Social media news filtering for underage users

A thesis
Submitted in partial fulfillment of the requirements for the Degree of
Bachelor of Science in Computer Science and Engineering

Submitted by

Parsia Akter	**170204011**
Nusrat Jahan Tonoya	**170204012**
Takia Maliha	**170204037**
Tahiya Ahmed Chowdhury	**170204048**

Supervised by

Prof. Dr. Md. Shahriar Mahbub

Department of Computer Science and Engineering
Ahsanullah University of Science and Technology

Dhaka, Bangladesh

June 30, 2022

CANDIDATES' DECLARATION

We, hereby, declare that the thesispresented in this report is the outcome of the investigation performed by us under the supervision of Prof. Dr. Md. Shahriar Mahbub, Department of Computer Science and Engineering, Ahsanullah University of Science and Technology, Dhaka, Bangladesh. The work was spread over two final year courses, CSE4100: Project and Thesis I and CSE4250: Project and Thesis II, in accordance with the course curriculum of the Department for the Bachelor of Science in Computer Science and Engineering program.

It is also declared that neither this thesis nor any part thereof has been submitted anywhere else for the award of any degree, diploma or other qualifications.

Parsia Akter
170204011

Nusrat Jahan Tonoya
170204012

Takia Maliha
170204037

Tahiya Ahmed Chowdhury
170204048

CERTIFICATION

This thesis titled, “**Social media news filtering for underage users**”, submitted by the group as mentioned below has been accepted as satisfactory in partial fulfillment of the requirements for the degree B.Sc. in Computer Science and Engineering in June 30, 2022.

Group Members:

Parsia Akter	**170204011**
Nusrat Jahan Tonoya	**170204012**
Takia Maliha	**170204037**
Tahiya Ahmed Chowdhury	**170204048**

Prof. Dr. Md. Shahriar Mahbub
Professor & Supervisor
Department of Computer Science and Engineering
Ahsanullah University of Science and Technology

Prof. Dr. Mohammad Shafiul Alam
Professor & Head
Department of Computer Science and Engineering
Ahsanullah University of Science and Technology

ACKNOWLEDGEMENT

We would like to express our sincere gratitude to our research supervisor Dr. Md. Shahriar Mahbub, Professor, Department of Computer Science and Engineering, Ahsanullah University of Science and Technology, for his exemplary mentoring, constant encouragement, constructive criticism and fervent support throughout this research work. His humble suggestions and interesting ideas always pushed us to take up new challenges and deliver our best.

We would like to earnestly thank Dr. Mohammad Shafiul Alam, Professor & Head of the Department of Computer Science and Engineering of Ahsanullah University of Science and Technology, and all our respected teachers, lab assistants, and friends for their moral support and helpful discussions.

We would also wish to thank our family members for their support and words of encouragement throughout our study.

Dhaka
June 30, 2022

Parsia Akter

Nusrat Jahan Tonoya

Takia Maliha

Tahiya Ahmed Chowdhury

ABSTRACT

Social media has been one of the most popular forms of communication between people of all ages, genders and classes all over the world for quite some time now. Millions of uncategorized and unfiltered news as well as information: fake, real, sensitive, censored, important, cautionary,etc., are being shared everyday. Although the minimum user age requirement on most social media platforms like Facebook, Twitter, Instagram, Google, Quora, Reddit, etc., is thirteen, however, all news is not appropriate for under eighteen users. Also, there is a trend among people to use fake credentials on social media. Thus the need for identifying under eighteen users and providing age-specific news for them arises. That being so, we have proposed an idea for displaying age-based news on social media platforms by simultaneously detecting the user as under eighteen or above eighteen by analyzing the texts from his posts and the incoming news as under eighteen appropriate or above eighteen appropriate by text analysis and lastly display or block the news for the respective user accordingly. The dataset for age classification was chosen from Kaggle and the dataset for news classification was created by merging another Kaggle dataset that contained adult news articles with a manually made dataset containing kids news articles by scraping. For user's age classification, we have used Logistic Regression, Support Vector Classifier, Random Forest Classifier, Decision Tree, Naive Bayes and K-NN. Whereas for news classification, we opted for the newly emerging artificial neural network LSTM and a functional model with the popular BERT pre-processor and encoder. Among the models for age classification, Random Forest displayed the highest accuracy of 91.31% while the LSTM sequential model displayed 96.16% accuracy, that is a bit higher than the functional BERT pre-processed model for news classification.

Contents

List of Figures

List of Tables

Social media news filtering for underage users

A thesis
Submitted in partial fulfillment of the requirements for the Degree of
Bachelor of Science in Computer Science and Engineering

Submitted by

Parsia Akter	**170204011**
Nusrat Jahan Tonoya	**170204012**
Takia Maliha	**170204037**
Tahiya Ahmed Chowdhury	**170204048**

Supervised by

Prof. Dr. Md. Shahriar Mahbub

Department of Computer Science and Engineering
Ahsanullah University of Science and Technology

Dhaka, Bangladesh

June 30, 2022

CANDIDATES' DECLARATION

We, hereby, declare that the thesispresented in this report is the outcome of the investigation performed by us under the supervision of Prof. Dr. Md. Shahriar Mahbub, Department of Computer Science and Engineering, Ahsanullah University of Science and Technology, Dhaka, Bangladesh. The work was spread over two final year courses, CSE4100: Project and Thesis I and CSE4250: Project and Thesis II, in accordance with the course curriculum of the Department for the Bachelor of Science in Computer Science and Engineering program.

It is also declared that neither this thesis nor any part thereof has been submitted anywhere else for the award of any degree, diploma or other qualifications.

Parsia Akter
170204011

Nusrat Jahan Tonoya
170204012

Takia Maliha
170204037

Tahiya Ahmed Chowdhury
170204048

CERTIFICATION

This thesis titled, “**Social media news filtering for underage users**”, submitted by the group as mentioned below has been accepted as satisfactory in partial fulfillment of the requirements for the degree B.Sc. in Computer Science and Engineering in June 30, 2022.

Group Members:

Parsia Akter	**170204011**
Nusrat Jahan Tonoya	**170204012**
Takia Maliha	**170204037**
Tahiya Ahmed Chowdhury	**170204048**

Prof. Dr. Md. Shahriar Mahbub
Professor & Supervisor
Department of Computer Science and Engineering
Ahsanullah University of Science and Technology

Prof. Dr. Mohammad Shafiul Alam
Professor & Head
Department of Computer Science and Engineering
Ahsanullah University of Science and Technology

ACKNOWLEDGEMENT

We would like to express our sincere gratitude to our research supervisor Dr. Md. Shahriar Mahbub, Professor, Department of Computer Science and Engineering, Ahsanullah University of Science and Technology, for his exemplary mentoring, constant encouragement, constructive criticism and fervent support throughout this research work. His humble suggestions and interesting ideas always pushed us to take up new challenges and deliver our best.

We would like to earnestly thank Dr. Mohammad Shafiul Alam, Professor & Head of the Department of Computer Science and Engineering of Ahsanullah University of Science and Technology, and all our respected teachers, lab assistants, and friends for their moral support and helpful discussions.

We would also wish to thank our family members for their support and words of encouragement throughout our study.

Dhaka
June 30, 2022

Parsia Akter

Nusrat Jahan Tonoya

Takia Maliha

Tahiya Ahmed Chowdhury

ABSTRACT

Social media has been one of the most popular forms of communication between people of all ages, genders and classes all over the world for quite some time now. Millions of uncategorized and unfiltered news as well as information: fake, real, sensitive, censored, important, cautionary,etc., are being shared everyday. Although the minimum user age requirement on most social media platforms like Facebook, Twitter, Instagram, Google, Quora, Reddit, etc., is thirteen, however, all news is not appropriate for under eighteen users. Also, there is a trend among people to use fake credentials on social media. Thus the need for identifying under eighteen users and providing age-specific news for them arises. That being so, we have proposed an idea for displaying age-based news on social media platforms by simultaneously detecting the user as under eighteen or above eighteen by analyzing the texts from his posts and the incoming news as under eighteen appropriate or above eighteen appropriate by text analysis and lastly display or block the news for the respective user accordingly. The dataset for age classification was chosen from Kaggle and the dataset for news classification was created by merging another Kaggle dataset that contained adult news articles with a manually made dataset containing kids news articles by scraping. For user's age classification, we have used Logistic Regression, Support Vector Classifier, Random Forest Classifier, Decision Tree, Naive Bayes and K-NN. Whereas for news classification, we opted for the newly emerging artificial neural network LSTM and a functional model with the popular BERT pre-processor and encoder. Among the models for age classification, Random Forest displayed the highest accuracy of 91.31% while the LSTM sequential model displayed 96.16% accuracy, that is a bit higher than the functional BERT pre-processed model for news classification.

Contents

List of Figures

List of Tables

Chapter 1

Introduction

1.1 Overview

In a world full of thrilling technological breakthroughs, social media is and has been one of the most persistent ones. As per the Global social media statistics research summary 2022, among the 58.4 percent social media users of the total global population [1], there exists people of all ages, gender, nationality, religion and background. Along with all the distinct ways the platform is being used by the world population, sharing of news and information contributes to a notable fraction. Also, as users may use false credentials, hence our initiative of classifying user's age and news articles to ensure the presentation of age-appropriate news to users arise.The proportion of research done solely on age classification is fairly good. But research on age based news classification falls far behind from that. Hence our research proposes a new fix to display adult news for eighteen plus users and kids news to under eighteen users by classifying user's age and news articles after exploring the texts from his posts and concurrently inspecting the news posts of his timeline or news feed.

