200. Friday's food news runs roughly 600. The quantity of news in the entertainment category on Friday is approximately 44k, while the amount of news in the business category is approximately 28k. The amount of sports news on Saturday is approximately 35000, whereas the amount of political news is approximately 3500. The number of health-related articles is approximately 2500, while the number of environment-related articles on Saturday is approximately 500. The food news on Saturday is approximately 300. The amount of entertainment news on Saturday is approximately 26k, whereas the amount of business news is approximately 11k.

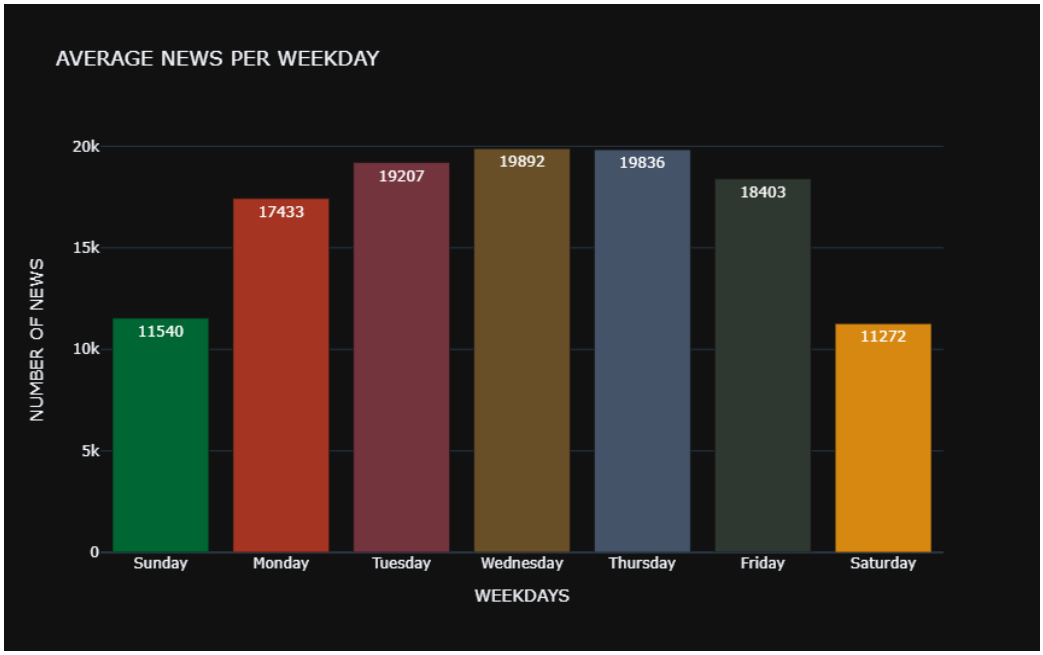

Figure 4.2(d): Average news per weekday

The figure 4.2(d) illustrates the average amount of news reported during the week. The information is broken up into seven different areas, which are as follows: business, politics, health, entertainment, and the environment. Figure 4.2(d) illustrates that there are 11540 pieces of news published on a typical Sunday. On Monday, there was a significant rise in the average number of news stories, which pushed it beyond the 15000. On Monday, there were 17433 news stories overall. The average number of news stories increased to 19207 on Tuesday, while on Wednesday and Thursday, they were 19892 and 19836, respectively. On Friday, there was a decline in the average quantity of news, which brought the average to 18403. On Saturday, the daily average number of news items fell even more to 11272.

The average number of new events that occur each month is displayed in Figure 4.2(e). As can be seen in figure 4.2(e), there were an average of 24527 pieces of news published during the month of May. The number dropped to 23007 in the month of June. The number of news items for the month of July rose by a modest amount, reaching 23303. The average number of news items continued to rise during the

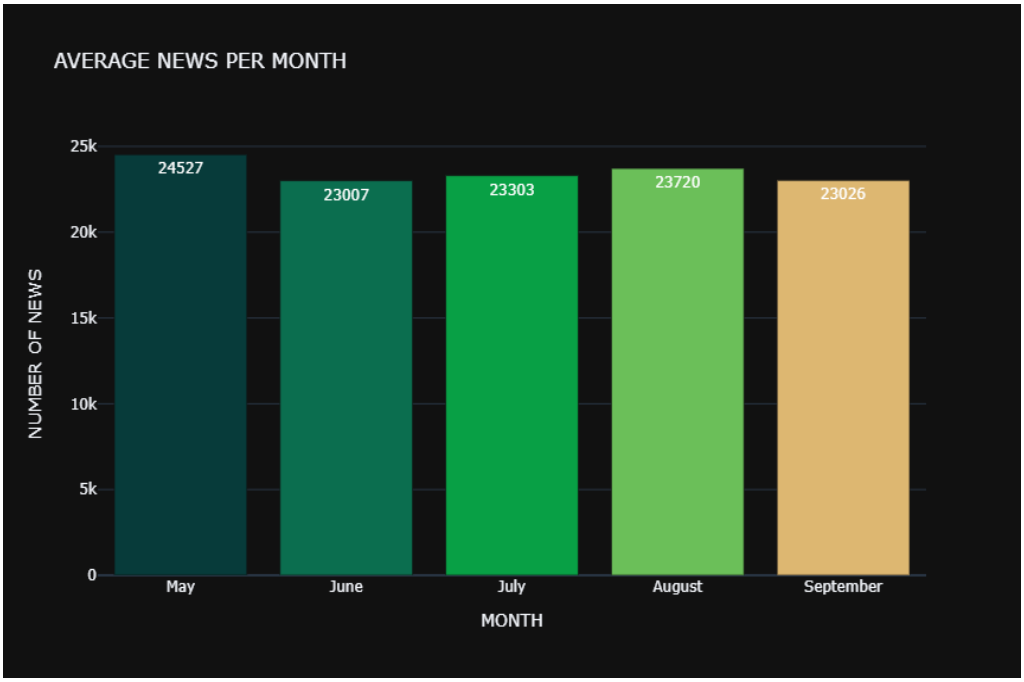

Figure 4.2(e): Average news per month

month of August, reaching 23720. The monthly average of news fell even more to 23026 during the month of September.

Figure 4.2(f) depicts the quantity of news per hour for each category of news. There are seven types of news, including business, entertainment, environment, food, health, and politics. The news range for the business news category is between 4k and 8k for 0 to 5 hours. Between five and ten hours, the news ranges between eight thousand and eleven thousand. The news stayed constant between 10 and 15 hours before beginning to decline between 15 and 20 hours. From 20 to 23 hours, the business news category decreased even further. Figure 4.2(f) depicts the quantity of news per hour for each category of news. There are seven types of news, including business, entertainment, environment, food, health, and politics. The news range for the business news category is between 4k and 8k for 0 to 5 hours. Between five and ten hours, the news ranges between eight thousand and eleven thousand. The news stayed constant between 10 and 15 hours before beginning to decline between 15 and 20 hours. From 20 to 23 hours, the business news category decreased even further. From 0 to 5 hours, the news for the entertainment sector runs between 8k and 20k items. After then, there is a significant surge in entertainment news, which reached around 12,000 at 10 hours. For the period between ten and fifteen hours, the information grew. At 15 hours, the amount of news is around 14 k, and from 15 to 20 hours, it began to decrease quickly. At 20 hours, the news number is approximately 12,000. For the period between 20 and 23 hours, the news deteriorated considerably. The environment news for the time range between 0 to 5 hours is between 200 and 400. For the time between 5 to 10 hours, the news started increasing, and at 10 hours, the environment news is around 400. Further, for the time range between 10 to 15 hours, the environment news increased even more; after 15 hours, the amount of environment news crossed 600. Around 16 hours, the environmental news decreased sharply; at 20 hours, the number is around 200. For the time between 20 to 23 hours, the amount of environmental information again started increasing slowly. After environment news category there is a food category. Between zero and five hours, the food news ranges between 100 and 200. In addition, the food news stayed between 100 and 200 for 5 to 15 hours. Then, between 15 and 20 hours, the food news abruptly surged to around 400, and between 20 and 23 hours, it massively reduced. From 0 to 5 hours, the health news ranges between 1000 and 1500. From five to ten hours, the number of health-related news stories grew to between 1,500 and 2000. Between the hours of 10 and 15, the number of health-related articles climbed

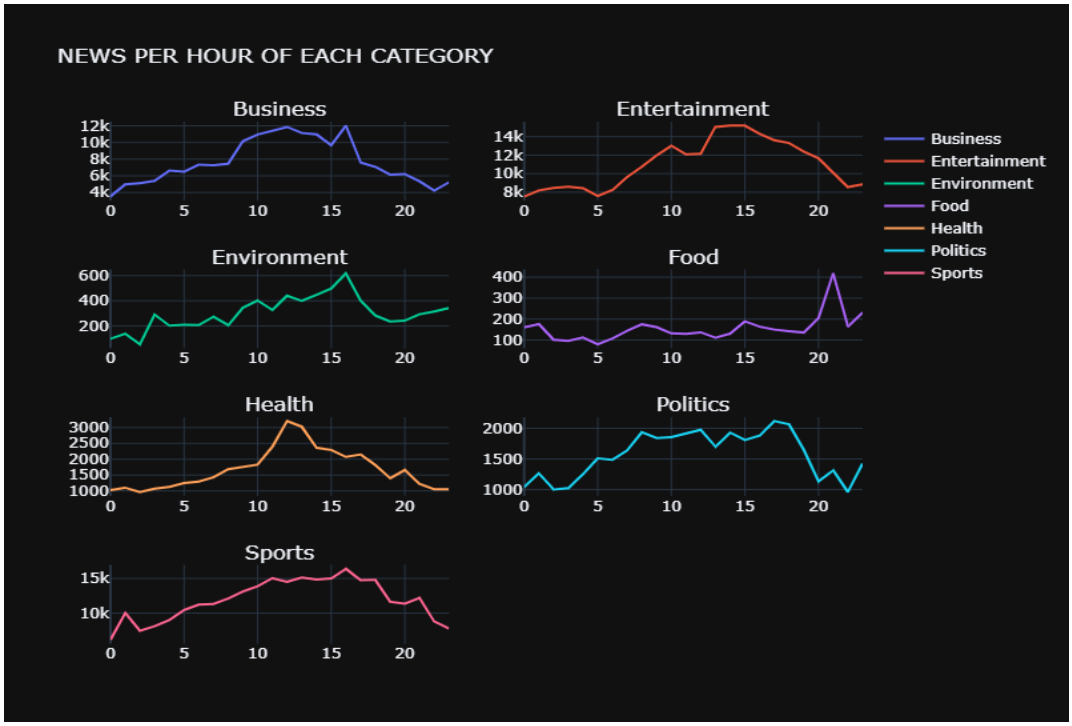

Figure 4.2(f): News of each category per hour

till 3000. Then, over the remaining hours, the health news began to diminish. The political news during the period between 05:00 and 15:00 is between 1000 and 1500. The news then grew between 5 and 10 hours, and between 1500 and 2000, politics news is reported. Between 10 and 15 hours, the political news remains steady; between 15 and 20 hours, the news declines; and between 20 and 23 hours, the news declines further. During the initial hours, between zero and five hours, the sports news reached around 10,000. Over the course of five to ten hours, the sports news maintained an increasing trend. After 15 hours, the sports news began to decline, and between the hours of 20 and 23, the decline accelerated.

Figure 4.2(g) depicts the news of each category occurred per month. As shown in the figure 4.2(g). In May, the sports news category has 59,94k articles, while the politics news category has 8,197 articles. May's health and food news totals are 8064 and 767, respectively. The number of environment news articles for the month of May is 1388, whereas the number of entertainment news articles is 54.83k and the number of business news articles is 38.50k. In June, the sports news category has 53.64k articles, while the politics news category has 7653 articles. June's health and food news totals are 8307 and 784, respectively. The number of environment news articles for the month of June is 1392, whereas the number of entertainment news articles is 53.27k and the number of business news articles is 35.99k. The sports news category contains 57.73k articles in July, whereas the politics news category has just 7,737 publications. In July, there were 8542 health news articles and 685 food news articles. The number of environment-related news stories for the month of July is 1596, whereas the number of entertainment-related news pieces is 51.34k and the number of business-related news articles is 35.49k. In August, the sports news category had 57.60k articles, whilst the politics news category had just 7508 publications. In August, there were 8500 health news stories and 727 food news pieces. For the month of August, there were 1541 environmental news items, 52.12k entertainment news pieces, and 38.04k business news articles. In September, the sports news category had 56.18k articles, whilst the politics news category had just 6636 publications. In September, there were 6797 health news stories and 793 food news pieces. For the month of September, there were 1364 environmental news items, 53.24k entertainment news pieces, and 36.15k business news articles.

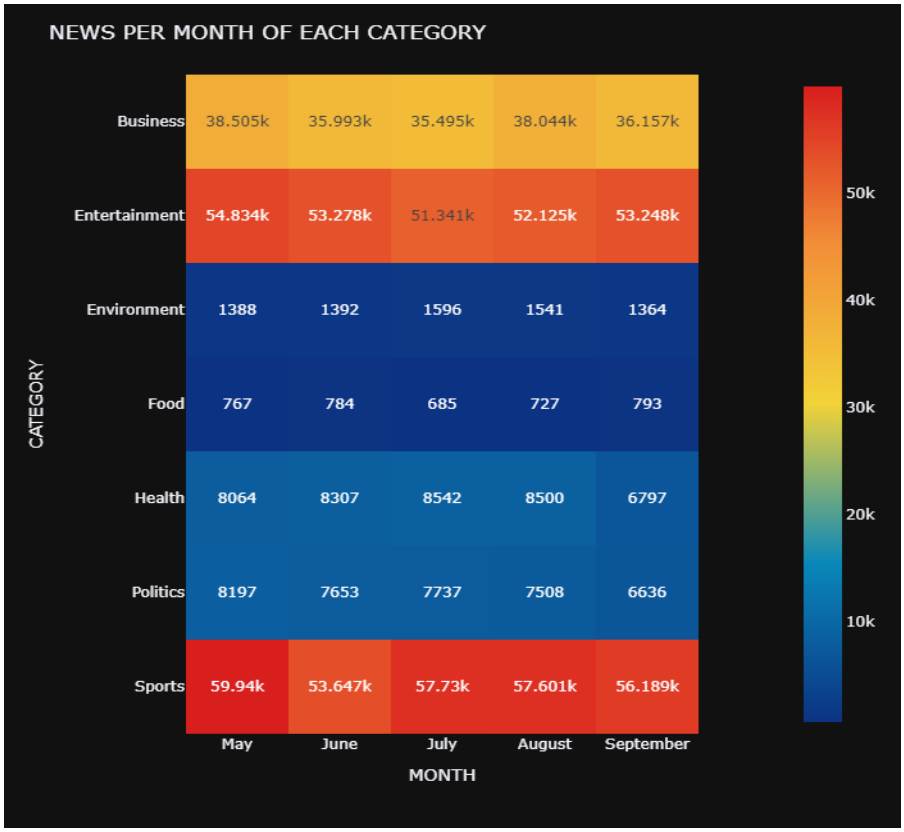

Figure 4.2(g): News of each category per month

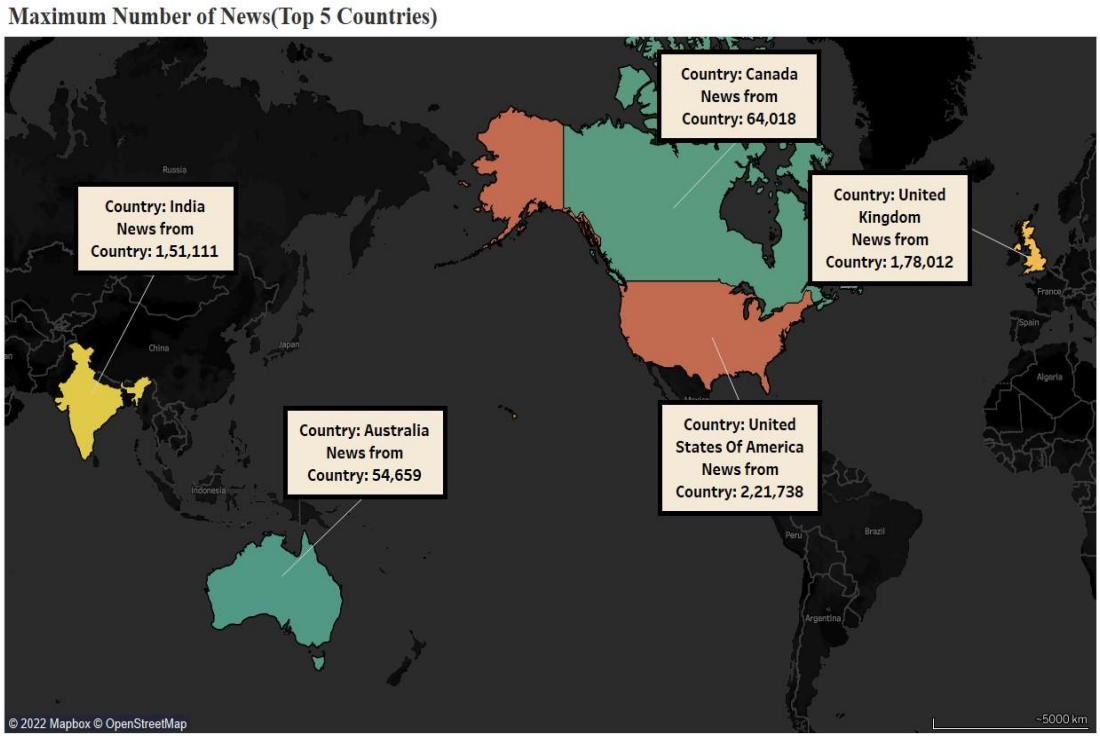

Figure 4.2(h): Maximum amount of news from the top 5 countries

Figure 4.2(h) shows the map of the countries producing the top news from May 2022 to September 2022 for the categories of business, food, health, environment, entertainment, politics and sports. As seen in figure 4.2(h), the United States of America has produced the highest amount of news which is

2,21,738. The USA is followed by the United Kingdom, which made 1,78,012 news. Then India and Canada produced 1,51,111 and 64,018 information, respectively. The fifth country is Australia which grew 54,659 news.

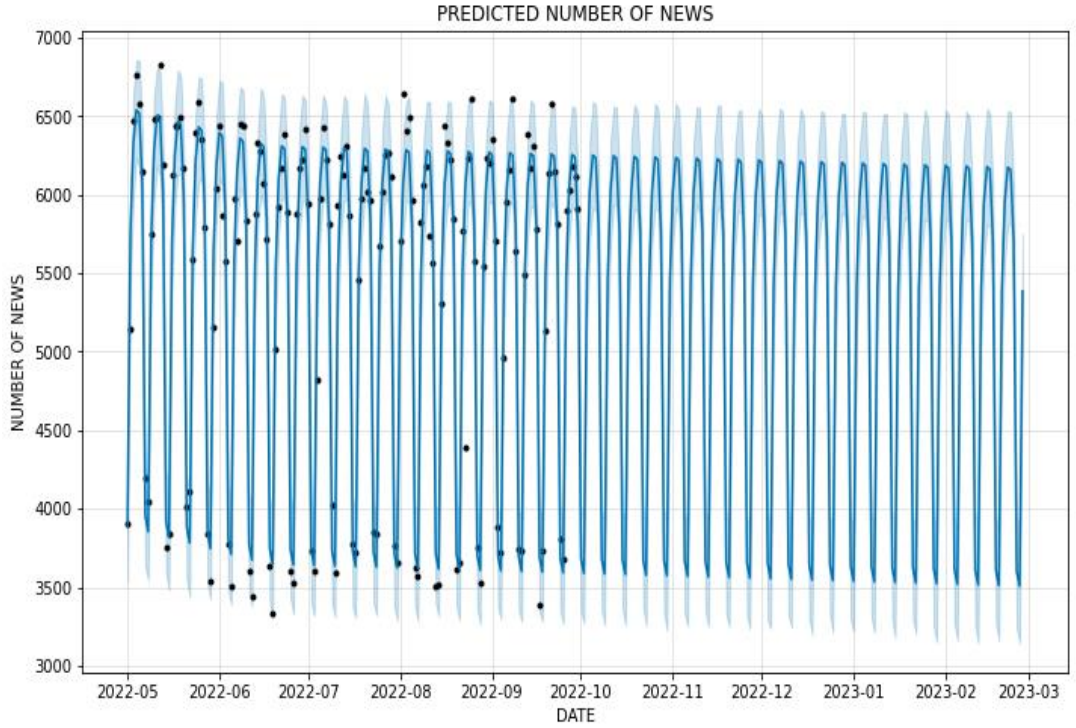

Figure 4.2(i): News prediction using prophet model

The date column is applied as the input and the news per day is taken as output for the prophet model. After that the prophet model is fitted on this data. The period for prediction is considered as the 150 days. The predicted news is shown in figure 4.2(i). As seen in figure 4.2(i) the predicted amount of news is from October 2022 till February 2023. The forecasted formed the sideways pattern which following the pattern for the news from May2022 till September 2022. The amount of news for each day is ranging between approximately between 2500 to around 6800. The predicted news which is from October 2022 till February is also seems to be in the range 2500 to around 6800.

4.3 Sentiment analysis model performance

4.3.1 Naïve bayes model performance (multinomial)

Parameters	Baseline Model	Tuned model
Accuracy	0.68	0.70
Precision Score	0.70	0.68
Recall Score	0.56	0.63
F1 Score	0.59	0.64
Fbeta Score	0.61	0.65
MAE	0.44	0.43
MSE	0.70	0.70
RMSE	0.83	0.84
Training Time(s)	0.37	0.97
ROC AUC Score	0.87	0.85

Table 4.3.1(I): Naïve bayes model (multinomial) performance as per the parameters

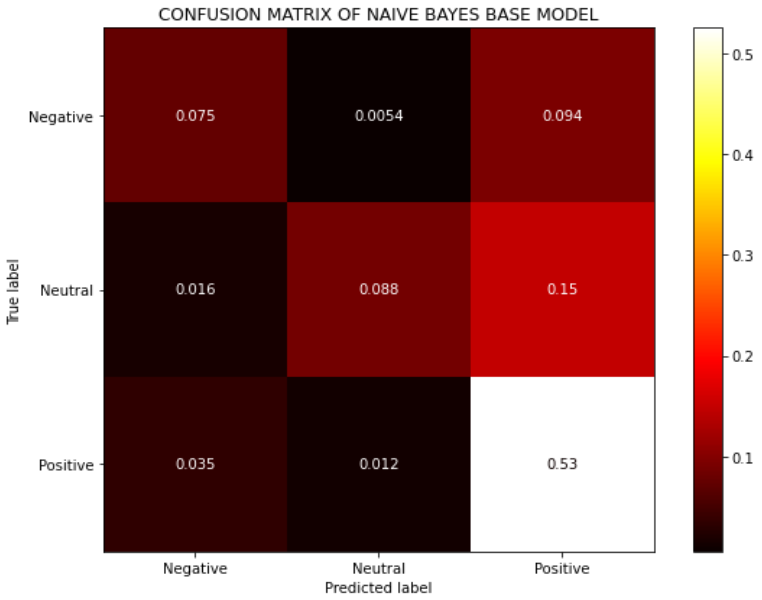

Figure 4.3.1(a): Confusion matrix of the naïve bayes base line model (multinomial)

As depicted in Figure 4.3.1(a), the naïve bayes base line model (Multinomial) confusion matrix consists of three actual and anticipated classes that are negative, neutral, and positive. The true positive value of the negative class in this instance is 0.075. True positive value of class neutral is 0.088, whereas true positive value of class positive is 0.53. The naive bayes baseline model (Multinomial) accurately predicted 0.075 for the negative class, 0.088 for the neutral class, and 0.53 for the positive class.

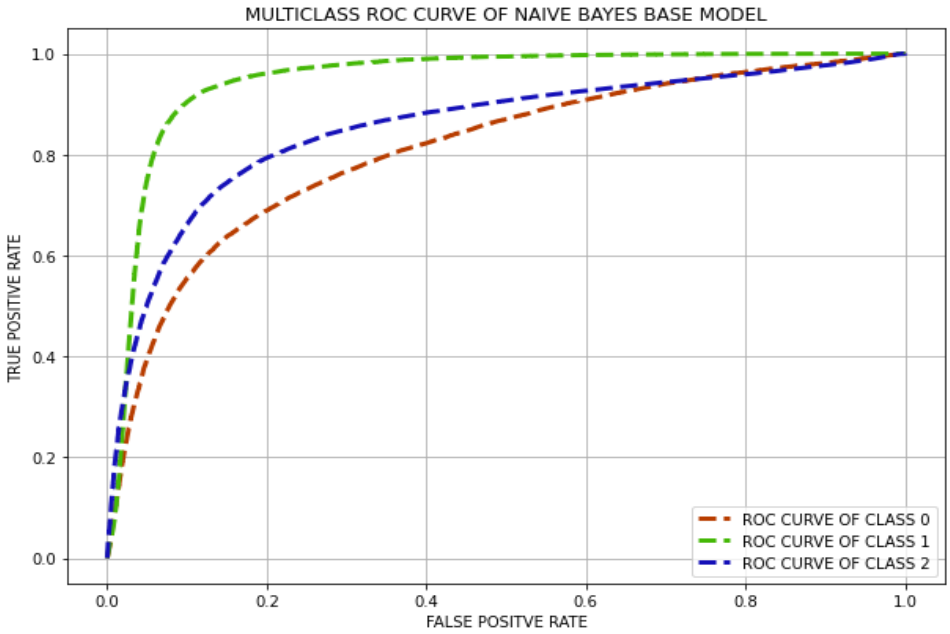

Figure 4.3.1(b): ROC curve for naïve bayes base line model (Multinomial)

The receiver operating characteristic curve for the naive bayes baseline model (Multinomial) is depicted in Figure 4.3.1(b). Here, the ROC curve for the three classes negative, neutral, and positive is presented. For the naive bayes baseline model (Multinomial), the ROC AUC value is 0.87. The curve for the negative class lies between 0.8 and 1, but the curves for the neutral and positive classes are close to 0.8 and 0.65, respectively.

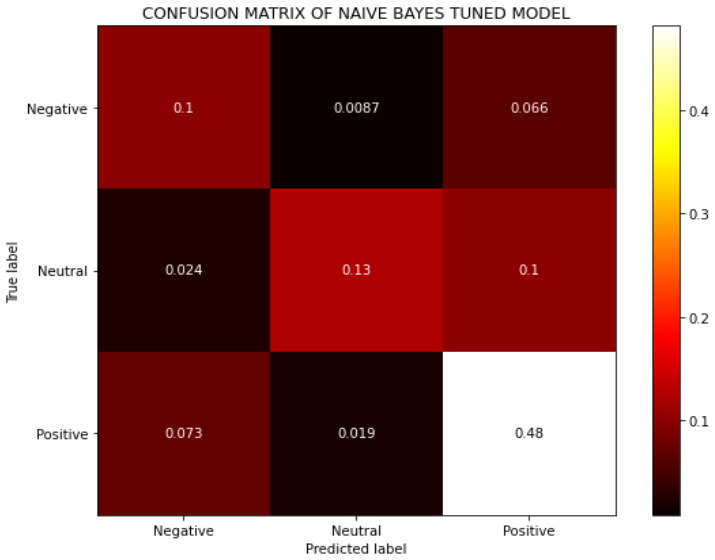

Figure 4.3.1(c): Confusion matrix of the naïve bayes tuned model (Multinomial)

After applying the hyperparameters which are alpha = 0.1, class_prior =None and the fit_prior = True the naïve bayes tuned model (multinomial) was created and its confusion matrix in figure 4.3.1(c). It comprises of three negative, neutral, and positive real and predicted classes. In this instance, the true positive value of the negative class is 0.1. Class neutral has a true positive value of 0.13, while class positive has a true positive value of 0.48. The naïve bayes tuned model properly predicted 0.1 for the negative category, 0.13 for the neutral category, and 0.48 for the positive category.

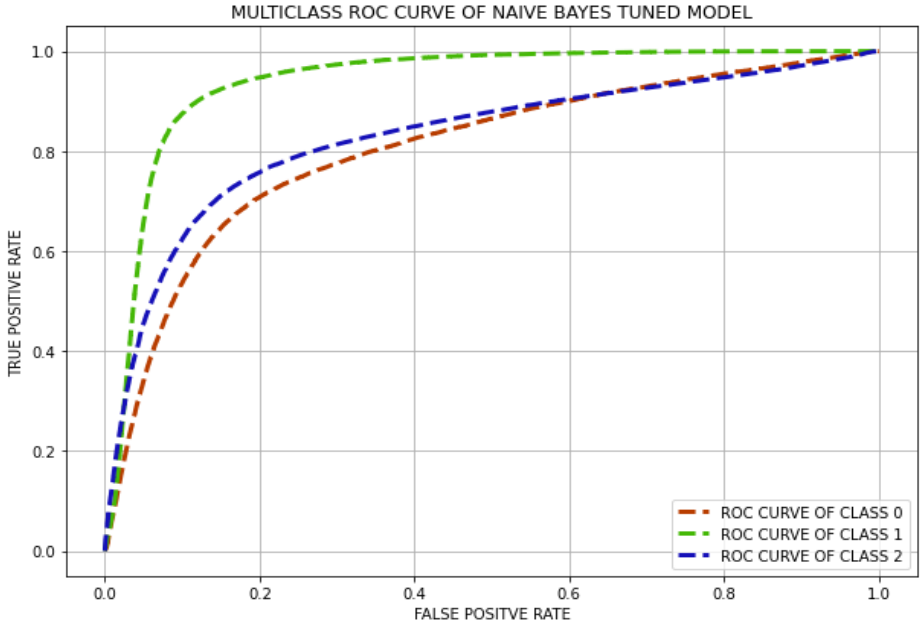

Figure 4.3.1(d): ROC curve for naïve bayes tuned model (Multinomial)

The receiver operating characteristic curve for the naïve bayes tuned model (Multinomial) is depicted in Figure 4.3.1(d). An ROC curve is shown here for the three possible outcomes (negative, neutral, and positive). The area under the ROC curve (AUC) is 0.85 for the naive bayes tuned model. The negative class has a curve between 0.8 and 1, whereas the neutral and positive classes have curves around 0.65 and 0.60, respectively.