1.2 Context

Social media is bridging the gap between a diverse group of people scattered all across the world. Among them are adults as well as children. Analyzing an article about E-Safety Information by Beechfield School, we note that on most social media sites, the eligible age for creating an account is thirteen. [2] And although all social media sites dictate a minimum user age requirement, yet underage users may create accounts by providing false information. Also, social media contains multidimensional content, including news articles, all of which may not be appropriate for users under a certain age limit, especially under eighteen ones. In fact, the age of majority is considered as the threshold of adulthood as recognized or declared in law. [3] The threshold is eighteen plus in most countries of the world. This may be visualized in the Figure 1.1 collected from [4]

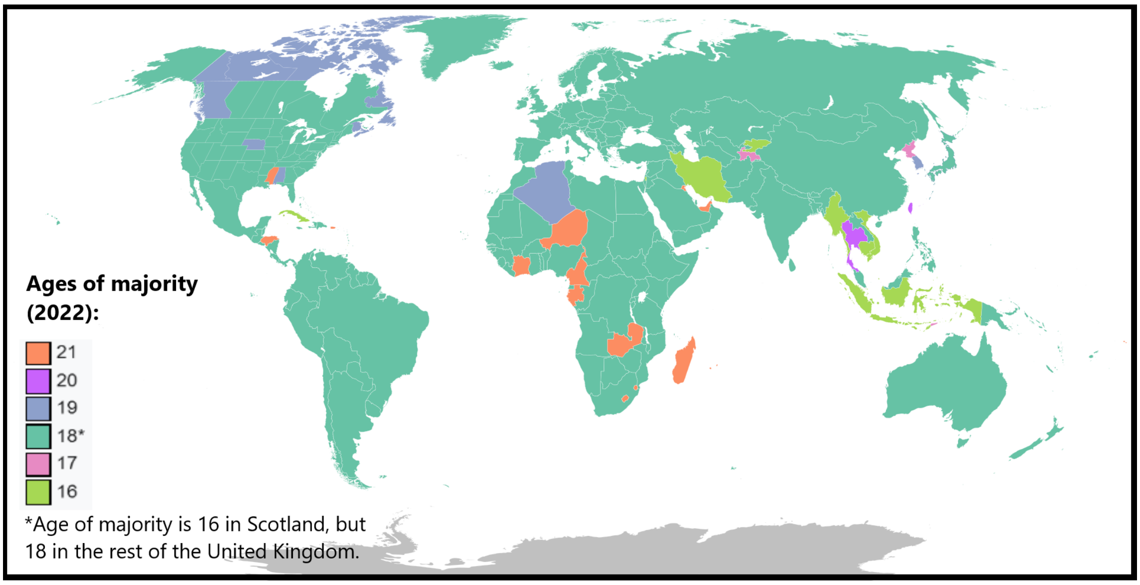

Ages of majority
(2022):
21
20
19
18*
17
16
*Age of majority is 16 in Scotland, but
18 in the rest of the United Kingdom.

Figure 1.1: Demographics of Global Age of Majority

Almost 5.6% of the users of the popular social media site Facebook are under eighteen ones [1], which may be noted in the Figure 1.2 below.

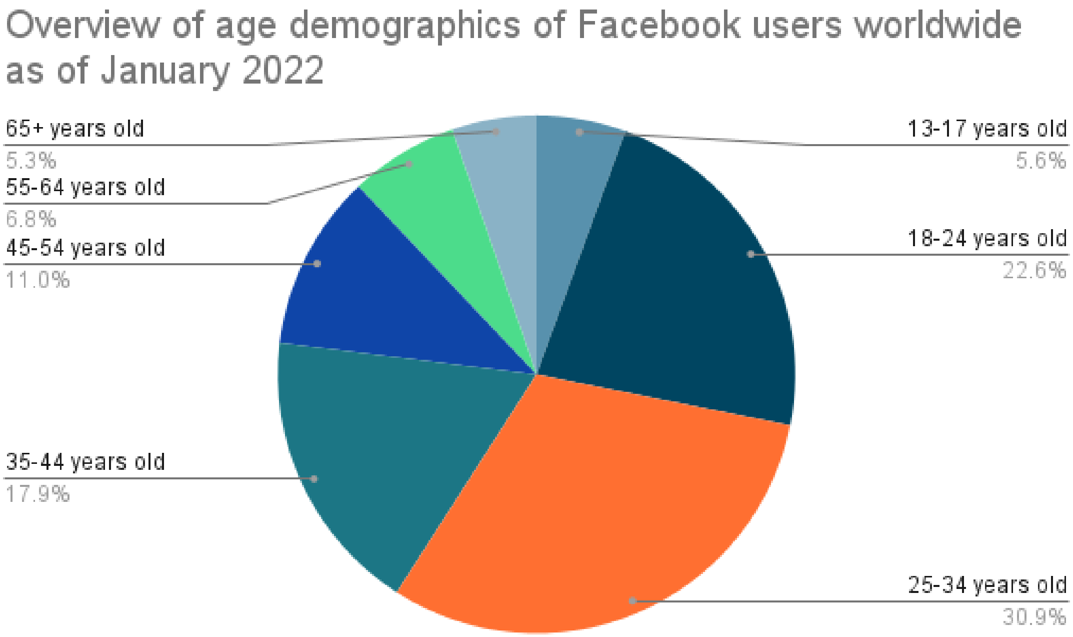


Figure 1.2: Age Distribution of Facebook Users

Some news may terrorize, misguide, frighten, expose too much information or simply mislead children. Thus news filtering may be explored.

1.3 Motivation

From an article about E-Safety Information by Beechfield School, we get to know that social media platforms like Facebook, Reddit, Pinterest, Tumblr, Linkedin, Myspace, etc., allow under eighteen people [2] . While others like YouTube and Flickr allow minors that are the ones under eighteen with adult permission. News usually contains beneficial information which may play a role in keeping children updated, improving their command over the language, enhancing their storehouse of knowledge, enriching their vocabulary, encouraging analytical thinking, shaping their writing skills and thus contributing to the most important aspects of child development like cognitive, social and emotional, speech and language, etc. According to a 2021 research by Pew Research Center, the percentage of US adults using social media sites for their daily news source is noteworthy. [5] An overview of which may be included below in Figure 1.3 :

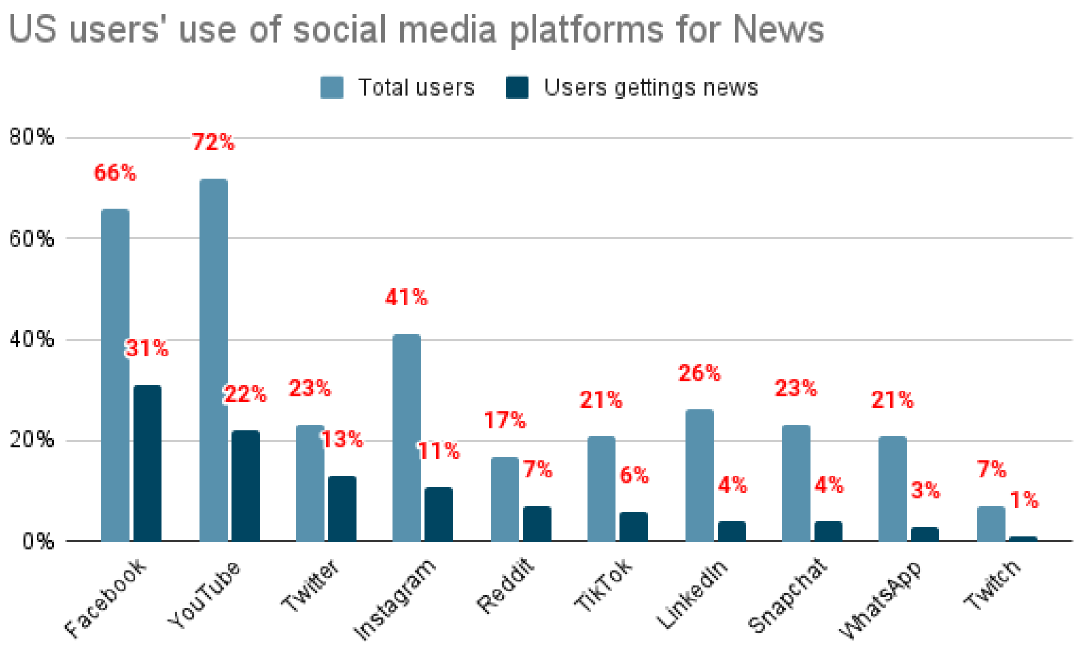


Figure 1.3: News Consumption of US Facebook Users from Social Media Sites

With the leaping advancement of technology, internet safety for kids is turning into an alarming issue. They are more susceptible to the potential dangers of malicious content online. Thus the urge to provide a safe space for them also motivated us to brainstorm solutions for the chalked-out problems. Although we cannot eliminate all problems at once yet, we believe that it is a good place to start. Hence the need for identifying the under eighteen users as well as the authentic above eighteen users and filtering news for the two age groups accordingly arises.

1.4 Summary

The first obstacle we faced while starting our research was the unavailability of any kids news dataset. All the prevailing news datasets contained only the standard news. Thus we had to collect kids news articles from different kids news websites by web scraping and form a complete dataset containing both adults and kids news proportionately. Then instead of filtering all posts based on some typically common features, it seemed wiser and meaningful to consider the context of the posts alongside the other prime aspects. Hence to solve one of the many problems, we have decided to utilize the applications of Natural Language Processing. We have proposed a news filtering approach that detects the user's age class as eighteen plus or under eighteen from the texts collected from his posts and parallelly classifies the news in his social media feed as standard or adult news and kids news to display accordingly. For users age classification, the machine learning models Logistic Regression, Support Vector Classifier, Random Forest Classifier, Decision Tree, Naive Bayes and K-nn were used. While for news classification, we decided to experiment between the artificial neural network Long short-term memory (LSTM) and a functional model with BERT pre-processing. The highest accuracy reported for the two models were 91.31% by Random Forest and 96.16% by LSTM respectively.

Chapter 2

State Of The Art

2.1 Overview

In this chapter, we will shortly discuss some of the previous works related to the subject matter at hand. A list of the state of the art we have found which are related to our work is provided as follows:

- Mining the Blogosphere: Age,gender and the varieties of self-expression [6]
- Predicting age and gender in online social networks [7]
- Age and Gender Identification in Social Media [8]
- Automatic Age Detection Using Text Readability Features [9]
- Effects of Age and Gender on Blogging [10]
- Twitter News Classification Using SVM [11]
- Web Filtering Using Text Classification [12]
- A Combined Topical/Non-topical Approach to Identifying Web Sites for Children [13]
- Detection of News Feeds Items Appropriate for Children [14]
- Applications of Deep Learning in News Text Classification [15]
- Deep Learning-Based Classification of News Texts Using Doc2Vec Model [16]
- Online news classification using Deep Learning Technique [17]
- Bengali News Classification Using Long Short-Term Memory [18]

2.2 Descriptions of Papers

Here, the details of the papers that we have studied for our research work are stated below:

- Factor analysis and machine learning techniques were applied in paper [6] .Thus a clear pattern of age and gender linked variation in writing topic and style were demonstrated by them. In this research, blogs were collected from blogger.com. Here, each blogger was categorized by age '10s'(ages 13-17) ,'20s'(ages 20-27) and so on. They removed the boundary ages. For classifying author groups, the machine learning algorithms Bayesian Multinomial Logistic Regression(BMR) and multi-class balanced real-valued Winnow(WIN) were applied. An accuracy of 77.4 % was found after using BMR and 75.0 % was found after using WIN for predicting author age. An accuracy of 79.3 % was found after using BMR and 80.5 % was found after using WIN for predicting author gender.

- Text categorization of short text is done in paper [7] , to identify the age and gender of a person. The information of the user's profile in a social media is matched with the predicted data and thus various false profiles are identified. The system also classifies adults and adolescents. It presents a study on a corpus of 1,537,283 Flemish-Dutch posts from the Belgian social networking site Netlog, which were able to obtain together with the user's profile data. It was their main objective to develop a useful component in a pedophile detective system. They use Natural Language Processing-text analysis. Their approach to this computational stylometry task is based on text categorization and involves the creation of document representations based on a selected set of patterns of features, feature selection using statistical techniques, and classification using Machine Learning algorithms. The stylistic difference in usage of non-dictionary words combined with content words allowed to predict the age group (10s, 20s, 30s or higher) with an accuracy of 80.32% and gender group with an accuracy of 89.18%. Here, the SVM classifier yielded an accuracy of 71.3% for classifying the age of min16 vs. plus16. This is to be compared to the random baseline of 50.0 %. As the distance between the age groups increases, the accuracy rose to 80.8 % for min16 vs. plus18 and even to 88.2 % for min16 vs. plus25.

- Paper [8] classify text as either blog post, tweet, hotel reviews or any other social media post. The collection of posts of a single user is present in both English and Spanish. Content based features and stylish based features, which includes various sub-categories, are extracted in the feature extraction step. LP and CC models are trained for each genre and language combination. LP(Label powerset transformation) turns a multi-label classification problem into a single label one by unifying levels. CC(Classifier chains) approach utilizes two single label classification in which the prediction made by the first is used as a feature in the second. SVM underlying learning algorithm has been used for both the models. The accuracy of the models were evaluated using the scikit-learn. A simple majority class baseline model has been used for comparison purposes.

- Paper [9] presents us the results of automatic age detection based on very short texts as about 100 words per author. 10 different models were evaluated and compared by calculating the f-scores and 10-fold cross validation. No specific categorization was made in selecting short text. Text was average 98 words long and collected in the same language (Estonian). Datasets were presented with two age groups ,that is young and adult. 14 different features were extracted. For classification they tested 6 popular machine learning algorithms,which are ,Logistic regression, SVM, K-nearest neighbor classifier, Naïve Bayes and Adaboost and used java implementation of them. Model generated by Support Vector Machine(SVM) with Adaboost yielded to f-score 0.94 and Logistic regression to 0.93. Later on, they implemented their prototype age detection application with Logistic regression, as it best performs without using Adaboost. Implemented feature extraction routine and classification function in client-side javascript. Used written natural language text as an input to extract features. A new and simple algorithm was created for syllable counting. Finally created a web application.

- From ten thousand blogs which contains almost 300 million words the difference in the writing styles of the male and female bloggers as well as bloggers of different ages were noted in paper [10] . Such differences has been exploited to find the age and gender of an unknown author on the basis of blog's vocabulary. Style based and content based features are considered in case of writing blog.Learning algorithm Multi-Class Real Winnow (MCRW) has been used to learn models that classify blogs according to author's gender and age respectively.Word class frequency (per 10000 words) and standard error by age and gender has been calculated.10-fold cross-validation experiments has been done which shows the accuracy results.Accuracy result for 10-fold cross-validation in case of gender is 80.01%. Age groups are divided into categories 10s (13-17), 20s (23-27) and 30s (33-42). Using content and style features together, 10s are distinguishable from 30's with an accuracy above 96% and 10s are distinguishable from 20s with an accuracy of 87.3%. However, many 30s are misclassified as 20s yielding overall accuracy as 76.2%.

- The research on paper [11] was conducted on a Twitter dataset containing news headlines with a goal to classify news into different groups so that the user could identify the most popular news group in a given country for a given time. The dataset was collected by extracting news from Srilankan Twitter news groups such as 'Ada Derana', 'Ceylon Today', 'lIN Sri Lanka', 'Lanka Breaking News', 'Lanka E News' and 'Sri

Lanka News Now' with the help of Twitter API. The headlines were then classified into twelve groups: war-terrorist-crime, economy, business, health, sports, development-government, politics, accident, entertain, disaster-climate, education, society and international. After that a model was created using the bag-of-words approach where each word is used as a feature and frequency of each word is used as a data. Like any other natural language processing model, the common words were removed from the data. As the categorized groups are not prone to any changes that regularly thus Support Vector Machine, a supervised learning approach was used to train the model. 90 percent data was used for training purposes while the rest 10 percent for testing purposes. For evaluating the performance of the model, average precision, average recall, F-measure and Fβ scores were observed for each of the twelve categories individually. The precision values ranged from 0.8171581 to 1 and recall from 0.3924145 to 0.9. While the F-measure ranged from 0.538703 to 0.8851583 and Fβ from 0.6939138 to 0.9233.

- A web filtering system in paper [12] was proposed, which will provide protection against inappropriate content by blocking access to pages that are against a defined policy and hence prevent misuse of the network. The similarity between the content of testing website was calculated with that of the training dataset containing contents from forbidden websites by finding out the cosine value. If the value exceeds the minimum threshold value then that webpage is blocked. For the threshold value, after creating a new dataset with contents from allowed pages and classifying the new dataset on the basis of the actual dataset, the similarity value which satisfied most members of the new dataset was chosen. As the filtering system is text-based thus it works faster compared to the filtering based on blacklists and whitelists, keyword blocking and rating systems.However, the system lacks the ability of filtering based on non textual elements. Various webpages were collected and the web filtering module was established which was then incorporated with the firewall toolkit (FWTK). The maximum average blocking rate i.e correctly blocked websites with respect to the actual number of websites to be blocked was 99.35 percent. While the minimum average over blocking rate i.e incorrectly blocked websites was 0 percent. Their system undoubtedly provided better results than the other discussed systems in their research. But in case of undecryptable network traffic by the firewall, it fails to access the webpage and hence fails to classify or filter it out.

- This research on paper [13] aims to improve child-appropriate web search engine performance by proposing an automatic way of distinguishing web pages for children from those for adults. One thing they emphasized on was that in case of identifying any children suitable content ensuring appropriateness exceeds filtering offensive content. The dataset was collected from the Open Directory Project where the ODP editors decided that a proper child web page should be informative, age-appropriate, non-commercial and for children, not about children. The website appropriateness for children was detected based a number of categories including complexity of text, average number of words per sentence, number of complex (3+ syllables) words, readability scores, part-of-speech features, html features, visual features as well as presentational, navigational and ethical aspects. A training set of 20,778 web pages (6225 for children and 14553 for adults) from a range of 1350 distinct topics was selected. Only 26.71 percent of the kids page seemed suitable exclusively for children. After using a logistic regression classifier trained on the categorized feature sets and evaluating using 10- fold stratified cross validation, the precision, recall, F0.5-measure and ROC was calculated for the different categories.

- The main goal of paper [14] was to ensure children's content that is appropriate, accessible and sensitive to their age, developmental stage, and level of understanding. They deemed classifying children's content far more complex than standard text based classification. For the dataset, two feeds: (i) standard news feed (referred to as BBC), and (ii) children's news feed (CBBC) were selected. But 51,833 entries were found for standard news feeds while only 1,832 entries were found for kids. In order to determine which features were the best at distinguishing between classes, information gain (IG) was calculated to assess the measures. Automated Readability Index came out with the highest IG i.e. 0.03. After discarding less frequent datas, the feature space was reduced to around 17,000 terms. Two features one based on bag-of-word tokens from the feed entries, and the other based on the eight readability measures was extracted. Aims to determine whether short snippets for individual feed entries are appropriate for children. The model was trained with both Naive Bayes and Support vector machine. Among both, SVM provided better results.