4.3.2 Random forest model performance

Parameters	Baseline Model	Tuned model
Accuracy	0.86	0.85
Precision Score	0.90	0.87
Recall Score	0.77	0.77
F1 Score	0.80	0.79
Fbeta Score	0.82	0.81
MAE	0.22	0.22
MSE	0.39	0.40
RMSE	0.63	0.63
Training Time(s)	43900.46	8830.47
ROC AUC Score	0.97	0.94

Table 4.3.2(I): Random forest model performance as per the parameters

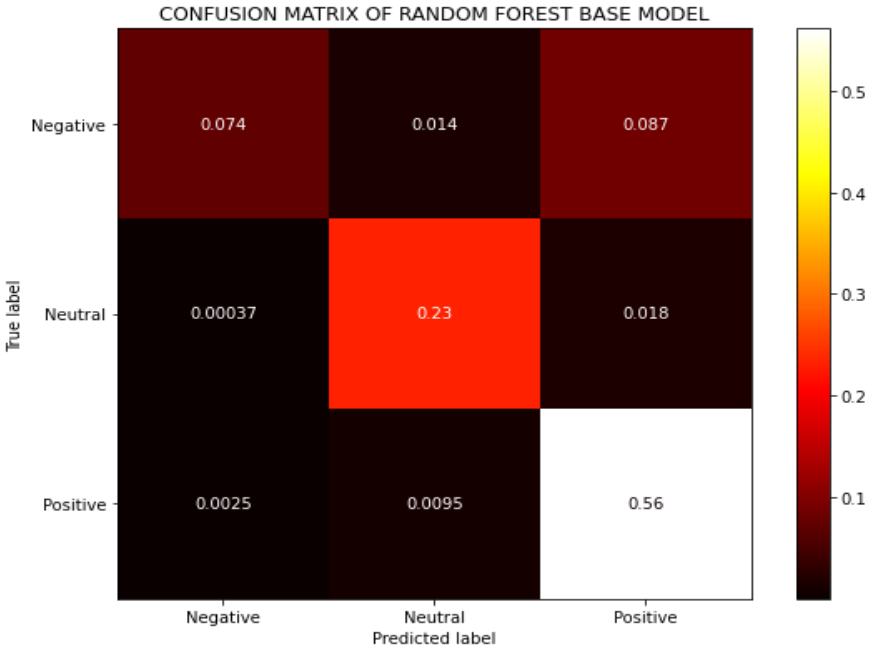

Figure 4.3.2(a): Confusion matrix of the random forest base line model

The confusion matrix for the random forest baseline model is shown in Figure 4.3.2(a), and it consists of the three classes (negative, neutral, and positive) that are actually present and that were expected. In this case, the true positive value of the negative category is 0.074. For class neutral, the genuine positive value is 0.23, while for class positive it is 0.56. The random forest baseline model predicted a value of 0.074 for the negative class, 0.23 for the neutral class, and 0.56 for the positive class.

Figure 4.3.2(b) shows the random forest baseline model's receiver operating characteristic curve. Here, the ROC curves for negative, neutral, and positive classes are shown. The ROC AUC value for the base model of a random forest is 0.97. The curve for the negative class is between 0.8 and 1, but the curves for the neutral and positive classes are also between 0.8 and 1.

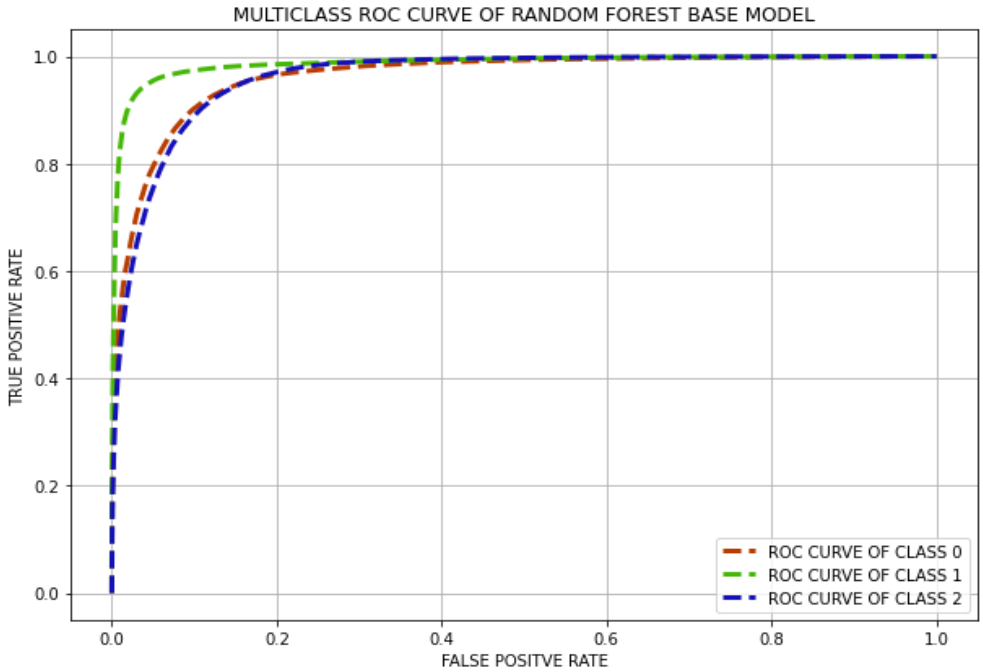

Figure 4.3.2(b): ROC curve for random forest base line model

.

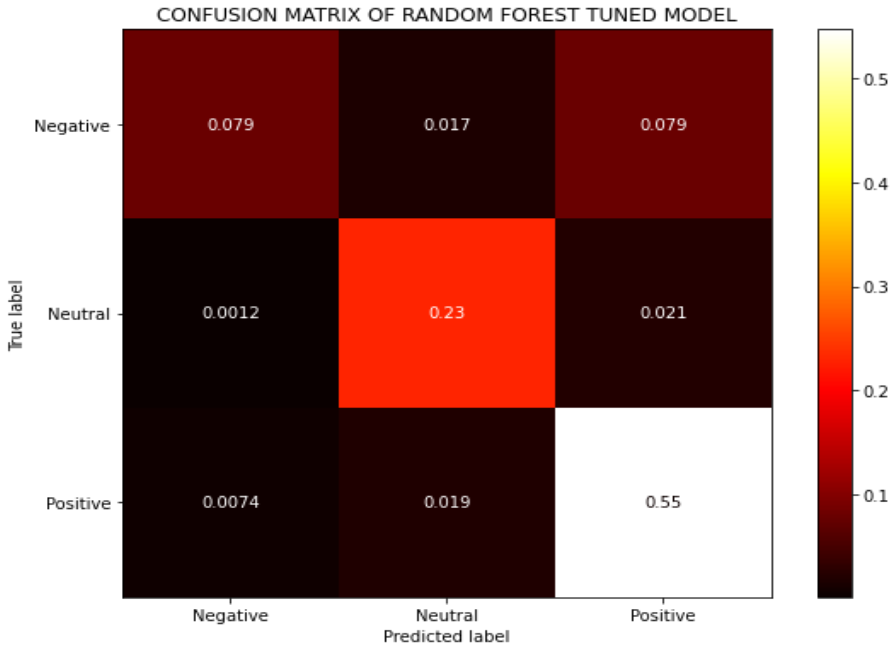

Figure 4.3.2(c): Confusion matrix of the random forest tuned model

After applying the hyperparameters which are n_estimators and max_features the random forest tuned model was created and its confusion matrix in figure 4.3.2(c). It comprises of three negative, neutral, and positive real and predicted classes. In this instance, the true positive value of the negative class is 0.079. Class neutral has a true positive value of 0.23, while class positive has a true positive value of 0.55. The random forest tuned model properly predicted 0.079 for the negative category, 0.23 for the neutral category, and 0.55 for the positive category.

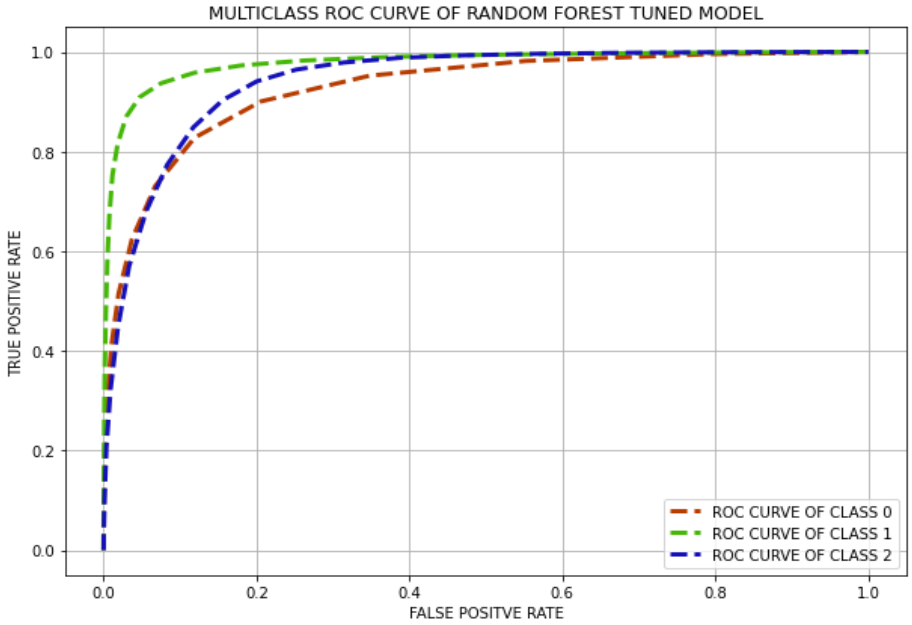

Figure 4.3.2(d): ROC curve for random forest tuned model

The receiver operating characteristic curve for the random forest tuned model is depicted in Figure 4.3.2(d). Here, the ROC curves for negative, neutral, and positive classes are shown. The ROC AUC value for the random forest tuned model is 0.94. The curve for the negative class is between 0.8 and 1, while the curves for the neutral and positive classes are close to 0.85 and 0.80, respectively.

4.3.3 Stochastic gradient descent model performance

Parameters	Baseline Model	Tuned model
Accuracy	0.91	0.93
Precision Score	0.90	0.91
Recall Score	0.87	0.91
F1 Score	0.88	0.91
Fbeta Score	0.88	0.91
MAE	0.12	0.09
MSE	0.20	0.15
RMSE	0.45	0.39
Training Time(s)	9.38	44.24
ROC AUC Score	0.97	0.96

Table 4.3.3(I): Stochastic gradient descent model performance as per the parameters

The confusion matrix for the stochastic gradient descent baseline model is shown in Figure 4.3.3(a), and it consists of the three classes (negative, neutral, and positive) that are actually present and that were expected. In this case, the true positive value of the negative category is 0.12. For class neutral, the genuine positive value is 0.25, while for class positive it is 0.54. The stochastic gradient descent baseline model predicted a value of 0.12 for the negative class, 0.25 for the neutral class, and 0.54 for the positive class.

Figure 4.3.3(b) shows the stochastic gradient descent baseline model's receiver operating characteristic curve. Here, the ROC curves for negative, neutral, and positive classes are shown. The ROC AUC value for the base model of a stochastic gradient descent baseline model is 0.97. The curve for the negative

class is between 0.8 and 1 and the curves for the neutral and positive classes are also between 0.8 and 1.

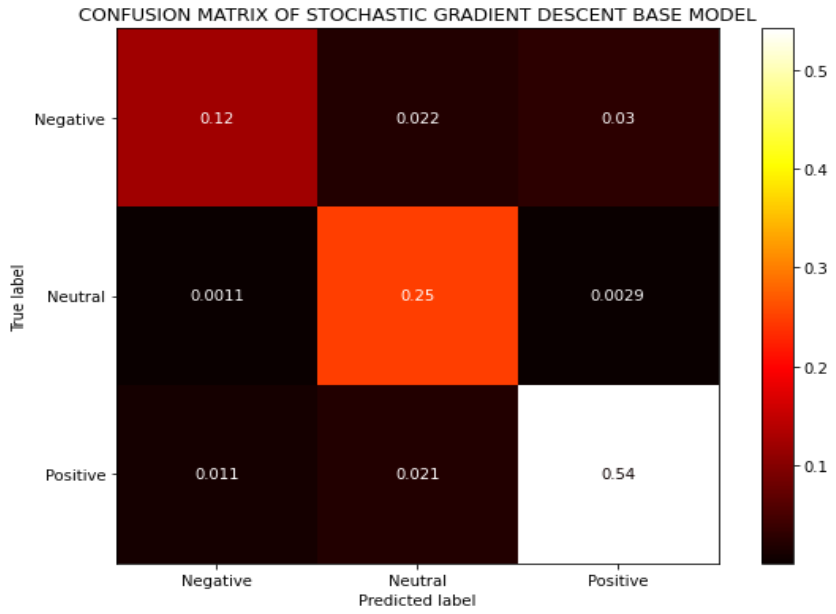

Figure 4.3.3(a): Confusion matrix of the stochastic gradient descent base line model

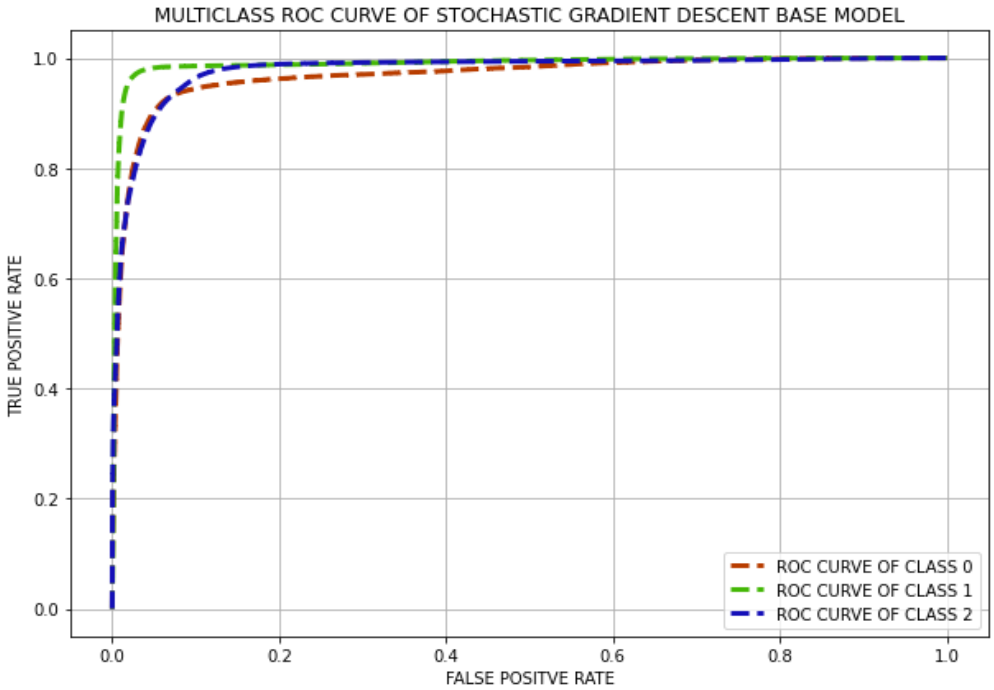

Figure 4.3.3(b): ROC Curve of the stochastic gradient descent base line model

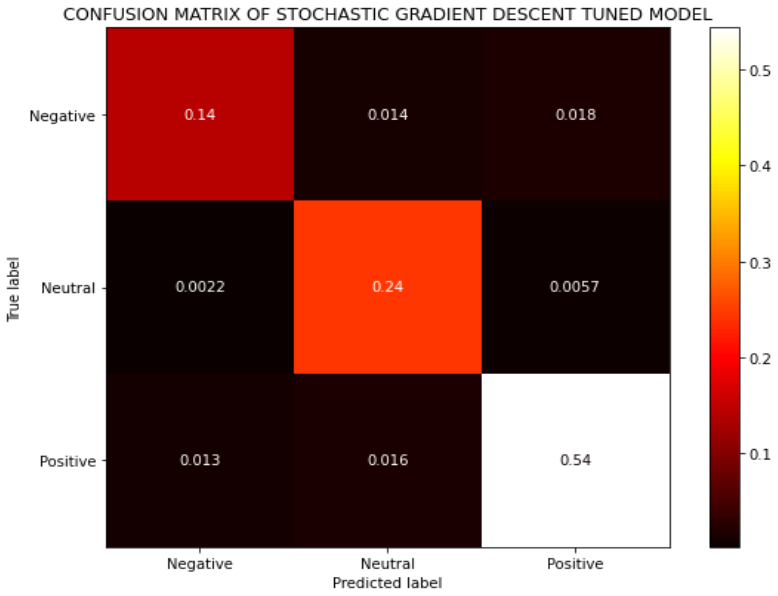

Figure 4.3.3(c): Confusion matrix of the stochastic gradient descent tuned model

For stochastic gradient tuning the hyperparameters such as penalty and alpha applied and its confusion matrix is shown in figure 4.3.3(c). It comprises of three negative, neutral, and positive real and predicted classes. In this instance, the true positive value of the negative class is 0.14. Class neutral has a true positive value of 0.24, while class positive has an actual positive value of 0.54. The stochastic gradient descent tuned model properly predicted 0.14 for the negative class, 0.24 for the neutral category, and 0.54 for the positive class.