- The classification of news text data were done in paper [15], using a customized algorithm known as DCLSTM-MLP model.This algorithm is the combination of DL algorithms such as CNN, LSTM, and MLP . Word vector and word dispersion were used to express the proposed model. Word vector expresents the relationship among words as an input of the CNN module whereas an input of the MLP module represents the relationship between words and categories. Multiple experiments were performed for checking the stability and performance of the proposed method. It solves the problems of text length, feature extraction problems in the news text and classify the news text in an effective way. It attained better accuracy, recall rate and comprehensive value as compared to the other models and it's accuracy is 94.82%.

- They created a Doc2Vec model in paper [16] and compared the success rates by classifying the datasets created by preprocessing the TTC-3600 and BBC-News datasets. Here, the documents were divided into 90% training and 10% testing data. Then, the datasets were classified using the deep learning model CNN and traditional machine learning methods GNB, RF, NB and SVM. For classifying Turkish news texts, the highest accuracy rate they obtained was 94.17% as a result of the CNN classification of the PV-DM model of the Clean+Stem-DS dataset. For classifying the English dataset, the highest accuracy rate they obtained was 96.41% by applying the PV-DM model of the C-DS data set with CNN.

- Paper [17] has the main aim to increase the accuracy in predicting the popularity of online news. Results found four types of categories, which are politics, financial and sports.Neural Network was used in classification for obtaining better results. It has obtained an accuracy of 99.93% and has provided good results in compared to the traditional methods.

- The research on paper [18] was developed for improving the results on Bengali news text classification. It had the target of showing the improved accuracy using long short-term memory (LSTM). They have used a dataset containing 13,445 news collected from various newspapers. They used the embedding layer, spatial dropout layer, LSTM layer and dense layer to made their model. This model has classified 5 categories of news and this experiment has achieved a good accuracy of 84%.

2.3 Research Gap

Many kinds of research on age, gender, nationality, etc., detection from user's social media post content have been performed. These allow us to detect various traits of the user. Also, news classification has been proposed mostly based on the domains like international, politics, sports, business, entertainment, etc. But very limited research on age-based news classification which basically refers to classifying the news as adult news, the standard news we mostly read everyday, and kids news that is news for children, has been done. Also no such proposition merging the two ideas have yet been suggested where the age-appropriate news will be shown to user by detecting their age based on the posts they share.

2.4 Summary

In this chapter, we have discussed the papers that we have studied related to our research work. We get similar kind of papers that has done the age and gender classification but not the exact one that has done the underage and 18 plus users classification. Similarly, in the case of news, we did not get any paper that has done the kids-friendly news and adult news classification using either ML or ANN models. We get papers classifying multiple categories of news.

Chapter 3

Background Study

3.1 Overview

In this chapter, we will discuss some important topics we need to cover before starting our research work. These terms include the idea of Machine Learning and its models, Neural Network and terms related to it, LSTM (ANN model), various feature selection and feature extraction algorithms and also many more. It is a prerequisite to provide an explanation for these terms.

3.2 Machine Learning

Machine learning is a branch of artificial intelligence (AI) and computer science. The use of data and algorithms is the main focus of it. It is used to imitate the way that humans learn by gradually improving its accuracy.

The ML models used for classification problems are discussed below.

3.2.1 Logistic Regression

Logistic Regression is also known as logit model. It is a supervised machine learning algorithm often used for classification and predictive analytics. It estimates the probability of occurring of an event.For example a person has voted or not. The probability of occurence is based on a given dataset of independent variables. The dependent variable is bounded between 0 and 1 as the outcome is a probability.
This model suffers from overfitting particularly when there is a high number of predictor variables within the model. When the model suffers from high dimensionality, regulariza-

tion is typically used to penalize parameters large coefficients. [19]
Logistic function or sigmoid function is used to calculate probability in logistic regression. It is a simple S-shaped curve used to convert data into a value between 0 and 1.
The equation is as follows : [20]

$$\text{h0(x)} = 1/1 + \text{e - } (\beta 0 + \beta 1X)$$

Here,
h0(x) : output of logistic function, where $0 \leq h0(x) \geq 1$
$\beta 1 : slope$
$\beta 0 : y - intercept$
$X : independent\ variable$

Figure 3.1 is showing the logistic function:

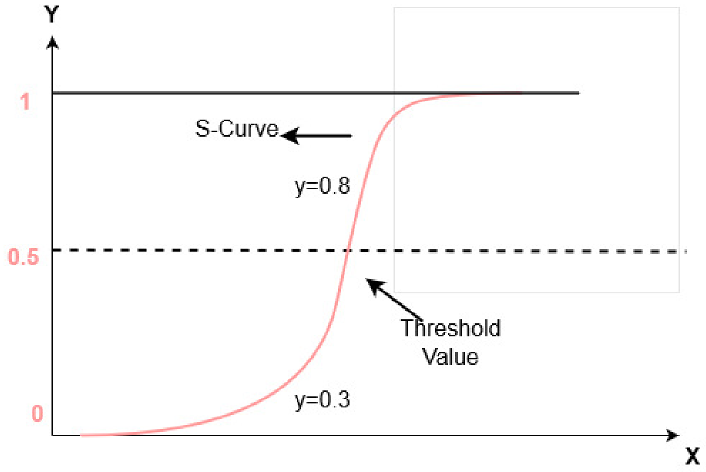


Figure 3.1: Logistic function

Based on categorical response, logistic regression models are divided into three types. These are **Binary logistic regression** which is dichotomous in nature ,i.e, it has only two possible outcomes, **Multinomial logistic regression** which has three or more possible outcomes and **Ordinal logistic regression** which also has three or more possible outcomes but in this case, these values do have a defined order.

3.2.2 K-Nearest Neighbor(KNN)

Based on supervised learning technique, K-NN is one of the simplest machine learning algorithm. K-NN algorithm assumes the similarity between the new case and available cases and put the new case into the category that is most similar to the available categories. It is used for both regression and classification problems but mostly for classification. It is also called a lazy learner algorithm because it does not learn from the training set immediately instead it stores the dataset and at the time of classification, it performs an action on the dataset. [21]

Suppose a new data point x shown in Figure 3.2, need to be assigned to either Category A or Category B. We can solve this problem using a K-NN algorithm. With the help of it, we can easily identify the category or class of a particular dataset. For this we need to select the number K of the neighbors. Then calculate the euclidean distance of K number of neighbors. Then, take the K nearest neighbors as per the calculated euclidean distance. Among these k neighbors, we need to count the number of the data points in each category and then assign the new data points to that category for which the number of the neighbors are maximum. Here, suppose the number of neighbors of the new data point in Category A is maximum. So, it will belong to Category A.

Figure 3.2 is showing the K-NN Classifier:

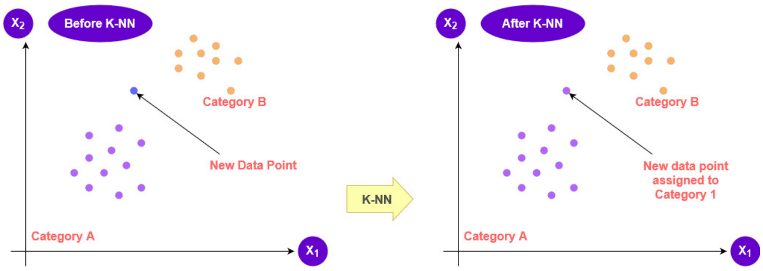


Figure 3.2: K-NN Classifier

3.2.3 Decision Tree

Though decision Tree is a supervised learning technique used mostly for classification, it can also be used for solving regression problems. It contains a decision Node and a leaf Node. Decision nodes make decisions whereas leaf nodes are the output of those decisions. It is a graphical representation for getting all the possible solutions to a problem based on given conditions. [22]

The main issue arises while implementing a Decision tree is how to select the best attribute for the root node and for sub-nodes. So, to solve such problems there is a technique known as Attribute selection measure or ASM and thus we can easily select the best attribute for the nodes of the tree. The two popular techniques for ASM are **Information Gain** and **Gini Index**.

Information Gain is the measurement of changes in entropy after the segmentation of a dataset based on an attribute. A decision tree algorithm always tries to maximize the value of information gain. It can be calculated by the following formula:

Information Gain= Entropy(S)- [(Weighted Avg) * Entropy(each feature)]

Entropy is a metric to measure the impurity in a given attribute and specifies randomness in data. It can be calculated as:

Entropy(s)= -P(yes)log2 P(yes)- P(no) log2 P(no)

Here,
S= total number of samples
P(yes)= probability of yes
P(no)= probability of no

Gini index measures the impurity or purity while creating a decision tree. If the value of Gini index is lower, than that attribute is preferred. It can be calculated using the below formula:

$$GiniIndex = 1 - \sum_{jPj^2}$$

Suppose, a candidate has a job offer. He/she has to decide whether he/she will accept the offer or Not. Figure 3.3 is showing the steps of the Decision Tree Classifier for this case.

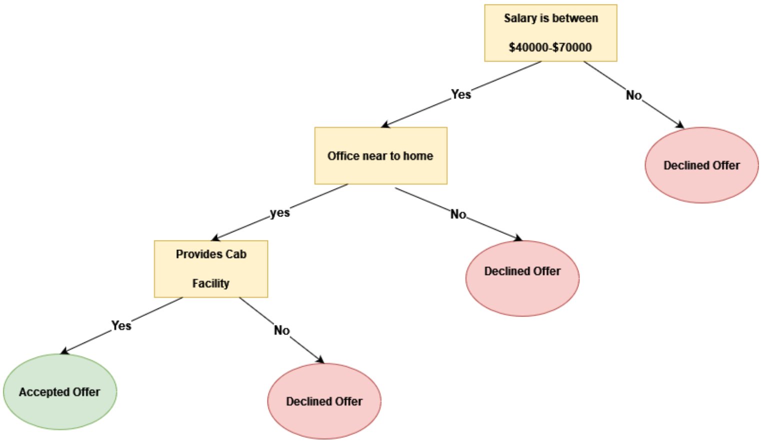


Figure 3.3: Decision Tree Classifier

3.2.4 Random Forest

Random Forest classifier chooses an average number of decision tree on various subsets of the given dataset to improve the result of the prediction. Higher accuracy can be obtained if there are greater number of decision trees in the forest. It also prevents the problem of overfitting. [23]

Suppose a dataset contains multiple fruit images as shown in Figure 3.4. We need to select random K data points from the training set and build the decision trees associated with the selected data points. Then, we need to choose the number N for decision trees that we want to build. During the training phase, each decision tree produces a prediction result. When a new data point occurs, based on the majority of results, it predicts the final decision.

Figure 3.4 is showing the Random Forest Classifier:

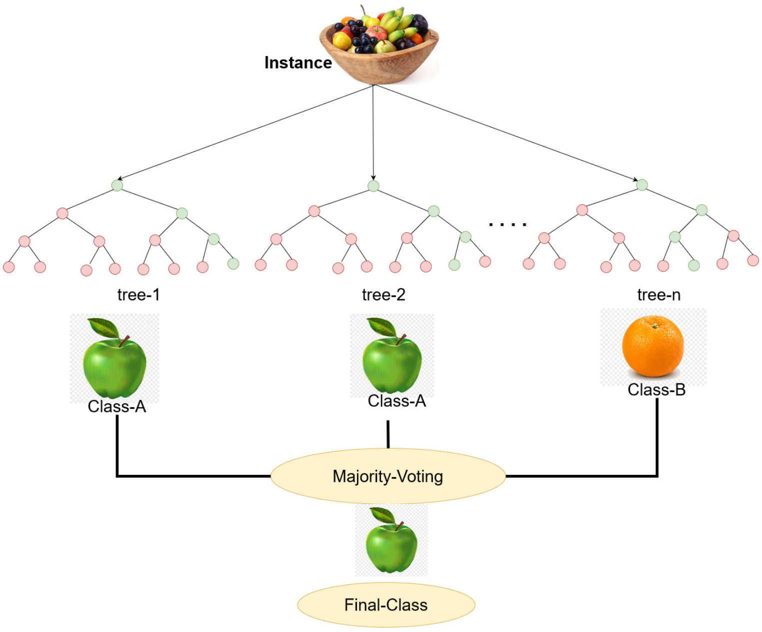


Figure 3.4: Random Forest Classifier

3.2.5 Multinomial Naive Bayes

Multinomial Naive Bayes algorithm is a probabilistic learning method.It is mostly used in text analysis, i.e, Natural Language Processing (NLP). [24] The concept of this algorithm came from the Bayes theorem [25].It predicts the tag of a text. Probability of each tag for a given sample is calculated by it. The tag with the highest probability as taken here as output.

As the Naive Bayes [26] concept came from the Bayes theorem, which was developed by Thomas Bayes, we need to know first about the Bayes theorem which is based on the formula given below:

$$\mathrm{P(A|B)} = \frac{P(A) * P(B|A)}{P(B)}.$$

This algorithm's prediction accuracy is lower than that of other probability algorithms and isn't appropriate for regression. It is used to classify textual input but cannot be used to estimate numerical values.

3.2.6 Support Vector Machine(SVM)

SVM is one of the most popular supervised machine learning algorithm. Though it is basically used for classification problems, it is also used for regression problems. [27]
SVM algorithm creates the best line or decision boundary known as hyperplane, that can separate n-dimensional space into classes. Thus we can easily put the new data point in the correct category in the future. Hyperplane dimension depends on the features present in the dataset. It means if there are 2 features (as shown in Figure 3.5), hyperplane will be a straight line and if there are 3 features, then hyperplane will be a 2-dimension plane. Hyperplane should have the maximum margin, which means that the distance between the data points should be maximum. SVM chooses the extreme points/vectors known as support vector, that creates the hyperplane and hence the algorithm is termed as Support Vector Machine.

Figure 3.5 is showing two different categories that are classified using a decision boundary or hyperplane:

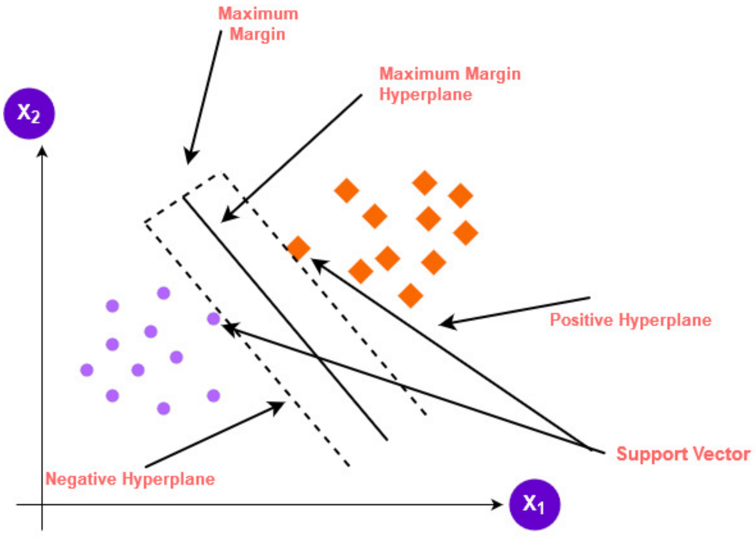


Figure 3.5: SVM Classifier

SVM can be categorized into two types.These are **Linear SVM** and **Non-linear SVM**. **Linear SVM** is used to classify linearly separable data. It means data can be classified by using a straight line. But if data are non linear, it cannot be separated by using a straight line. In that case, **Non-linear SVM** is used. In this case, by applying kernel functions, data are mapped into a high-dimensional feature space, in which the linear classification is possible.

3.3 TF-IDF

TF-IDF stands for term frequency-inverse document frequency.It is mainly used in text analysis and is very useful for scoring words in machine learning algorithms for Natural Language Processing (NLP). It measures how much a word is relevant to a document in a collection of documents. This is done by multiplying two metrics. TF is calculated by dividing how many times a word appears in a document by the total words of that particular document. IDF calcultes the logarithm of the result obtained by dividing the total number of documents by the word appears in how many documents. So, if the word is very common and appears in almost all the documents, the IDF result of that word will be zero. So, by multiplying TF with IDF, the ultimate result of TF-IDF of that word will also be zero. [28]

The formula of TF is given below:

$$\text{TF} = \frac{Number\ of\ times\ a\ word\ appears\ in\ a\ document}{Total\ number\ of\ words\ in\ that\ document}$$

The formula of IDF is given below:

$$\text{IDF} = \log(\frac{The\ total\ number\ of\ documents}{The\ word\ appears\ in\ how\ many\ documents})$$

Therefore, the final formula obtained by multiplying TF with IDF is given below:

TF-IDF = TF * IDF

3.4 Word Embeddings in NLP

Word Embedding is an approach for representing words in a lower-dimensional space. Word with similar meaning can have a similar representation in this case. A word vector with x values can represent x number of unique features. It reduces dimensionality and uses a word to predict similar kind of word.

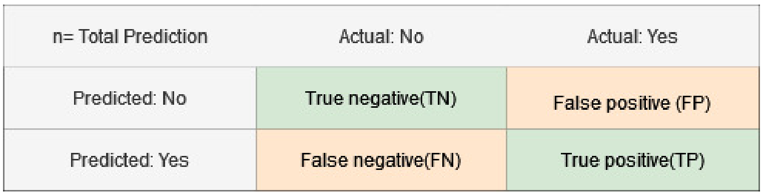
It is an feature extracting algorithm. It extracts linguistic features out of text. Then, we input those features into machine learning or deep learning models. It tries to preserve syntactical and semantic information.The other methods such as Bag of Words(BOW), CountVectorizer and TFIDF counts word in a sentence but they do not preserve any syntactical or semantic information. These algorithms creates the vector size equal to the number of elements in the vocabulary and we get a sparse matrix with most of the elements are zero. Large input

vectors results in high computation during the training. But word embedding gives a solution to these problems by reducing the input vector.Thus reduces computation.
So far, we have studied two different approaches to get Word Embeddings.These are **Word2Vec** and **GloVe**.
A vector is assigned to every word in **Word2Vec** and starts with a random vector or one-hot vector. **One-Hot vector** represents 1 bit in a vector as 1. There are two neural embedding methods in **Word2Vec** that after assigning vectors to each word take a window size and iterate through the entire corpus. These are **Continuous Bowl of Words(CBOW)** and **Skip Gram**. In **CBOW** model, tries to fit the neighboring words in the window to the central word and whereas **Skip Gram** tries to make the central word closer to the neighboring words which is the complete opposite of the **CBOW** model.
GloVe method take the corpus and iterate through it. It gets the co-occurrence of each word with other words in the corpus. A co-occurrence matrix can be found through this.
All modern applications of NLP uses Word Embedding technique. [29]

3.5 Confusion Matrix

Confusion matrix determines the performance of the classification models for a given set of test data. For predicting 2 classes the matrix uses 2*2 table, for predicting 3 classes it uses 3*3 and so on.It is consists of two dimensions. One is the predicted value and the other is the actual value among the total number of predictions. [30]

Figure 3.6 is showing the confusion matrix:



n= Total Prediction	Actual: No	Actual: Yes
Predicted: No	True negative(TN)	False positive (FP)
Predicted: Yes	False negative(FN)	True positive(TP)

Figure 3.6: Confusion Matrix

In a confusion matrix, **True Negative** indicates that the model has predicted no and the actual class was also no. **True Positive** indicates that the model has predicted yes and the actual class was also yes. **False Negative** indicates that the model has predicted no but the actual class was yes. **False Positive** indicates that the model has predicted yes but the actual class was no.