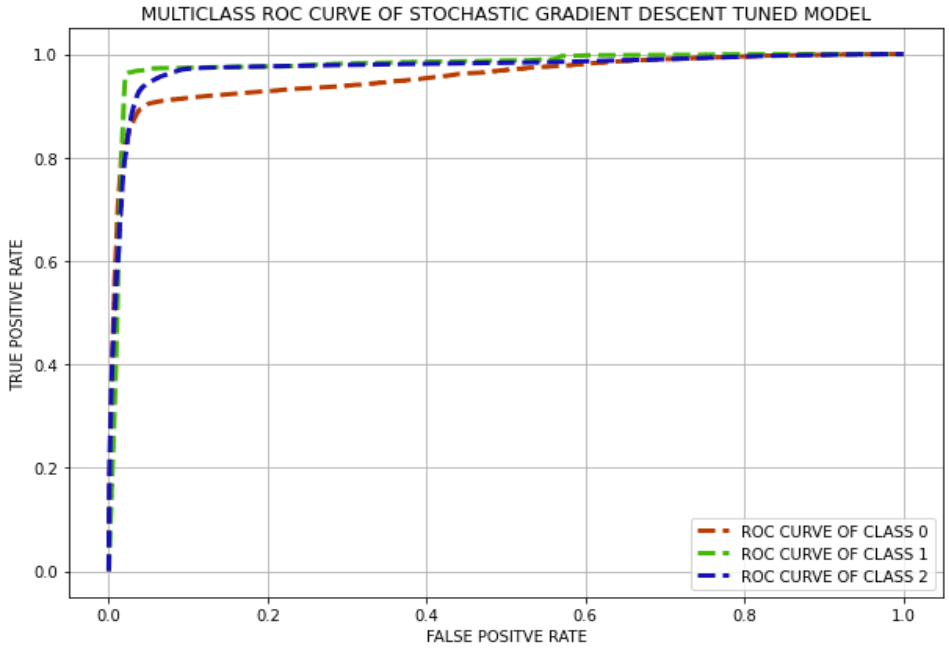

Figure 4.3.3(d): ROC Curve of the stochastic gradient descent tuned model

The receiver operating characteristic curve for the stochastic gradient descent tuned model is depicted in Figure 4.3.3(d). Here, the ROC curves for negative, neutral, and positive classes are shown. The ROC AUC value for the stochastic gradient descent tuned model is 0.96. The curve for the negative class is between 0.8 and 1, while the curves for the neutral and positive classes are also between 0.8 and 1.

4.3.4 XGBoost model performance

Parameters	Baseline Model	Tuned model
Accuracy	0.74	0.74
Precision Score	0.74	0.74
Recall Score	0.67	0.67
F1 Score	0.64	0.64
Fbeta Score	0.65	0.64
MAE	0.32	0.32
MSE	0.45	0.45
RMSE	0.67	0.67
Training Time(s)	550.65	524.17
ROC AUC Score	0.86	0.86

Table 4.3.4(I): XGBoost model performance as per the parameters

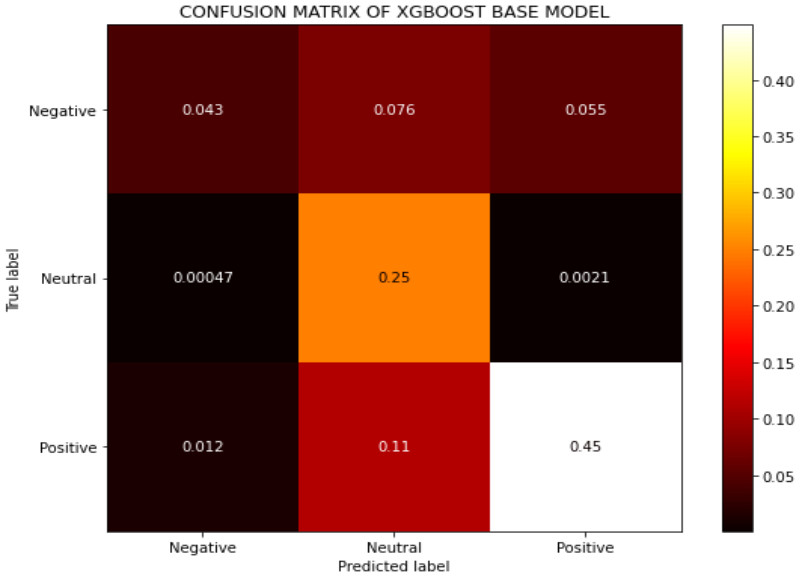

Figure 4.3.4(a): Confusion matrix of the xgboost base line model

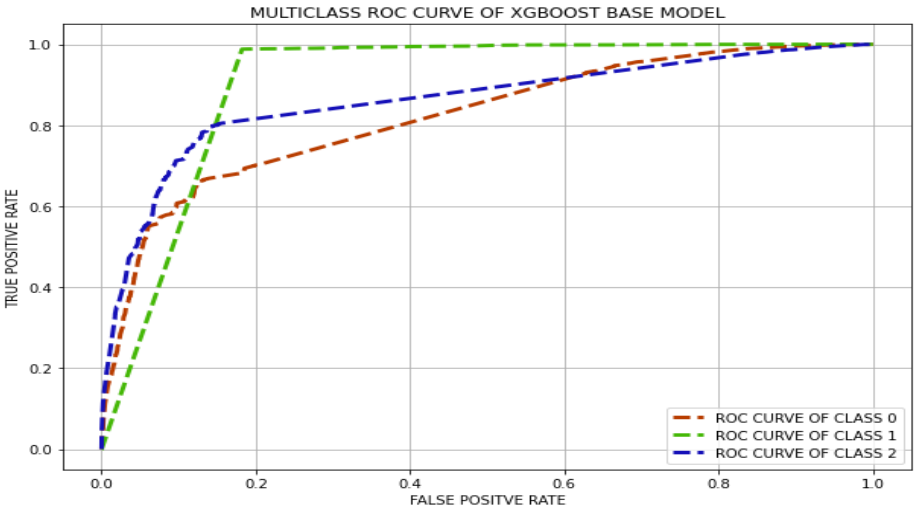

Figure 4.3.4(b): ROC Curve of the xgboost base line model

The confusion matrix for the xgboost baseline model is shown in Figure 4.3.4(a), and it consists of the three classes (negative, neutral, and positive) that are actually present and that were expected. In this case, the true positive value of the negative category is 0.043. For class neutral, the true positive value is 0.25, while for class positive it is 0.45. The xgboost baseline model predicted a value of 0.043 for the negative class, 0.25 for the neutral class, and 0.45 for the positive class.

Figure 4.3.4(b) shows the xgboost baseline model's receiver operating characteristic curve. Here, the ROC curves for negative, neutral, and positive classes are shown. The ROC AUC value for the base model of a xgboost baseline model is 0.86. The curve for the negative class is around 0.6 and the curves for the neutral and positive classes are around 1 and 0.8, respectively.

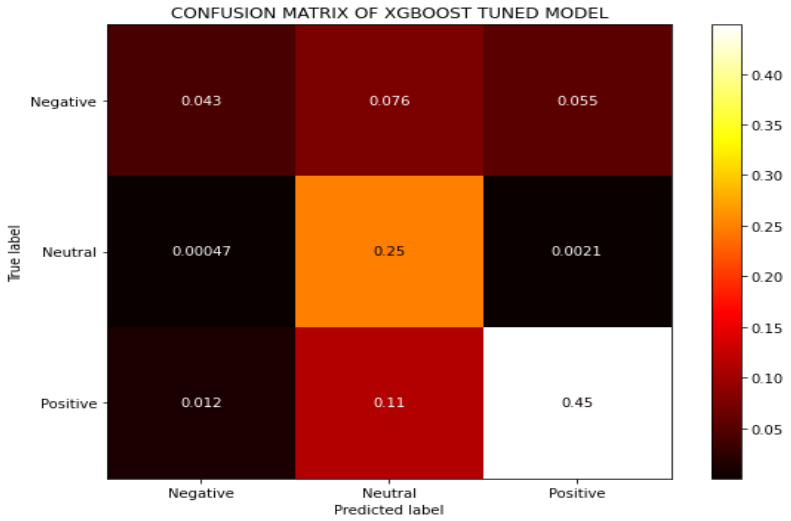

Figure 4.3.4(c): Confusion matrix of the xgboost tuned model

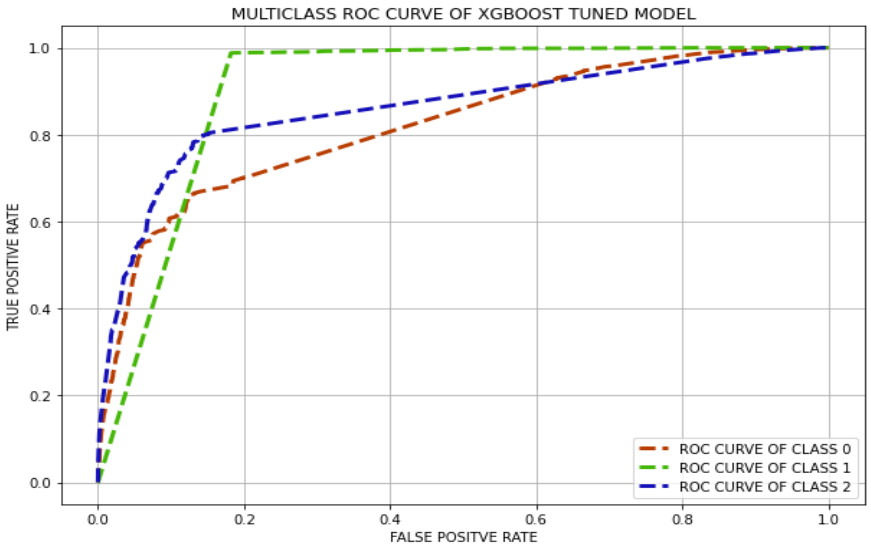

Figure 4.3.4(d): ROC Curve of the xgboost tuned model

The confusion matrix for the xgboost tuned model is shown in Figure 4.3.4(c), and it consists of the three classes (negative, neutral, and positive) that are actually present and that were expected. In this case, the negative category's true positive value is 0.043. The true positive value for class neutral is0.25,

while it is 0.45 for class positive. The xgboost tuned model predicted that the negative class has a value of 0.043, the neutral class would have a value of 0.25, and the positive class would have value of 0.45. For creating the xgboost tuned model the hyperparameters such as gamma and max_depth is used.

Figure 4.3.4(d) shows the xgboost tuned model's receiver operating characteristic curve. Here, the ROC curves for negative, neutral, and positive classes are shown. The ROC AUC value for the base model of a xgboost tuned model is 0.86. The curve for the negative class is around 0.6 and the curves for the neutral and positive classes are around 1 and 0.8, respectively.

4.4 News classification analysis model performance

4.4.1 CNN model performance

```
Model: "sequential"
_________________________________________________________________
 Layer (type)                Output Shape              Param #
=================================================================
 embedding (Embedding)       (None, 3800, 128)         1920128

 conv1d (Conv1D)             (None, 3798, 64)          24640

 dropout (Dropout)           (None, 3798, 64)          0

 max_pooling1d (MaxPooling1D  (None, 1899, 64)         0
 )

 conv1d_1 (Conv1D)           (None, 1897, 64)          12352

 dropout_1 (Dropout)         (None, 1897, 64)          0

 max_pooling1d_1 (MaxPooling  (None, 948, 64)          0
 1D)

 conv1d_2 (Conv1D)           (None, 946, 64)           12352

 global_max_pooling1d (Globa  (None, 64)               0
 lMaxPooling1D)

 dense (Dense)               (None, 64)                4160

 dense_1 (Dense)             (None, 7)                 455

=================================================================
Total params: 1,974,087
Trainable params: 1,974,087
Non-trainable params: 0
```

Figure 4.4.1(a): Trainable parameters of cnn baseline model

Figure 4.4.1(a) shows the training parameters that can be used. As shown, the model gets its information from the parameters. Then, these parameters went through the first layer of convolution. After the first layer's parameters were convolved, they were sent to the first dropout layer and the first max pooling layer. The parameters then went through the second convolution layer, the second dropout layer, and the second maximum pooling layer. The parameters then went through the third convolution layer and the global maximum pooling layer. In the end, the parameters went through a couple of dense layers, and the total number of parameters that could be used was 1,974,087.

Figure 4.4.1(b) shows the information about the training loss of the cnn baseline model. The cnn baseline model was trained for five iterations, as shown. After the first epoch is applied, the loss went to a value between 0.27 and 0.26. CNN's baseline model has a loss of about 0.24 in the second epoch. The decline in the third epoch ranges from 0.22 to 0.23. Approximately 0.21 is lost in the fourth epoch, and roughly 0.2 is lost in the final epoch. It can be seen that the training loss for the cnn baseline model is reducing as more and more epochs are accumulated.

Figure 4.4.1(c) shows the information about the training accuracy of the cnn baseline model. As it can be seen the model was trained on five epochs. When the 1st epoch was applied the training accuracy of the cnn baseline model went to approximately 0.91. After applying the 2ndepoch the training accuracy of the cnn baseline model was between 0.92 to 0.925. When the 3rd epoch was applied the training accuracy of the cnn baseline model is between 0.925 to 0.93. After applying the 4thepoch the training

accuracy further increased and went close to 0.93 and when the last epoch was applied the training accuracy went close to 0.935. For training accuracy, when epochs are increasing the training accuracy also increases which is the exactly opposite pattern as compare to the training loss.