The calculations that are done using a confusion matrix are stated below:

3.5.1 Accuracy

It is the ratio of the total number of correct predictions to the total number of predictions by the classifier/model.

The formula of accuracy is :

$$\text{Accuracy} = \frac{TP + TN}{TP + FP + FN + TN}$$

3.5.2 Precision

It is the ratio of the number of positive classes correctly predicted by the model to the total number of positive classes predicted by the model.

The formula of precision is :

$$\text{Precision} = \frac{TP}{TP + FP}$$

3.5.3 Recall

It is the ratio of the number of positive classes correctly predicted by the model to the total number of actual positive classes.

The formula of recall is :

$$\text{Recall} = \frac{TP}{TP + FN}$$

3.5.4 F-measure

It is the harmonic mean of precision and recall. It has the maximum value, if precision is equal to recall.

The formula of F-measure is :

F-measure $= \dfrac{2 * Precision * Recall}{Precision + Recall}$

3.6 Artificial Neural Network(ANN)

ANN is similar to human brain. [31] Human brain consists of neurons which are interconnected by synapses.ANN consists of input and output layer that are build up with many neurons and weighted synapses.These two layers are connected with each other with the help of hidden layer neurons and weighted synapses.The input values are passed to the hidden layer(can be multiple) to process the result and finally send to the output layer to get the final result.

3.6.1 Necessity of Neural Network

It is used to solve both Classification and Regression problems. In our research we are using Neural Network in classification purpose as it gives us more better performance than that of the Machine Learning models.

3.6.2 Classification using Neural Network

If a problem has n classes, then the output layer of the neural network will have n number of neurons. That is in case of binary classification,the output layer will have 2 neurons and so on. The activation function that is used in the output layer is known as the Softmax activation function. [32] The formula of softmax function is defined below:

$$\sigma(z)_i = \frac{\exp(z_i)}{\sum_{j=1}^{K} \exp(z_j)}$$

Here,
$\sigma = softmax\, function$
$z = input\, vector$
$exp(z_i) = standard\, exponential\, function\, for\, input\, vector$
$K = number\, of\, classes\, in\, the\, multiclass\, classifier$
$exp(z_j) = standard\, exponential\, function\, for\, output\, vector$

3.6.3 Necessary Terms for Neural Network

In this section,we will discuss some of the necessary terms of the neural network.

Batch

Batch divides the dataset equally into some subsets over which the neural network computes gradient and updates weight.

Iterations

Iteration is the number of times a batch of data has passed through the algorithm.

Epoch

One forward and one backward propagation together make one epoch. It is the number of times an algorithm goes through the whole dataset.

Forward Propagation

Forward Propagation is a uni-directional process that propagates in the forward direction. Here at first , the input layer passes its data to the 1st layer of the neural network. The neuron present in that layer processes the data and updates the weight. This process continues. Then ultimately the processing data is send to the output layer for final update of the weight. It is the calculation and storage of intermediate variables (including outputs) for a neural network in order from the input layer to the output layer. [33]

Backward Propagation

Backward Propagation is also a uni-directional process that propagates in the backward direction.It is calculated to minimize the loss occurred during the forward propagation. It adjusts the values of weights and biases by calculating the gradient of cost function so that the cost function is minimized. [34]

Activation Function

The output of a node given an input or set of inputs is defined by the activation function of that node. [35] It calculates the weighted sum of the inputs and adds bias to it. It then decides whether a neuron should be fired or not. They can be linear or non-linear. Some of the activation functions include Linear activation [36],ReLU activation [37],tanh activation [38],sigmoid activation [39],softmax activation [32],etc. ReLU activation function is popularly used in the hidden layer. Output layer contains the softmax or sigmoid activation function.

Optimizer

Algorithms that work towards optimizing different attributes of neural networks, such as weights, are called optimizers [40] . The goal of these algorithms is to optimize weights in such a way that the cost function is minimized. That is, reducing the overall loss and improving the accuracy. Some of the optimizers are Adam [41],Adargrad [42], SGD [43],etc.

3.6.4 Long Short Term Memory (LSTM)

LSTM is an artificial neural network used in deep learning applications, such as natural language processing(NLP). LSTM architecture was motivated by an analysis of error flow in existing RNNs(Recurrent Neural Network) [44]. Results after the analysis found that long time lags were inaccessible to existing architectures, because backpropagated error either blows up or decays exponentially. [45] LSTM is a special kind of RNN and it has the capability of handling long-term dependencies by handling the vanishing gradient problem that occurred in RNN. RNNs works similarly like LSTM. Though they remember the previous information and use it for processing the current input, they can not remember Long term dependencies due to vanishing gradient. To avoid long-term dependency problems, the concept of LSTM has arised and are designed.

Figure 3.7 shows the internal structure of the LSTM network.

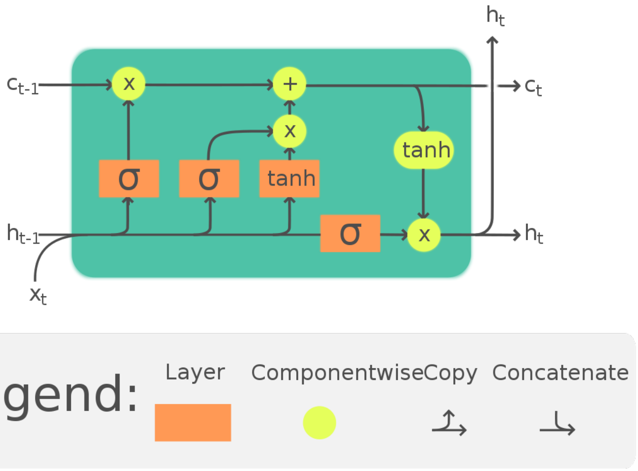


Figure 3.7: Internal structure of LSTM

LSTM contists of three parts,which are known as **the Forget gate**,**the Input gate** and **the Output gate**.

Forget Gate is the first part which decides whether we should keep the information that is coming from the previous timestamp or we are to forget it. The eqaution of the forget gate is given below:

$$\mathrm{f}_t = \sigma(X_t * U_f + H_{t-1} * W_f)$$

Here,
X_t : *input to the current timestamp*
U_f : *weight associated with the input*
H_{t-1} : *hidden state of the previous timestamp*
W_f : *weight matrix associated with hidden state*

ft is multiplied with a sigmoid function [39] that make the range of ft between 0 and 1. Then after that, ft is multiplied with the cell state of the previous timestamp(C_{t-1}).

The product will be zero if the value of ft is 0.Then the forget gate will forget everything. The product will be equal to the value of the cell state of the previous timestamp if the value of ft is 1. Then the forget gate will remember everthing.

$C_{t-1} * f_t = 0, if f_t = 0$
$C_{t-1} * f_t = C_{t-1}, if f_t = 1$

Input Gate is the second part in which the cell tries to learn new information from the input to this cell. The equation of the input gate is given below:

$i_t = \sigma(X_t * U_i + H_{t-1} * W_i)$

Here,
$X_t : current\, timestamp\, input$
$U_i : weight\, matrix\, of\, input$
$H_{t-1} : hidden\, state\, at\, the\, previous\, timestamp$
$W_i : Weight\, matrix\, of\, input\, that\, is\, associated\, with\, the\, hidden\, state$

Also in this step sigmoid function is applied to keep the value between 0 to 1.

Here, in this case the formula of new information is given below :

$N_t = tanh(X_t * U_c + Ht_1 * W_c)$

This new information is needed to be passed to the cell state.It is a function of a hidden state at the previous timestamp t-1 and input x at timestamp t. Tanh activation function [38] is used here and it will bring the value of new information between -1 and 1. The information is subtracted from the cell state, if the value is of Nt is negative and in the other case it is added to the cell state at the current timestamp.

Then we can get the updated equation which is as follows :

$C_t = f_t * C_{t-1} + i_t * N_t$

Output Gate is the third part where the cell passes the updated information from the current timestamp to the next timestamp. The equation of the output gate is given below:

$o_t = \sigma(X_t * U_o + H_{t-1} * W_o)$

It also uses sigmoid function to keep the value between 0 to 1.After that,the current hidden state need to be calculated as shown below:

$\mathrm{H}_t = o_t * tanh(C_t)$

To calculate the output of the current timestamp, the SoftMax activation [32] is just applied on the hidden state Ht.

$\text{Output} = \text{SoftMax}(\mathrm{H}_t)$

Here, the output is the prediction of the LSTM model.

3.6.5 BERT

BERT stands for Bidirectional Encoder Representations from Transformers.It is a new language representation model. BERT is designed to pre-train deep bidirectional representations.It is done from unlabeled text by jointly conditioning on both left and right context in all layers. As a result, the pre-trained BERT model can be fine-tuned with just one additional output layer to create state-of-the-art models for a wide range of tasks.