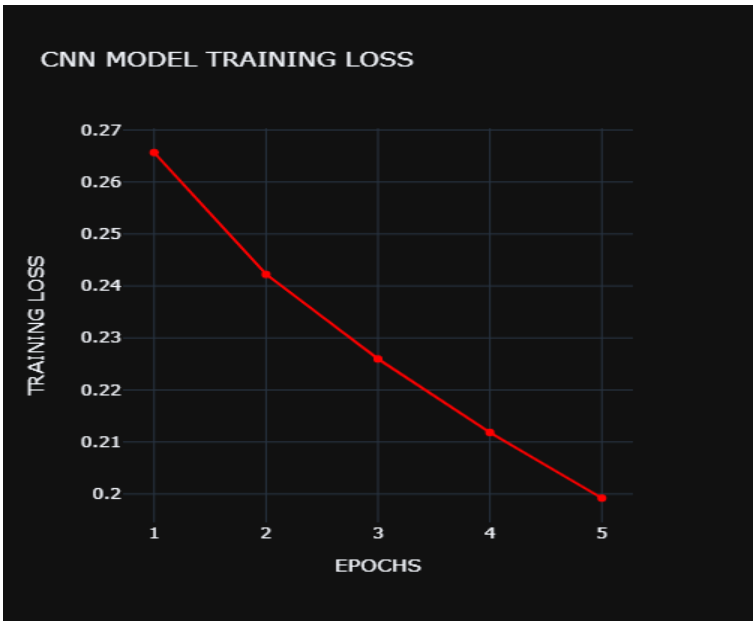

Figure 4.4.1(b): CNN baseline model training loss

Figure 4.4.1(c): CNN baseline model training accuracy

Parameters	Training Performance	Testing Performance
Accuracy	0.93	0.57
Precision Score	0.87	0.13
Recall Score	0.80	0.11
F1 Score	0.82	0.12
Fbeta Score	0.83	0.12
MSE	0.90	6.01
RMSE	0.95	2.45

Table 4.4.1(I): CNN baseline model performance as per the parameters

Parameters	Training Performance	Testing Performance
Loss	0.19	20.22
Accuracy	0.94	0.55

Table 4.4.1(II): CNN 1st tuned model performance as per the parameters (tuning type = hyperband, epochs=10, batch size = 512)

Parameters	Training Performance	Testing Performance
Loss	0.17	25.91
Accuracy	0.95	0.53

Table 4.4.1(III): CNN 2nd tuned model performance as per the parameters (tuning type = hyperband epochs=15, batch size = 512)

Parameters	Training Performance	Testing Performance
Loss	1.2	7.77
Accuracy	0.43	0

Table 4.4.1(IV): CNN 3rd tuned model performance as per the parameters (tuning type = random search epochs=10, batch size = 512)

Parameters	Training Performance	Testing Performance
Loss	0.35	27.48
Accuracy	0.90	0.59

Table 4.4.1(V): CNN 4th tuned model performance as per the parameters (tuning type = random search epochs=15, batch size = 512)

4.4.2 LSTM model performance

```
Model: "sequential"
_________________________________________________________________
 Layer (type)                 Output Shape              Param #
=================================================================
 embedding (Embedding)        (None, 3800, 128)         1920000

 dropout (Dropout)            (None, 3800, 128)         0

 cu_dnnlstm (CuDNNLSTM)       (None, 128)               132096

 dense (Dense)                (None, 64)                8256

 dropout_1 (Dropout)          (None, 64)                0

 dense_1 (Dense)              (None, 7)                 455

=================================================================
Total params: 2,060,807
Trainable params: 2,060,807
Non-trainable params: 0
_________________________________________________________________
```

Figure 4.4.2(a): Trainable parameters of lstm baseline model

The lstm model training parameters are shown in Figure 4.4.2(a). In the beginning, the parameter was used on the embedding layer. The parameters then went via a dropout layer. Next, the parameter is used on the cudnnlstm layer and the first dense layer. Next, the parameters are applied on the dense layer and the second dropout layer. There are a grand total of 2,060,807 trainable parameters at the end.

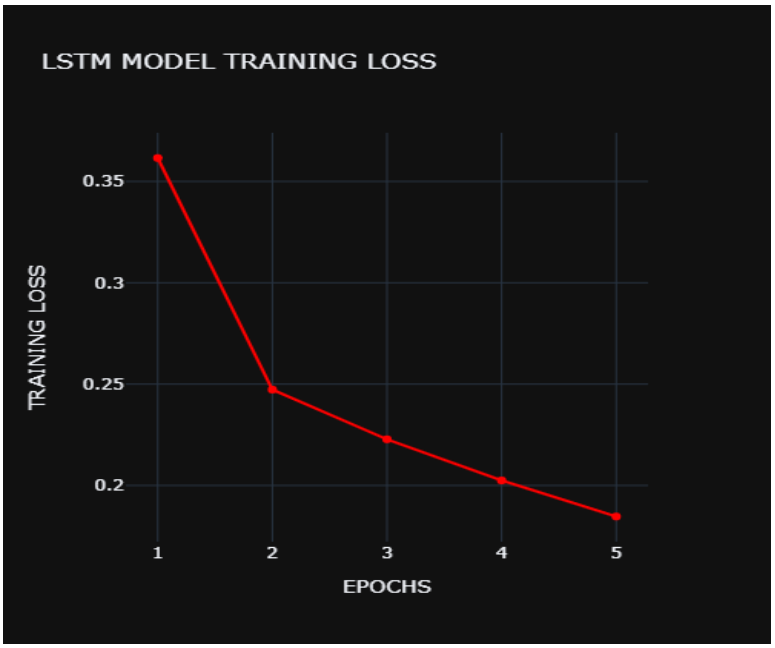

Figure 4.4.2(b): LSTM baseline model training loss

The lstm baseline model's training loss is displayed in Figure 4.4.2(b). As can be seen, the training loss of the lstm baseline model is around 0.35 after the first epoch has been applied. The lstm baseline model's training loss is nearly 0.25 when the second epoch is applied. The training loss is dropped below 0.25 when the third epoch is implemented. Loss dropped nearly to 0.2 once the fourth epoch was applied. Last but not least, the loss has dropped to less than 0.2 when the last epoch was applied.

Figure 4.4.2(c): LSTM baseline model training accuracy

The lstm baseline model's training accuracy data is displayed in Figure 4.4.2(c). The lstm baseline model's accuracy is close to 0.88 when the first epoch is applied, as can be seen here. Applying the second epoch resulted in a near 0.92 improvement in training accuracy. Third epoch training resulted in a training accuracy of 0.92 to 0.93, which is a significant improvement over previous iteration. The training accuracy reached 0.93 after the fourth epoch, and then exceeded that threshold after the fifth epoch.

Parameters	Training Performance	Testing Performance
Accuracy	0.94	0.58
Precision Score	0.87	0.13
Recall Score	0.83	0.11
F1 Score	0.85	0.12
Fbeta Score	0.85	0.12
MSE	0.93	5.97
RMSE	0.96	2.44

Table 4.4.2(I): LSTM baseline model performance as per the parameters

Parameters	Training Performance	Testing Performance
Loss	0.16	10.41
Accuracy	0.96	0.53

Table 4.4.2(II): LSTM 1st tuned model performance as per the parameters (tuning type = hyperband, epochs=10, batch size = 512)

Parameters	Training Performance	Testing Performance
Loss	0.18	14.01
Accuracy	0.97	0.54

Table 4.4.2(III): LSTM 2nd tuned model performance as per the parameters (tuning type = hyperband epochs=15, batch size = 512)

Parameters	Training Performance	Testing Performance
Loss	0.23	17.79
Accuracy	0.92	0.56

Table 4.4.2(IV): LSTM 3rd tuned model performance as per the parameters (tuning type = random search epochs=10, batch size = 512)

Parameters	Training Performance	Testing Performance
Loss	0.24	28.46
Accuracy	0.91	0.55

Table 4.4.2(V): LSTM 4th tuned model performance as per the parameters (tuning type = random search epochs=15, batch size = 512)

4.4.3 GRU model performance

```
Model: "sequential"
_________________________________________________________________
 Layer (type)                Output Shape              Param #
=================================================================
 embedding (Embedding)       (None, 3800, 128)         1920000

 dropout (Dropout)           (None, 3800, 128)         0

 cu_dnngru (CuDNNGRU)        (None, 128)               99072

 dense (Dense)               (None, 64)                8256

 dropout_1 (Dropout)         (None, 64)                0

 dense_1 (Dense)             (None, 7)                 455

=================================================================
Total params: 2,027,783
Trainable params: 2,027,783
Non-trainable params: 0
_________________________________________________________________
```

Figure 4.4.3(a): Trainable parameters of gru baseline model

Figure 4.4.3(a) displays the parameters used to train the gru model. The parameter was first applied to the embedding layer. After that, a dropout layer was used on the parameters. Following this, the parameter is applied to the cudnngru layer and the initial dense layer. Parameters are then used on the dense layer and the second dropout layer. Finally, there are 2,027,783 parameters that can be trained.

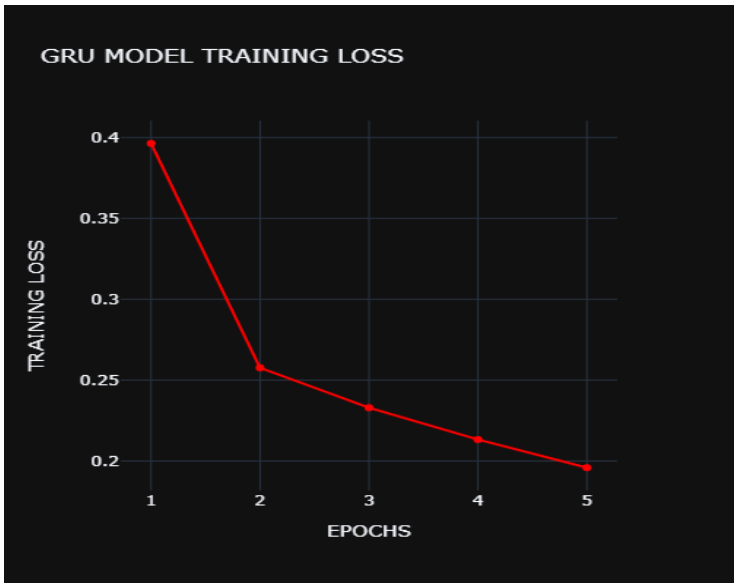

Figure 4.4.3(b): GRU baseline model training loss

Training loss for the gru baseline model is shown in Figure 4.4.3(b). The gru baseline model's training loss is roughly 0.40 after the first epoch, as shown. Training loss for the gru baseline model is almost 0.25 after the second epoch. When the third epoch was used, the training loss decreased to below 0.25. Applying the fourth epoch led to a greater reduction in loss. Finally, after the most recent period was implemented, the loss was about at 0.2.

Figure 4.4.3(c): GRU baseline model training accuracy

Training accuracy statistics for the gru baseline model are shown in Figure 4.4.3(c). As can be observed, when the first epoch is used, the gru baseline model achieves an accuracy of nearly 0.87. When the second epoch was implemented, the training accuracy increased by nearly 0.92. An increase in training accuracy from 0.88 to 0.93 was achieved in the third period of training, compared to the preceding iteration. After the fourth epoch, the training accuracy was nearly at 0.93, and by the fifth epoch, it had surpassed it.

Parameters	Training Performance	Testing Performance
Accuracy	0.93	0.58
Precision Score	0.85	0.13
Recall Score	0.80	0.11
F1 Score	0.82	0.12
Fbeta Score	0.83	0.12
MSE	0.98	5.99
RMSE	0.99	2.44

Table 4.4.3(I): GRU baseline model performance as per the parameters

Parameters	Training Performance	Testing Performance
Loss	0	0
Accuracy	0.27	0

Table 4.4.3(II): GRU 1st tuned model performance as per the parameters (tuning type = hyperband, epochs=10, batch size = 512)

Parameters	Training Performance	Testing Performance
Loss	0	0
Accuracy	0.27	0

Table 4.4.3(III): GRU 2nd tuned model performance as per the parameters (tuning type = hyperband epochs=15, batch size = 512)

Parameters	Training Performance	Testing Performance
Loss	0	0
Accuracy	0.27	0

Table 4.4.3(IV): GRU 3rd tuned model performance as per the parameters (tuning type = random search epochs=10, batch size = 512)

Parameters	Training Performance	Testing Performance
Loss	0	0
Accuracy	0.27	0

Table 4.4.3(V): GRU 4th tuned model performance as per the parameters (tuning type = random search epochs=15, batch size = 512)

4.4.4 MLP model performance

```
Model: "sequential"
_________________________________________________________________
 Layer (type)                Output Shape              Param #
=================================================================
 dense (Dense)               (None, 256)               384256

 dropout (Dropout)           (None, 256)               0

 dense_1 (Dense)             (None, 128)               32896

 dropout_1 (Dropout)         (None, 128)               0

 dense_2 (Dense)             (None, 7)                 903

=================================================================
Total params: 418,055
Trainable params: 418,055
Non-trainable params: 0
_________________________________________________________________
```

Figure 4.4.4(a): Trainable parameters of mlp baseline model

The multilayer perceptron models' trainable parameters are depicted in Figure 4.4.4(a). The dense layer receives the parameters first. Parameters are then sent to the initial dropout layer. Once more, the parameters are routed through the two dense layers. a total of 418,055 parameters are available for training.

The training loss for the mlp baseline model is displayed in Figure 4.4.4(b). The mlp baseline model's loss is close to 0.42 once the first epoch is applied. After incorporating the second epoch, the mlp baseline model's loss was nearly 0.36. The third epoch brought the training close to 0.34; the fourth and fifth epochs brought the loss down even more, to roughly 0.33 and 0.32, respectively.

Training accuracy for the mlp baseline model is displayed in Figure 4.4.4(c). As can be seen, the training accuracy is roughly 0.86 after the first epoch is applied. The training accuracy of the mlp baseline model is between 0.875 and 0.88 after using the 2ndepoch. When using a third training session, the mlp baseline model achieves a training accuracy of 0.88–0.885. The training accuracy improved even better once the 4thepoch was applied, rising to a new high of between 0.88 and 0.89. The training accuracy nearly reached 0.89 after the last epoch was applied.

Figure 4.4.4(b): MLP baseline model training loss

Figure 4.4.4(c): MLP baseline model training accuracy

Parameters	Training Performance	Testing Performance
Accuracy	0.88	0.53
Precision Score	0.79	0.13
Recall Score	0.70	0.10
F1 Score	0.73	0.11
Fbeta Score	0.74	0.12
MSE	2.01	6.77
RMSE	2.02	2.60

Table 4.4.4(I): MLP baseline model performance as per the parameters

Parameters	Training Performance	Testing Performance
Loss	0.27	7.84
Accuracy	0.90	0.53

Table 4.4.4(II): MLP 1st tuned model performance as per the parameters (tuning type = hyperband, epochs=10, batch size = 512)

Parameters	Training Performance	Testing Performance
Loss	0.26	11.01
Accuracy	0.91	0.49

Table 4.4.4(III): MLP 2nd tuned model performance as per the parameters (tuning type = hyperband epochs=15, batch size = 512)

Parameters	Training Performance	Testing Performance
Loss	0.23	8.59
Accuracy	0.91	0.58

Table 4.4.4(IV): MLP 3rd tuned model performance as per the parameters (tuning type = random search epochs=10, batch size = 512)

Parameters	Training Performance	Testing Performance
Loss	0.20	11.28
Accuracy	0.92	0.58

Table 4.4.4(V): MLP 4th tuned model performance as per the parameters (tuning type = random search epochs=15, batch size = 512)

4.5 Conclusion of sentiment analysis and news classification

The section 4.3 has helped in solving the 3rd research question. It was found most of the sentiment analysis models did the better performance than other researchers. Especially machine learning models have given best performance on the huge data also, when hyperparameters was applied the performance of the machine learning models increased. Similarly, the section 4.4 has helped in solving the 4th research question. It was found that the deep learning models have successfully was able to classify the news categories. Just for the gru model when hyper parameters applied it failed and only gru baseline model performed well.

Chapter 5. Discussion

In this part of the report, all of the goals of this project will be reviewed, and it will be determined whether or not the results met the goals.