3.7 Summary

In this chapter, we have discussed the knowledge we need to gather before starting our research work.We have discussed about Machine Learning and various ML models, Neural Network and various terms that are related to Neural Network, LSTM which is an ANN model, TF-IDF a feature extraction algorithm which is applied to extract features before applying the ML models in our research, Word embedding which is used in the feature extraction step before applying the LSTM model in our research , confusion matrix and performance metrices. So, after studying these terms we have started doing our research work.

Chapter 4

Dataset Description

4.1 Age Classification Dataset

4.1.1 Collection

This Blog Authorship Corpus dataset was collected from a Kaggle link [46]. The total dataset consisted of posts collected from blogger.com. This huge dataset contains 681,288 rows containing the seven columns: id, gender, age, topic, sign, date and text.

We have chosen 8000 rows from the Blog Authorship Corpus dataset and formed our age classification dataset by altering the columns as per our research need.

4.1.2 Description

A brief overview of our age classification dataset has been provided below in table 4.1:

Table 4.1: Table showing summary of Age Dataset

Column Number	Column Name	Data Type
1	age	int64
2	text	object
3	Age_divide	object
4	encoded_category	int64

Now the columns present in our age classification dataset, along with their description, are stated below:

1. age: The age column contained the original age of the user such as 12,28,etc.
2. text: This column contained the texts collected from user's post.
3. Age_divide: Age_divide was introduced to mark the users as Under Aged or Aged based on whether they are under eighteen or over eighteen.
4. encoded_category: This column basically is the label encoded version of the Age_divide category where Under Aged users were labelled as 1 and Aged users as 0.

An example of under 18 versus above 18 users' posts are provided below:

- **Under 18 user's post:** An underage user's post is shown below:
  "*Today was a good day... Starting at about 10:30 last night... hee hee. Rose Dillon and I moused Chantele. again and we ate lots and lots and lots of chocolate cake. ca-ake. In Newspaper we played 20 Questions... that was fun. I love Newspaper. I really do. you too and we practiced for mass today, and sang my lil' heart out. Singing makes me in a good mood... unless I'm in a crappy mood. .:giggle:. Yesterday Mr. Thomas asked me if I would do journalism practicum next year. That was cool.*"
- **Above 18 user's post:** An 18 plus user's post is shown below:
  "*and i didnt make it to my doctor's appt...the one i prolly wouldve tried to change treatment to bi-polar...i should probably reschedule that*"

4.1.3 Data Overview

Our age classification dataset was modeled to be able to classify user's age based on the posts the individual writes. That being so, the eight thousand rowed dataset was made, merging 4575 posts posted by eighteen plus users and 3425 posts posted by under eighteen users. That is, the dataset comprised of 57.20% and 42.80% posts from over eighteen and under eighteen users, respectively. Hence a brief overview of the percentage of above eighteen and under eighteen users posts present in our dataset may be analysed from the pie chart Figure 4.1 below:

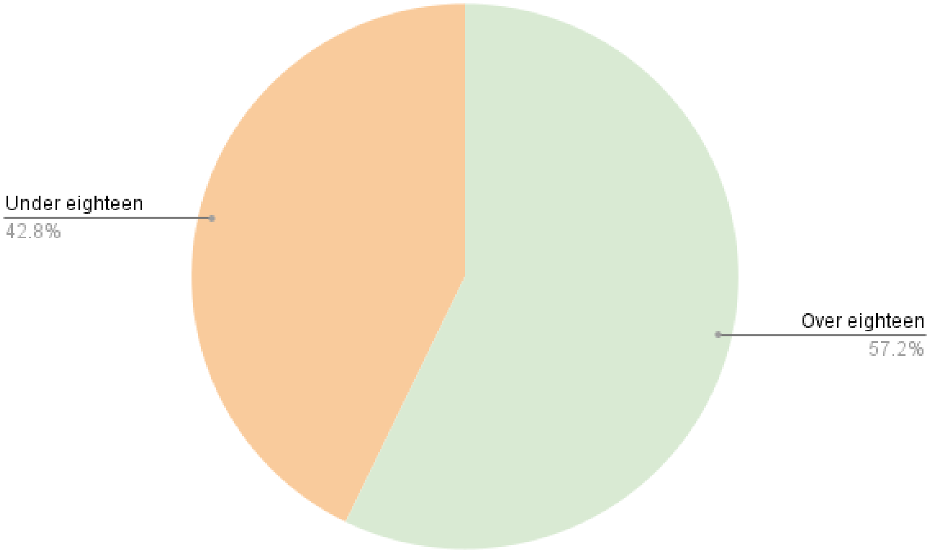


Figure 4.1: User Age Category

Since our research topic is on Natural Language Processing, a WordCloud will help to get a pen picture of our text data. Figure 4.2 shows a brief overview of under eighteen user's posts and Figure 4.3 reflects the overview of above eighteen user's posts:

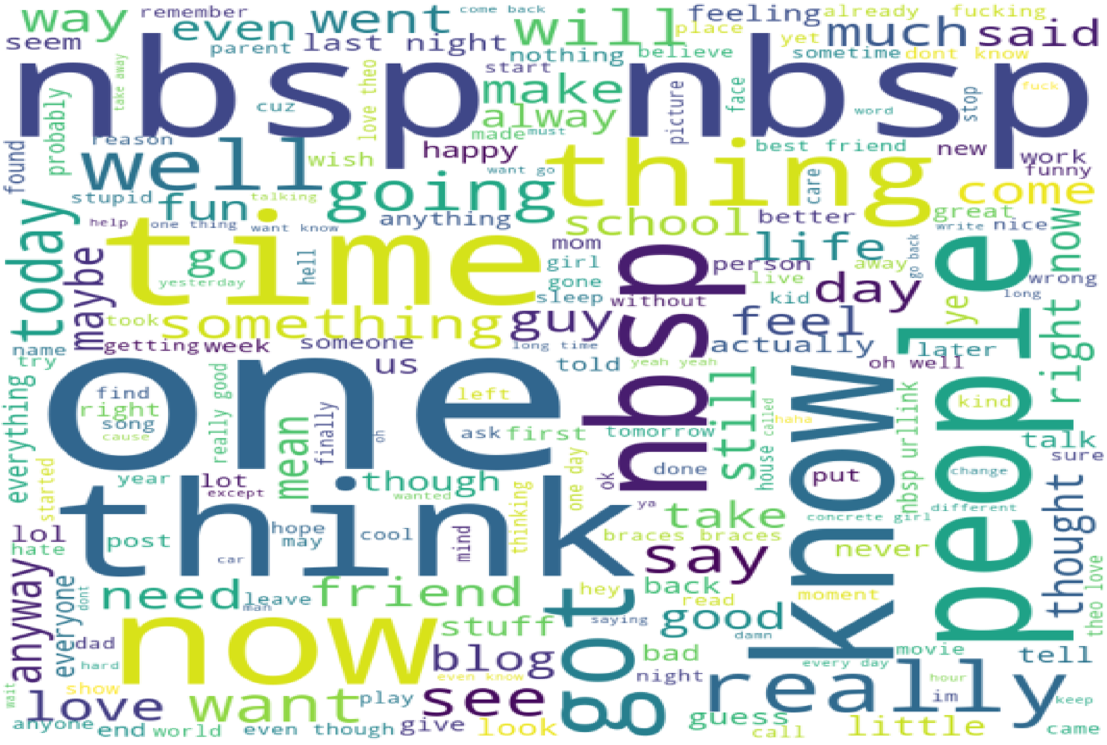


Figure 4.2: Under Eighteen Users posts

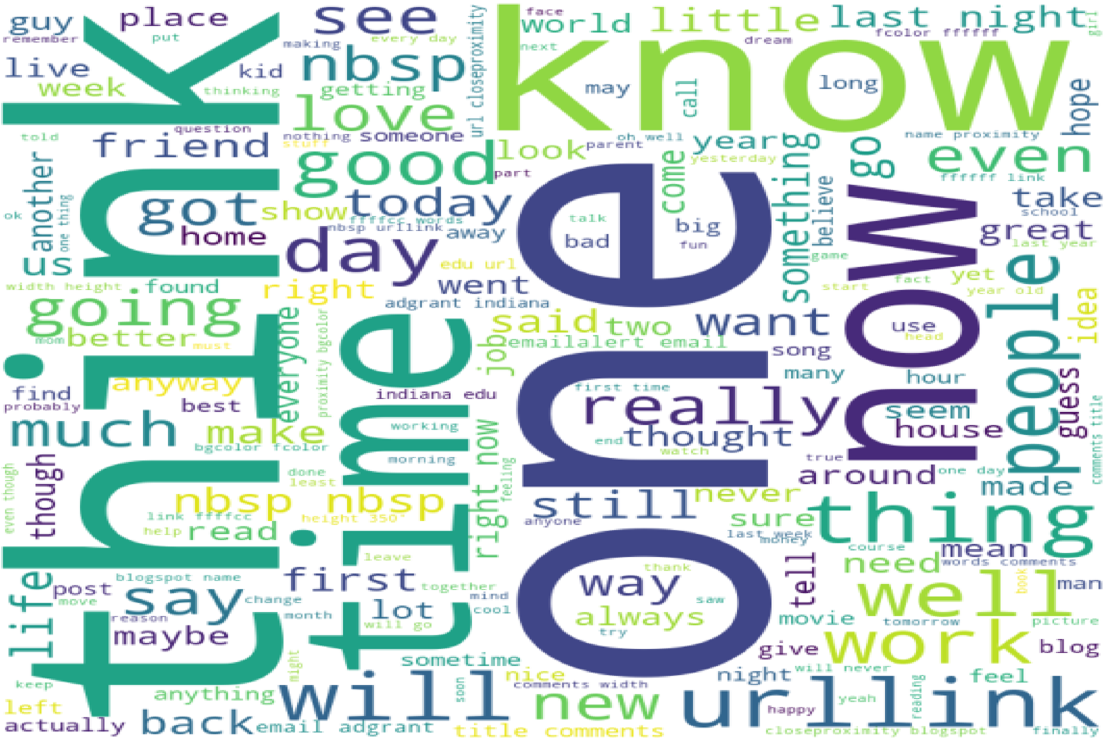


Figure 4.3: Above Eighteen Users posts

4.2 News Classification Dataset

4.2.1 Collection

The adult news were collected from a Kaggle dataset. [47] This dataset contained news articles from newspapers like New York Times, Breitbart, CNN, Business Insider, the Atlantic, Fox News, Buzzfeed News, New York Post, the Guardian, etc. The columns of the kaggle dataset were: unique id, news title, publication, author, date, year, month, url and content. The total dataset contained two lacs entries approximately.