5.1 Examination of the data collection

It was intended that the enormous dataset be utilised for this research; however, the huge dataset necessary to solve the text problem was not readily available. As a result, it was decided to scrape the data from a variety of websites and then conduct the analysis and modelling on the collected information.

5.2 Examination of the data cleaning

Since the data that has been scraped from various online sources has not been cleansed, this is an extremely crucial step to take before analysing the data for its accuracy. As a result of this, numerous different procedures for cleaning the data were carried out. Text cleaning techniques, in particular, such as removing stop words, lemmatization, removing punctuation, removing numbers from the text, removing emoticons from the text, and tokenization, were the primary area of focus. After successfully cleaning each dataset on its own, the seven datasets were combined, and additional data visualisation was applied to the combined dataset.

5.3 Examination of the data visualisation

5.3.1 Examination of sentiment analysis visualisation

The first graph that was plotted for this part was the percentage and the count of the sentiments. Both the count of the sentiments and the percentage of the sentiments were plotted on the charts. The doughnut chart displayed the percentage of the sentiments, and the bar chart displayed the count of the sentiments. After that, the sentiments associated with each category were plotted on a graph. After that, a graph of the average feelings throughout each county was created. The first research question has been answered thanks to the plots that were used in the visualisation of the sentiment analysis.

5.3.2 Examination of news visualisation

The number of news items and their respective percentages formed the basis of the initial graphic for the news classification. After that, the average news for each country was determined, followed by the average news for each weekday of the category, followed by the average news for each week, and finally, the average news for each month was determined. After that, the news was collected category by category, hour by hour. After that, a plot was created to determine the maximum amount of news for each country. The second research question has been answered as a result of these plots.

5.4 Examination of the sentiment analysis models

In this project, for the purpose of the classification of the sentiment, four models were built. These models are the multinomial naive bayes, random forest, stochastic gradient descent, and the xgboost. In order to accomplish the goals of the project, every machine learning model was altered in some way, and these new versions were compared to models that had been examined in the past. All of the baseline machine learning models had been constructed and trained on the massive amount of data, and all of the machine learning models were operating without any errors and all of the machine learning models performed exceptionally well. Now for the multinomial baseline naïve bayes the performance parameters are Accuracy-0.68, Precision Score-0.70, Recall Score-0.56, F1 Score-0.59, Fbeta Score-0.61, MAE-0.44, MSE-0.70, RMSE-0.83, Training Time(s)-0.37, ROC AUC Score-0.87. Also, after completing the naïve bayes it was found for the random forest baseline model have Accuracy-0.86, Precision Score-0.90, but the baseline model of random forest required the long training time which was 43900.46 seconds. But when the tuned model was trained the training time got reduce to 8830.47 seconds. Also, after obtaining various parameters such as accuracy score, precision score, recall score, mean absolute error, mean square error, root mean square error and the training time it was observed that the training time for the baseline stochastic gradient model is 9.38 seconds and after an applying

hyperparameter the training time increased to the 44.24. After completing the stochastic gradient descent, it was decided to do the xgboost, after designing the xgboost baseline model the performance parameters are Accuracy-0.74, Precision Score-0.74, Recall Score-0.67, F1 Score-0.64, Fbeta Score-0.65, MAE-0.32, MSE-0.45, RMSE-0.67, Training Time(s)-550.65, ROC AUC Score-0.8 but when the hyper parameters were applied the performance of the xgboost model remained unchanged xgboost model.

5.5 Examination of the news classification analysis models

For this project the objective was to use deep learning models for the classification of the news. For this purpose, four models were selected which are cnn, lstm, gru, and the mlp. After doing the designing of the cnn the baseline model of the cnn was tuned and the best parameters was obtained. For the cnn baseline mode is that the training loss is going down steadily as the epoch was increase. Similarly, the training accuracy for the cnn baseline mode was obtained. The performance parameters obtained for the cnn baseline model are Accuracy-0.93, Precision Score-0.87, Recall Score-0.80, F1 Score-0.82, Fbeta Score-0.83, MSE-0.90, RMSE-0.95. It was observed that the training accuracy was increasing steadily as the epoch was increasing. After this the lstm model was designed and the performance parameter obtained for it are Accuracy-0.94, Precision Score-0.87, Recall Score-0.83, F1 Score-0.85, Fbeta Score-0.85, MSE-0.93, RMSE-0.96, and it was tuned with the best hyperparameters. For lstm model that the training loss for the 1st epoch decreased rapidly but since second epoch the training loss was started decreasing steadily. Similarly, the training accuracy for the lstm model increased rapidly for the 1st epoch and the after1st epoch the training accuracy increased steadily. After this the gru baseline model was designed and its performance parameters are Accuracy-0.93, Precision Score-0.85, Recall Score-0.80, F1 Score-0.82, Fbeta Score-0.83, MSE-0.98, RMSE-0.99 but when the hyperparameters was applied to the baseline gru model the tuned model of the gru failed and gave the inappropriate results. For gru model that the training loss for the 1st epoch decreased rapidly but since second epoch the training loss was started decreasing steadily. Similarly, the training accuracy for the gru model increased rapidly for the 1st epoch and the after1st epoch the training accuracy increased steadily. Also, for the gru model when the hyperparameters was applied to the gru baseline model the model failed. After that the mlp baseline model was designed and then the tuned model for the mlp was designed. The performance of the mlp baseline model is Accuracy-0.88, Precision Score-0.79, Recall Score-0.70, F1 Score-0.73, Fbeta Score-0.74, MSE-2.01, RMSE-2.02.

Chapter 6. Evaluation, Reflections, and Conclusions

The primary objective of this project is to do an analysis of the data pertaining to the news and the sentiments, as well as to classify the sentiments and the news category respectively. In the beginning of this research project, the massive amount of data that had been collected was cleansed, and after that, a data visualisation of the positive, negative, and neutral feelings was performed. Following that, the visualisation for the news category is obtained. After that, the four machine learning models known as multinomial naive bayes, random forest, stochastic gradient descent, and the xgboost model are applied. On the other hand, the deep learning models known as cnn, lstm, gru, and mlp are used for the news classification. Throughout the course of the literature review, the numerous research papers were read. It was discovered that those researchers had included the baseline models in their study; nevertheless, they did not proceed with their research past the baseline model. The aspect of this study that is novel is that the model design for the deep learning model is novel, and for both the deep learning and the machine learning models, hyper parameter tuning techniques are applied. These techniques have resulted in a significant improvement in the models' ability to perform their tasks.

6.1 Project plan

The plan for the project was very different when it first got started, but as it moved forward, new things kept turning up. After the initial stage, it was determined to scrape the existing data set, however at the initial stage, it was chosen to use the existing data set. Also, it was determined that the classification of the Spanish text will be done, but later on, it was amended so that only the English language would be done, and the English news dataset was scraped from the various data sources. Additionally, it was agreed to simply perform the classification for the data pertaining to the news, but later on, it was decided to also perform the classification for the opinions. Also, considering the nature of the text data, it was presumed that the data visualisation would not be possible; however, the data visualisation component was introduced later on. In addition, it was determined to utilise an SVM model for the text categorization in the beginning; nevertheless, the data set that was obtained for this project was quite extensive. As a direct result of this, the idea was scrapped.

6.2 Reflections

The proportion and percentage of negative, neutral, and positive attitudes was discovered through the insights, and the sentiments of people all over the world regarding news from categories such as business, food, the environment, entertainment, sports, politics, and health came into picture as a result of the insights that were obtained. The insights are gathered by the amount of time spent reading various forms of online news on a daily, weekly, and monthly basis in various parts of the world. Through the use of the prophet model, one can also gain this information for maximum news throughout the world and the anticipated rate of the news data moving forward. It was found that the baseline models were working better than the work proposed by the other researchers. Also, the deep learning models performed better than other studies. The hyperparameter training techniques such as hyper band from keras and random search from keras is used which tuned the model rapidly as compare to other hyperparameter tuning technique.

6.3 Future work

As most part of the project is working well but there are still fields regarding this project which need to be improved or applied in the future are as follows:

1. As the gru baseline model worked perfectly but when the hyperparameter tuning techniques was applied to it the model failed. This issue will be address in the future.

2. The models will be deployed using flask to create the web app of this project.

3. The accuracy score of the multinomial naïve bayes was less as compare to the other models. Therefore, efforts will be made to improve the accuracy.

6.4 Conclusion

This project successfully classified the sentiments and the news on the real-world huge data set. Also, deep learning model used for this research are unique and the all the models have obtained the performance parameters such as the accuracy, precision, recall, mse, mae, rmse, training accuracy, training loss.

Glossary

Name	Definition
MSE	The term "Mean Squared Error" (MSE) refers to a common statistical measure of error. It's a metric for contrasting two numbers, such as the values predicted by a model with the actual numbers.
MAE	The acronym MAE means "Mean Absolute Error." It's a metric for contrasting two numbers, such as the values predicted by a model with the actual numbers in the real world.
RMSE	Root Mean Squared Error is short for "RMSE." It's a metric for contrasting two numbers, such as the values predicted by a model with the actual numbers in the real world. To calculate RMSE, take the root of the average squared error between the predicted and actual values.
Fbeta	The F-beta score is a metric used in machine learning and statistics to evaluate the precision of a model. Here, recall is given more importance than precision, yielding a weighted harmonic mean.
ROC	Receiver operating characteristic (ROC) curves are used in machine learning and statistics to visualise the diagnostic accuracy of a binary classifier system across different values of the discrimination threshold.
Training	Training is the procedure wherein a model is constructed by having it "fit" to a collection of training data. Training a model allows it to discover hidden patterns in data and make accurate predictions on previously undiscovered information.
Testing	Testing is a crucial element of machine learning since it allows us to see how well our trained models perform on new data.

References

Abebe Azime, I. and Mohammed, N. (2021). 'An Amharic News Text classification Dataset', (online) NASA ADS. Available at: https://ui.adsabs.harvard.edu/abs/2021arXiv210305639A/abstract (Accessed 12 Dec. 2022).

Ahmed, F., Akther, N., Hasan, M., Chowdhury, K. and Mukta, Md.S.H. (2021). 'Word Embedding based News Classification by using CNN', (online) IEEE Xplore. Available at: https://doi.org/10.1109/ICSECS52883.2021.00117

Agarwal, A., Sharma, V., Sikka, G. and Dhir, R. (2016). 'Opinion mining of news headlines using SentiWordNet', *2016 Symposium on Colossal Data Analysis and Networking (CDAN)*. Available at: https://doi.org/10.1109/cdan.2016.7570949

Antonio, V.D., Efendi, S. and Mawengkang, H. (2022). 'Sentiment analysis for covid-19 in Indonesia on Twitter with TF-IDF featured extraction and stochastic gradient descent', *International Journal of Nonlinear Analysis and Applications*, (online) 13(1), pp.1367–1373. Available at: https://doi.org/10.22075/ijnaa.2021.5735

Barua, A., Sharif, O. and Hoque, M.M. (2021). 'Multi-class Sports News Categorization using Machine Learning Techniques: Resource Creation and Evaluation', *Procedia Computer Science*, 193, pp.112–121. Available at: https://doi.org/10.1016/j.procs.2021.11.002

Clercq, O.D., Bruyne, L.D. and Hoste, V. (2020). 'News topic classification as a first step towards diverse news recommendation', *Computational Linguistics in the Netherlands Journal*, (online)10, pp.37–55. Available at: https://www.clinjournal.org/clinj/article/view/103 (Accessed 11 Dec. 2022).

Deshmukh, R. (2021). 'A Document Classification using NLP and Recurrent Neural Network', *International Journal of Engineering and Advanced Technology*, 8, pp.632-636. Available at: https://doi.org/10.35940/ijeat.F8087.088619

Deng, L., Ge, Q., Zhang, J., Li, Z., Yu, Z., Yin, T. and Zhu, H. (2022). 'News Text Classification Method Based on the GRU_CNN Model', *International Transactions on Electrical Energy Systems*, (online) 2022, p.e1197534. Available at: https://doi.org/10.1155/2022/1197534

Dien, T.T., Loc, B.H. and Thai-Nghe, N. (2019). 'Article Classification using Natural Language Processing and Machine Learning', *2019 International Conference on Advanced Computing and Applications (ACOMP)*. Available at: https://doi.org/10.1109/acomp.2019.00019

Duan, J., Zhao, H., Qin, W., Qiu, M. and Liu, M. (2020). 'News Text Classification Based on MLCNN and BiGRU Hybrid Neural Network', (online) IEEE Xplore. Available at: https://doi.org/10.1109/SmartBlock52591.2020.00032

Hama Aziz, R.H. and Dimililer, N. (2021). 'SentiXGboost: enhanced sentiment analysis in social media posts with ensemble XGBoost classifier', *Journal of the Chinese Institute of Engineers*, pp.1–11. Available at: https://doi.org/10.1080/02533839.2021.1933598

Joyce, B. and Deng, J. (2019). 'Sentiment Analysis Using Naive Bayes Approach with Weighted Reviews - A Case Study', *2019 IEEE Global Communications Conference (GLOBECOM)*. Available at: https://doi.org/10.1109/globecom38437.2019.9013588

Kalcheva, N., Karova, M. and Penev, I. (2020). 'Comparison of the accuracy and the execution time of classification algorithms for Bulgarian literary work', [online] IEEE Xplore. Available at: https://doi.org/10.1109/ICAI50593.2020.9311373

Karthika, P., Murugeswari, R. and Manoranjithem, R. (2019). 'Sentiment Analysis of Social Media Network Using Random Forest Algorithm', (online) IEEE Xplore. Available at: https://doi.org/10.1109/INCOS45849.2019.8951367

Kaur, S. (2016). 'Online news classification using Deep Learning Technique', (online) www.semanticscholar.org. Available at: https://www.semanticscholar.org/paper/Online-news-classification-using-Deep-Learning-Kaur/44e701c61381ea208c468ccc7fd6cff1c7bba447 (Accessed 12 Dec. 2022).

Li, Z., Shang, W. and Yan, M. (2016). 'News text classification model based on topic model', *2016 IEEE/ACIS 15th International Conference on Computer and Information Science* Available at: https://doi.org/10.1109/icis.2016.7550929

Luo, X. (2021). 'Efficient English text classification using selected Machine Learning Techniques', *Alexandria Engineering Journal*, 60(3), pp.3401–3409. Available at: https://doi.org/10.1016/j.aej.2021.02.009

Matta, P., Sharma, S. and Uniyal, N. (2022). 'Comparative Study Of Various Scraping Tools: Pros And Cons', *2022 IEEE Delhi Section Conference (DELCON)*. Available at: https://doi.org/10.1109/delcon54057.2022.9753358

Rana, M.I., Khalid, S. and Akbar, M.U. (2014). 'News classification based on their headlines: A review', *17th IEEE International Multi Topic Conference 2014*. Available at: https://doi.org/10.1109/inmic.2014.7097339

Rana, S. and Singh, A. (2016). 'Comparative analysis of sentiment orientation using SVM and Naive Bayes techniques', *2016 2nd International Conference on Next Generation Computing Technologies (NGCT)*. Available at: https://doi.org/10.1109/ngct.2016.7877399

Rao, V. and Sachdev, J. (2017). 'A machine learning approach to classify news articles based on location', (online) IEEE Xplore. Available at: https://doi.org/10.1109/ISS1.2017.8389300

Rustam, F., Khalid, M., Aslam, W., Rupapara, V., Mehmood, A. and Choi, G.S. (2021). 'A performance comparison of supervised machine learning models for Covid-19 tweets sentiment analysis', *PLOS ONE*, 16(2), p.e0245909. Available at: https://doi.org/10.1371/journal.pone.0245909

Saifullah, S., Fauziyah, Y. and Aribowo, A.S. (2021). 'Comparison of machine learning for sentiment analysis in detecting anxiety based on social media data'. *Jurnal Informatika*, 15(1), p.45. Available at: https://doi.org/10.26555/jifo.v15i1.a20111

Shahare, F.F. (2017). 'Sentiment analysis for the news data based on the social media', (online) IEEE Xplore. Available at: https://doi.org/10.1109/ICCONS.2017.8250692

Shiguihara, P. and Berton, L. (2022). 'Exploring Deep Neural Networks and Decision Tree for Spanish Text Classification', [online] IEEE Xplore. Available at: https://doi.org/10.1109/INTERCON55795.2022.9870087

Shirsat, V.S., Jagdale, R.S. and Deshmukh, S.N. (2017). 'Document Level Sentiment Analysis from News Articles', (online) IEEE Xplore. Available at: https://doi.org/10.1109/ICCUBEA.2017.8463638

Singh, Y.V., Naithani, P., Ansari, P. and Agnihotri, P. (2021).' News Classification System using Machine Learning Approach', (online) IEEE Xplore. Available at: https://doi.org/10.1109/ICAC3N53548.2021.9725409

Soni, S., Chouhan, S.S. and Rathore, S.S. (2022). 'TextConvoNet: a convolutional neural network based architecture for text classification', *Applied Intelligence*. Available at: https://doi.org/10.1007/s10489-022-04221-9

Stein, A.J., Weerasinghe, J., Mancoridis, S., Greenstadt, R., Stein, A.J., Weerasinghe, J., Mancoridis, S. and Greenstadt, R. (2020). 'News Article Text Classification and Summary for Authors and

Topics', *Computer Science & Information Technology (CS & IT)*, (online) 10(14), p.1. Available at: https://doi.org/10.5121/csit.2020.101401

Taj, S., Shaikh, B.B. and Fatemah Meghji, A. (2019). 'Sentiment Analysis of News Articles: A Lexicon based Approach', (online) IEEE Xplore. Available at: https://doi.org/10.1109/ICOMET.2019.8673428

Tang, R., Zhu, Z., Yao, H., Li, Y., Sun, X., Hu, G., Xie, G. and Li, Y. (2022). 'Integrating Medical Code Descriptions and Building Text Classification Models for Diagnostic Decision Support', [online] IEEE Xplore. Available at: https://doi.org/10.1109/ICHI54592.2022.00122

Tao, W. and Chang, D. (2019). 'News Text Classification Based on an Improved Convolutional Neural Network', *Tehnički Vjesnik*, (online) 26(5), pp.1400–1409. Available at: https://doaj.org/article/6b1fe8abbfb44ec79df38cc6a4a4165d (Accessed 12 Dec. 2022).

Vimali, J.S. and Murugan, S. (2021). 'A Text Based Sentiment Analysis Model using Bi-directional LSTM Networks', (online) IEEE Xplore. Available at: https://doi.org/10.1109/ICCES51350.2021.9489129

Wang, Z., Shan, X., Zhang, X. and Yang, J. (2022). 'N24News: A New Dataset for Multimodal News Classification', *arXiv:2108.13327 (cs)*. (online). Available at: https://arxiv.org/abs/2108.13327#:~:text=In%20this%20paper%2C%20we%20propose (Accessed 11 Dec. 2022).

Weng, L., Li, Q. and Xuehai, D. (2019). 'GRU based Convolutional Neural Network with Initialized Filters for Text Classification', (online) Available at: https://www.semanticscholar.org/paper/GRU-based-Convolutional-Neural-Network-with-Filters-Weng-Li/85b9995f5fcedc29e4df23877e5c4dd6c5b3285c (Accessed 12 Dec. 2022).

Wu, M.-J., Fu, T.-Y., Chang, Y.-C. and Lee, C.-W. (2020). 'A Study on Natural Language Processing Classified News', (online) IEEE Xplore. Available at: https://doi.org/10.1109/Indo-TaiwanICAN48429.2020.9181355

Yasaswi, K., Kambala, V.K., Pavan, P.S., Sreya, M. and Jasmika, V. (2022). 'News Classification using Natural Language Processing', (online) IEEE Xplore. Available at: https://doi.org/10.1109/ICIEM54221.2022.9853174

Zhao, W., Zhu, L., Wang, M., Zhang, X. and Zhang, J. (2022). 'WTL-CNN: a news text classification method of convolutional neural network based on weighted word embedding', *Connection Science*, 34(1), pp.2291–2312. Available at: https://doi.org/10.1080/09540091.2022.2117274

Appendix - A

Extra plot

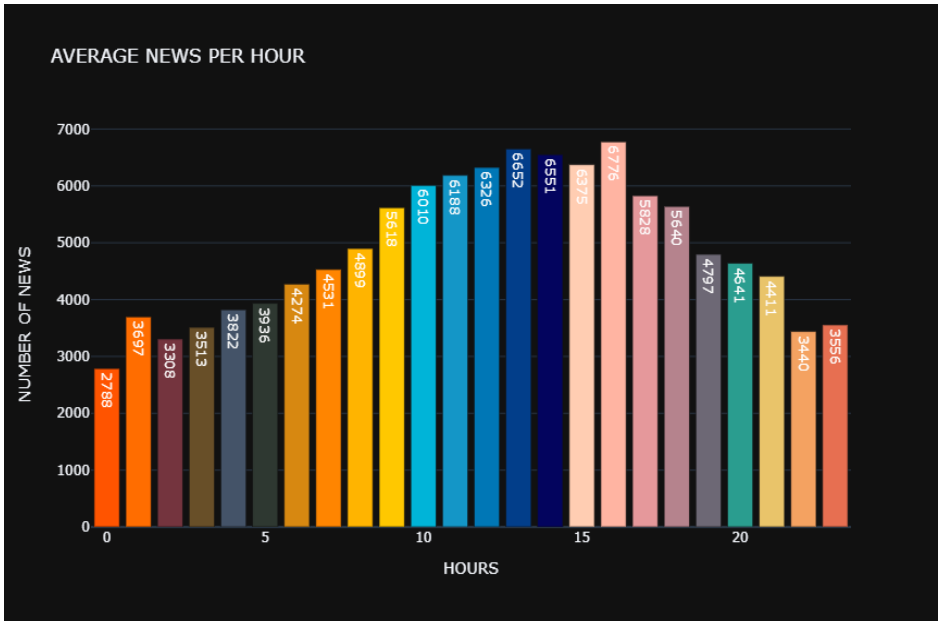

Fig A. Average news hour

It shows the average news per hour. To make this figure the hours and the category columns were grouped.

Appendix – B

Code running instructions:

The google colab link is provided below:

https://drive.google.com/drive/folders/1Z4AehVm79Z9NSb60YFCgQhLgUQ-DNUrD?usp=sharing

1. Enter the google colab link and the download the seven datasets which are “business.csv”, “food.csv”, “health.csv”, “entertainment.csv”, “environment.csv”, “sports.csv”, “politics.csv”

2. Run the file “data_collection_and_data_cleaning.ipynb” on jupyter notebook.

3. Upload the file in to the colab and run file “data_visualisation.ipynb”

4. Now run the file “naivebayes_base_and_tuned.ipynb”

5. Now run the file “random_forest_base_and_tuned.ipynb”

6. Now run the file “stochastic_gradient_descent_base_and_tuned.ipynb”

7. Now run the file “xgboost_base_and_tuned.ipynb”

8. Now run the file “cnn_base_and_tuned.ipynb”

9. Now run the file “lstm_base_and_tuned.ipynb”

10. Now run the file “gru_base_and_tuned.ipynb”

11. Now run the file “mlp_base_and_tuned.ipynb”

Appendix – C

Classification of Multi-Category Sport News Using NLP

Name: Vighnesh Mane| Email: vighnesh.mane@city.ac.uk

1. INTRODUCTION

In Natural Language Processing (NLP), the task of assigning a label from a specified tag list to a given document is a key challenge that must be solved. With labelled data, the computer may learn the patterns of previously seen data and predict labels for new, unlabeled data. Many NLP applications, such as news filtering, spam detection, sentiment analysis, and document organization, rely heavily on text classification (Varasteh et al., 2021). The main aim of this work is to apply the BERT model that can analyze Spanish sport news and multi-class classification should be possible outputs of the other models which will be applied to the data which is an existing sport news dataset which is obtained from Kaggle website. A classification model could be used to classify news into multiple categories. Spanish is a very famous language and many previous works has been done regarding in NLP and they have used pre trained BERT model. This study will not only translate the sport news from Spanish to English but it also performs the analysis of the sentiments and filtering of the particular sport news. Finally, the products will be generated are BERT model, trained features and several parameters after testing the models.

The beneficiaries from this product are as follows:

- Many members of the groups find difficulty in reading the translated news. So, if certain members want to read any specific sport news which is translated from Spanish to English will be benefitted from this product.
- Secondly, as sentiment analysis on a lexical level examine the flow of the words or phrases in the sentence to get a sense of the overall feeling of the piece (Shirsat et al., 2017). So, considering this some members of the certain group will get benefitted from this as they will be able to read the news based on the sentiments or reviews of other members on a selected sport news.

The following is the hypothesis major research question:

Will the models able to show the highest performance for the parameters which are to be selected?

2. CRITICAL CONTEXT

NATURAL LANGUAGE PROCESSING

Natural language processing (NLP) is a branch of computer science that allows people to converse with machines in their own language. Until the early 2000s, Alan Turing predicted that computers will be able to grasp and respond to conversational language in such a way that you wouldn't be able to tell the two apart (Sri et al., 2021). Processing natural language correctly is a goal of Natural Language Processing (NLP). In this way, computers will be able to understand human language. As a result, humans and computers were able to interact more

effectively. Many firms rely heavily on NLP for success. Life without Google search, Alexa, YouTube recommendations and so on is nearly impossible to fathom. Today, NLP can be found just about everywhere (Sri et al., 2021). It is possible to classify natural language processing as a subset of speech and language processing. So, there are parallels between NLP and other fields like computational linguistics, which uses rule-based models to model language (Beysolow et al., 2018). The challenge of processing natural language grows in significance. The topic includes natural language processing, data collection, and data mining (Wu et al., 2020).

MECHANISM OF THE TEXT CLASSIFICATION

It is the task of classifying a data instance into one or more known classes in machine learning. Data points can be in text, voice, image, or numeric form. To classify text, one or more input data points are used, each of which is categorized into one or more buckets (referred to as classes) from a preset bucket set (classes). The "text" can be as short as a single character or as long as a full document. Think about a case in which we wish to divide all customer reviews for a product into three categories: good, negative, and neutral. When it comes to categorization, the problem is to "learn" this categorization from a collection of samples for each category and then forecast the categories for fresh, unseen items and customer reviews. It's not always necessary to come up with a single category, and there can be any number of them (Vajjala et al., 2020).

REQUIREMENT OF THE TEXT CLASSIFICATION

Text classification is an essential part of many machine learning applications, including sentiment analysis, spam detection, and intent detection. This feature is particularly useful for recognizing languages in order to better understand client feedback and inform future efforts.

FEATURE EXTRACTION

It's not uncommon for NLP pre-processing to be more complex than just encoding a category feature with labels. In order to encode a fixed number of features across all of these observations, we need a technique to express individual observations of texts as a row. As a result, feature extraction is now the most significant part of preparing text (Beysolow et al., 2018). The selection of a feature extraction model is also influenced by the descriptive and target features that are pertinent to our objective. The performance of models is mostly based on feature extraction and engineering (Varasteh et al., 2020).

TRADITIONAL NEURAL NETWORKS

Models known as multilayer perceptron models (MLPs) have replaced single-layer perceptron models (SLPs). Originally, the SLP model couldn't properly handle data that wasn't linearly separable, therefore MLPs were invented to address this issue. There is a standard normal distribution for the weights that connect each layer of an MLP model. Each feature in a neural network is represented by a node in the input layer. It's up to the user how many hidden levels there are, but they all have the same number of nodes. The MLP classifier

makes use of a fully connected layer and a softmax classification layer. TF-IDF technique is used to convert text into TF-IDF vector in the MLP model (Li et al., 2019). There is only one node in the output layer when doing regression analysis. It has K nodes, where K is the number of classes, in classification (Beysolow et al., 2018). There are a variety of functions to choose from, depending on the nature of the problem or the data. activation functions are another term for this type of action. The types are described as follows:

- Linear activation function: The output of a linear neuron can range from a negative infinity all the way up to an infinite number (Kulkarni et al., 2019).
- Nonlinear activation function: Some of these are the most commonly used ones, and they limit the output to a certain range (Kulkarni et al., 2019).

Scaling down the result between 0 and 1 with a log function makes classification problems easier to solve. It is similar to sigmoid in that it calculates the probability of an occurrence over n different classes, which aids in multiclass categorization. Other than that, it's much the same thing as using an analogous sigmoid function. The activation function of a rectified linear unit transforms everything less than zero to zero. As a result, the range is now from 0 to infinity (Kulkarni et al., 2019).

SUPERVISED LEARNING

The aim behind supervised techniques is to manually annotate a corpus in order to automatically extract patterns and rules (Kurdi et al., 2017). For supervised classification, feature vectors and their accompanying class labels are commonly used to train the classifier (Métais et al., 2014). A supervised learning algorithm learns from labelled training data and enables you to predict the outcomes of unlabelled data. Building, developing, and deploying a precise supervised machine learning Data science model requires time. Tamil movie reviews were classified using supervised machine learning algorithms (Ramraj et al., 2020).

SUPERVISED EMBEDDINGS

It is feasible to extract more accurate embeddings in supervised situations by adding dependent variable information into the optimization model. Consider the situation in which some rows of the n d data matrix D are associated with class labels yi taken from the interval {1, 1} to {1, 1} inclusive (Aggarwal et al., 2018). Consequently, the set S of observed labels is defined as follows:

$$S = \{i : \text{Label } y_i \text{ of ith row of D is observed}\}$$

3. APPROACHES: METHODS & TOOLS

METHODS

LITERATURE SURVEY

The City University Library provided many of the cited articles, and a significant number of papers from various websites were also obtained. The articles that were judged to be the most helpful were chosen, and those with the most useful information were selected. This was done so that the techniques described in those articles may be used to keep the performance of products stable.

DATA COLLECTION

Following the completion of a literature review, the data set that was decided upon, which is the Spanish news data set that is acquired from Kaggle, was chosen. Before attempting to apply the models to this data, it is required to clean the data first because it is not in an organised format. In the subsequent stage, a description regarding data transformation and data pre-processing will be provided.

DATA TRANSFORMATION

A contextual representation with left and right contexts is represented by the BERT algorithm. It is easy to understand intellectually and in practise. BERT is superior than previous methods because it is the first unsupervised deep bidirectional system for NLP pre-training with domain-adaptive features. This distinction gives BERT a significant competitive advantage (Kulkarni et al., 2019). BERT is a masked word prediction and next sentence prediction system that is built on a multi-layer bidirectional Transformer and is trained on a huge corpus of plain text. These tasks are performed by BERT (Varasteh et al., 2021). As (Varasteh et al., 2021) used the ParsBERT which transformed the Persian language similar to that we will use BERT model related to the Spanish language which will transform the Spanish language to English language.

DATA PREPROCESSING

Stopwords are eliminated for every record in the database. Data has been de-identified and de-duplicated. Each token will have a unique index number allocated to it. Tokens are represented as vectors in fast text. Using the appropriate vectors, an embedding matrix is created for each token (Ramraj et al., 2020). This similar process of data pre-processing and data cleaning will be followed for this work.

BUILDING MODELS

After data pre-processing and extracting the features, the models will be constructed. The two models which will built and compared are support vector machine and multilayer perceptron. Several parameters will decide the performance of these models.

SVM

Regularization in a support vector machine (SVM) has a unique geometric meaning that results in a margin-based separation of points belonging to the two classes. When an SVM constructs two parallel hyperplanes on either side of the decision boundary, the majority of points fall on either side of these two margin hyperplanes on the right side (Aggarwal et al., 2018). Data can be classified using SVM's big margin classifier, which it learns to create and also trained SVM with TFIDF (Ramraj et al., 2020). For this work firstly the base SVM model will be created and after selecting the model with the best parameters the model will be further trained.

MLP

The MLP (multilayer perceptron) model consists of three layers: an input layer, a hidden layer, and an output layer. Two types of data are available: training data and data that has been confirmed. The data was then processed by the input layer. It sent the text to a matrix (Wu et al., 2020). Similarly for this study, the base MLP model will created and the model will be further trained after the optimum parameters have been selected.

EVALUATION

The evaluation of the products depends upon various parameters such as accuracy, precision, f1score, cross validation score. The results of the models will be updated to the report and if the results are not appropriate then to tackle this situation additional time period is been allocated to reperform the task where ever it is required.

TOOLS

It is been observed that the studies which are referred have performed the entire work on python. Therefore, for this work python is sufficient. But if there is requirement of various other tools apart from python then those tools will be utilized to get the results.

REPORT

The report will be initiated once the models are running successfully and the models are providing the estimated results. The report will be verified from the supervisor as per the scheduled meetings.

PROJECT MEETINGS

Meetings with the supervisor will take place and the supervisor will be updated about the progress of the project and feedback from the supervisor will be noted.

4. WORK PLAN

Flow diagram for the work plan is depicted in fig. 1.

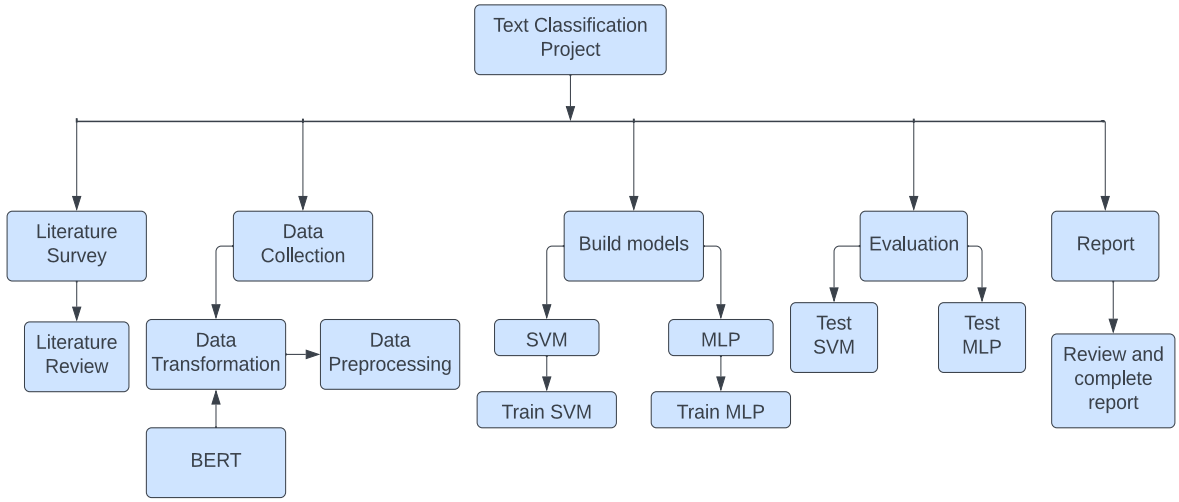

Figure 1: Flow Diagram of work plan

Work plan timeline is depicted in fig. 2.

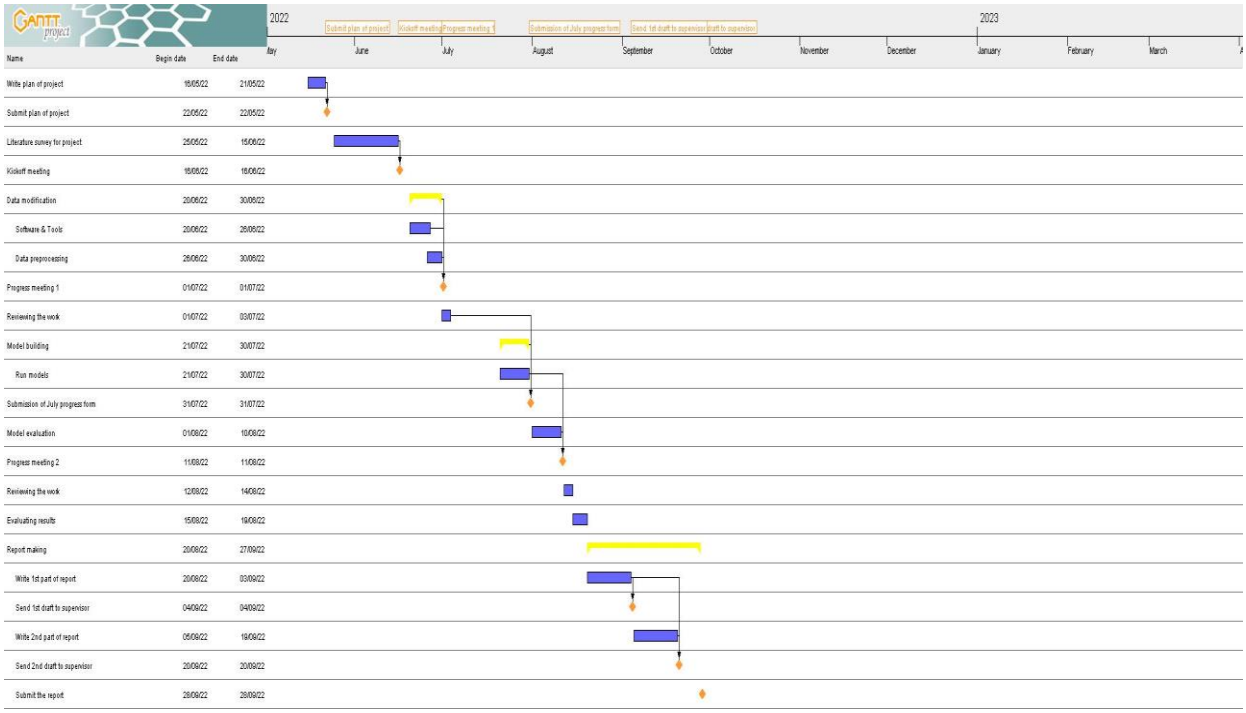

Figure 2: Work plan timeline

5. RISKS

Table 1 depicts the risk analytics of the project

Risk	Likelihood(L) (1-3)	Consequence(C) (1-5)	Impact (L * C)	Mitigation
Data might not be transformed properly.	1	2	2	Various BERT models will be utilized
Data will be not cleaned as per the requirement	2	2	4	Many text data cleaning techniques are available and almost majority of them will be applied.
Unwanted features can get extracted which can fail feature extraction	2	3	6	Feature extraction will be handled carefully as the model performance depends upon it. Consultation of supervisor is required.
Models cannot give estimated performances	3	3	9	If the model fails or cannot perform as per the requirement it will be trained and tested once again using different modifications.

Neural networks require high memory	2	2	4	Model layers will be reduced or systems will be upgraded.

Table 1: Risk register

REFERENCES

Sri, M. (2021). *Practical Natural Language Processing with Python*. Berkeley, CA: Apress. pp. 1-12.

Beysolow II, T. (2018). *Applied Natural Language Processing with Python*. Berkeley, CA: Apress. pp. 1-74.

Métais, E., Roche, M. and Teisseire, M. eds., (2014). *Natural Language Processing and Information Systems*. *Lecture Notes in Computer Science*. Cham: Springer International Publishing.

Kurdi, M.Z. (2017). *Natural Language Processing and Computational Linguistics 2*. Hoboken, NJ, USA: John Wiley & Sons, Inc. pp. 169-196.

Kulkarni, A. and Shivananda, A. (2019). *Natural Language Processing Recipes*. Berkeley, CA: Apress. pp. 135-249.

Aggarwal, C.C. (2018). *Machine Learning for Text*. [online] Cham: Springer International Publishing. pp. 235-258.

Vajjala, S., Majumder, B., Gupta, A. and Surana, H. (2020). *Practical Natural Language Processing*. O'Reilly Media, Inc. pp.119-157.

Wu, M.-J., Fu, T.-Y., Chang, Y.-C. and Lee, C.-W. (2020). *A Study on Natural Language Processing Classified News*. [online] IEEE Xplore. doi:10.1109/Indo-TaiwanICAN48429.2020.9181355.

Varasteh, M. and Kazemi, A. (2021). *Using ParsBert on Augmented Data for Persian News Classification*. [online] IEEE Xplore. doi:10.1109/ICWR51868.2021.9443119.

Rahman, Md.M., Khan, Md.A.Z. and Biswas, A.A. (2021). *Bangla News Classification using Graph Convolutional Networks*. [online] IEEE Xplore. doi:10.1109/ICCCI50826.2021.9402567.

Davoudi, S. and Mirzaei, S. (2021). *A Semantic-based Feature Extraction Method Using Categorical Clustering for Persian Document Classification*. [online] IEEE Xplore. doi:10.1109/CSICC52343.2021.9420602.

Shirsat, V.S., Jagdale, R.S. and Deshmukh, S.N. (2017). *Document Level Sentiment Analysis from News Articles*. [online] IEEE Xplore. doi:10.1109/ICCUBEA.2017.8463638.

Li, Z., Zhu, J., Luo, Z. and Liu, S. (2019). *Research on Tibetan Text Classification Method Based on Neural Network*. [online] IEEE Xplore. doi:10.1109/IALP48816.2019.9037706.

Ramraj, S., Arthi, R., Murugan, S. and Julie, M.S. (2020). Topic categorization of Tamil News Articles using PreTrained Word2Vec Embeddings with Convolutional Neural Network. *2020 International Conference on Computational Intelligence for Smart Power System and Sustainable Energy (CISPSSE)*. doi:10.1109/cispsse49931.2020.9212248.

https://www.kaggle.com/datasets/mdamsterdam/marca-spanish-sports-news (accessed 5.20.22).

Research Ethics Review Form: BSc, MSc and MA Projects

Computer Science Research Ethics Committee (CSREC)

http://www.city.ac.uk/department-computer-science/research-ethics

Undergraduate and postgraduate students undertaking their final project in the Department of Computer Science are required to consider the ethics of their project work and to ensure that it complies with research ethics guidelines. In some cases, a project will need approval from an ethics committee before it can proceed. Usually, but not always, this will be because the student is involving other people ("participants") in the project.

In order to ensure that appropriate consideration is given to ethical issues, all students must complete this form and attach it to their project proposal document. There are two parts:

PART A: Ethics Checklist. All students must complete this part. The checklist identifies whether the project requires ethical approval and, if so, where to apply for approval.

PART B: Ethics Proportionate Review Form. Students who have answered "no" to all questions in A1, A2 and A3 and "yes" to question 4 in A4 in the ethics checklist must complete this part. The project supervisor has delegated authority to provide approval in such cases that are considered to involve MINIMAL risk. The approval may be ***provisional*** – *identifying the planned research as* likely to involve MINIMAL RISK. In such cases you must additionally seek ***full approval*** from the supervisor as the project progresses and details are established. ***Full approval*** must be acquired in writing, before beginning the planned research.

A.1 If you answer YES to any of the questions in this block, you must apply to an appropriate external ethics committee for approval and log this approval as an External Application through Research Ethics Online - https://ethics.city.ac.uk/		*Delete as appropriate*
1.1	Does your research require approval from the National Research Ethics Service (NRES)? *e.g. because you are recruiting current NHS patients or staff?* *If you are unsure try - https://www.hra.nhs.uk/approvals-amendments/what-approvals-do-i-need/*	**NO**
1.2	Will you recruit participants who fall under the auspices of the Mental Capacity Act? *Such research needs to be approved by an external ethics committee such as NRES or the Social Care Research Ethics Committee - http://www.scie.org.uk/research/ethics-committee/*	**NO**

1.3	Will you recruit any participants who are currently under the auspices of the Criminal Justice System, for example, but not limited to, people on remand, prisoners and those on probation? *Such research needs to be authorised by the ethics approval system of the National Offender Management Service.*	**NO**
A.2 If you answer YES to any of the questions in this block, then unless you are applying to an external ethics committee, you must apply for approval from the Senate Research Ethics Committee (SREC) through Research Ethics Online - https://ethics.city.ac.uk/		*Delete as appropriate*
2.1	Does your research involve participants who are unable to give informed consent? *For example, but not limited to, people who may have a degree of learning disability or mental health problem, that means they are unable to make an informed decision on their own behalf.*	**NO**
2.2	Is there a risk that your research might lead to disclosures from participants concerning their involvement in illegal activities?	**NO**
2.3	Is there a risk that obscene and or illegal material may need to be accessed for your research study (including online content and other material)?	**NO**
2.4	Does your project involve participants disclosing information about special category or sensitive subjects? *For example, but not limited to: racial or ethnic origin; political opinions; religious beliefs; trade union membership; physical or mental health; sexual life; criminal offences and proceedings*	**NO**
2.5	Does your research involve you travelling to another country outside of the UK, where the Foreign & Commonwealth Office has issued a travel warning that affects the area in which you will study? *Please check the latest guidance from the FCO* - http://www.fco.gov.uk/en/	**NO**
2.6	Does your research involve invasive or intrusive procedures? *These may include, but are not limited to, electrical stimulation, heat, cold or bruising.*	**NO**
2.7	Does your research involve animals?	**NO**
2.8	Does your research involve the administration of drugs, placebos or other substances to study participants?	**NO**
A.3 If you answer YES to any of the questions in this block, then unless you are applying to an external ethics committee or the SREC, you must apply for approval from the Computer Science Research Ethics Committee (CSREC) through Research Ethics Online - https://ethics.city.ac.uk/ **Depending on the level of risk associated with your application, it may be referred to the Senate Research Ethics Committee.**		*Delete as appropriate*
3.1	Does your research involve participants who are under the age of 18?	**NO**

3.2	Does your research involve adults who are vulnerable because of their social, psychological or medical circumstances (vulnerable adults)? *This includes adults with cognitive and / or learning disabilities, adults with physical disabilities and older people.*	**NO**
3.3	Are participants recruited because they are staff or students of City, University of London? *For example, students studying on a particular course or module.* *If yes, then approval is also required from the Head of Department or Programme Director.*	**NO**
3.4	Does your research involve intentional deception of participants?	**NO**
3.5	Does your research involve participants taking part without their informed consent?	**NO**
3.5	Is the risk posed to participants greater than that in normal working life?	**NO**
3.7	Is the risk posed to you, the researcher(s), greater than that in normal working life?	**NO**
A.4 If you answer YES to the following question and your answers to all other questions in sections A1, A2 and A3 are NO, then your project is deemed to be of MINIMAL RISK. **If this is the case, then you can apply for approval through your supervisor under PROPORTIONATE REVIEW. You do so by completing PART B of this form.** **If you have answered NO to all questions on this form, then your project does not require ethical approval. You should submit and retain this form as evidence of this.**		*Delete as appropriate*
4	Does your project involve human participants or their identifiable personal data? *For example, as interviewees, respondents to a survey or participants in testing.*	**NO**