For the kids news, we collected 15000 kids news by performing web scraping on five kids news websites which are:

1. News for kids: This US based news website was created by a teacher to create articles easily understandable by kids.
2. ROBINAGE: An Indian newspaper and news website specially designed for kids aged 4 to 15 years.
3. Science News for Students: Science News for Students is an online publication wholeheartedly dedicated to providing age-appropriate, science news.
4. DOGO News: This is an US based news website containing news for students from grade 1 to 9.
5. Time for Kids: Time for Kids, a division magazine of Time magazine is also an US based news website built for children.

Finally after data cleaning we picked out 12570 adult news and 12430 kids news. And hence we formed a complete news classification dataset containing 25000 rows. The whole collection process maybe briefly summarized by the Figure 4.4 below:

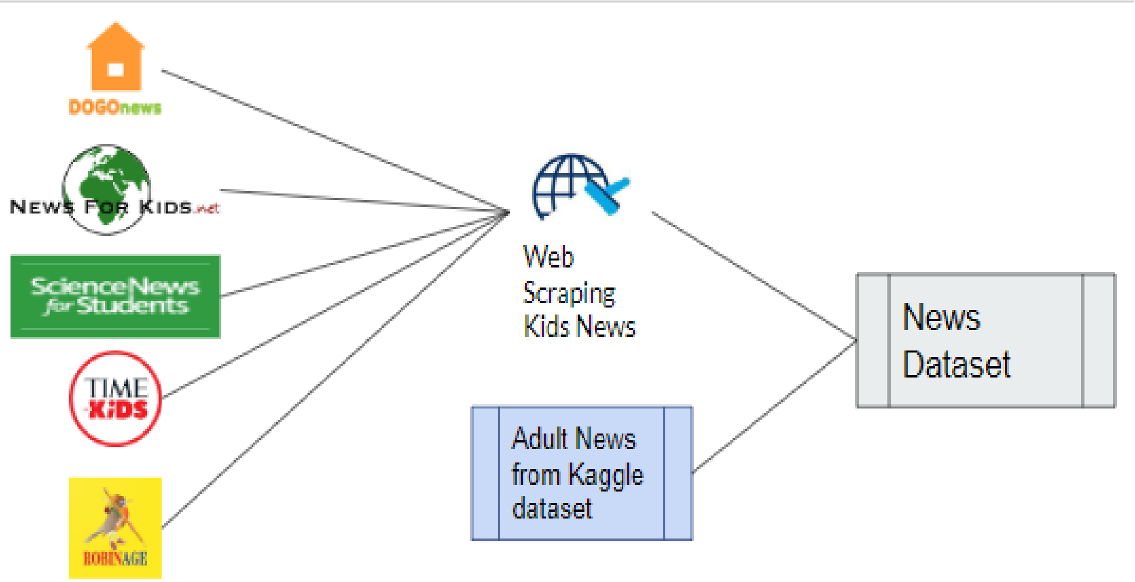


Figure 4.4: News Dataset Formation

4.2.2 Description

A brief overview of our news classification dataset has been provided below in table 4.2:

Table 4.2: Table showing summary of News Dataset

Column Number	Column Name	Data Type
1	name	object
2	url	object
3	title	object
4	article	object
5	kidFriendly	int64

As we can see from the table, the final dataset used for news classification contained a total of five columns. The column names along with their contents are:

1. name: The name column contained the Newspaper or Publication names.
2. url: This column contained the link to the news articles.
3. title: The title held the headlines of the news articles.
4. article: The textual contents of the news articles were kept in the article column.
5. kidFriendly: The last column, kidFriendly, was labelled as 1 for kids news articles and 0 for adult news articles.

A comparison of kids news and adult news from our news classification dataset is provided below:

- **Kids friendly news:** A kids friendly news is shown below:
  "*Ruth Bader Ginsburg died on Friday, September 18, in Washington, D.C., at the age of 87. She had served as a United States Supreme Court justice since 1993. Ginsburg was the second woman to be named to the nations top court. She will be remembered as a tireless fighter for gender equality. Memorials for Ginsburg were held at courthouses around the country. People left candles and flowers in memory of the late justice. They held signs thanking her for her service. Jennifer Berger joined a crowd that gathered Saturday night outside the Supreme Court building, in Washington, D.C. think it is important for us to recognize such a trailblazer, she told the AP. It is amazing to see how many people are feeling this loss tonight and saying goodbye. Read more about Ginsburg and the Supreme Court in the October 2, 2020, issue of TIME for Kids.*"

- **Adult or standard news:** An adult or standard news is shown below:
  "” ’Palantir, the extremely secretive Silicon Valley startup has had a rough year with lots of turnover and customers canceling their deals, .’ ’That article, based on confidential information and interviews with six past and current employees, provides our best look yet at the famously cagey Palantir.’ ’Here are its highlights:’ ’Palantir did not immediately respond to a request for comment from Business Insider. But Palantir representatives tell BuzzFeed, according to that report, that its business is strong and growing — and that, sometimes, .’ ’The company has raised $2. 42 billion in investment capital since its founding in 2004. Famed PayPal cofounder Peter Thiel and investor Joe Lonsdale are Palantir cofounders and remain involved in the company.’ ” ’Palantir, the extremely secretive Silicon. ..”"

4.2.3 Data Overview

Out of the 25000 news, the dataset contained 12570 adult news articles and 12430 kids news articles. That is 50.28% of the dataset consisted of adult news and the rest 49.72% kids news which is a represented in the following bar graph Figure 4.5.

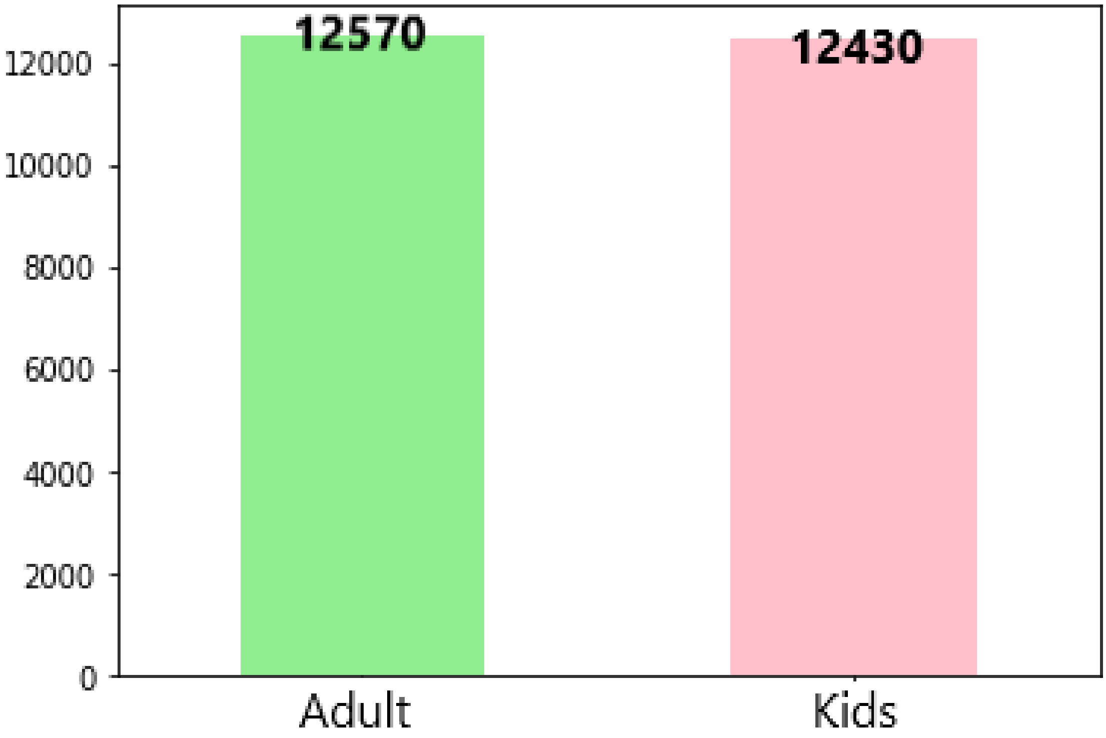


Figure 4.5: News Article Category

This equal distribution of the two categories ensures an unbiased and balanced dataset.

An overall review of the different Newspaper websites or online publications whose news articles are present in the dataset can be depicted in the Figure 4.6 below:

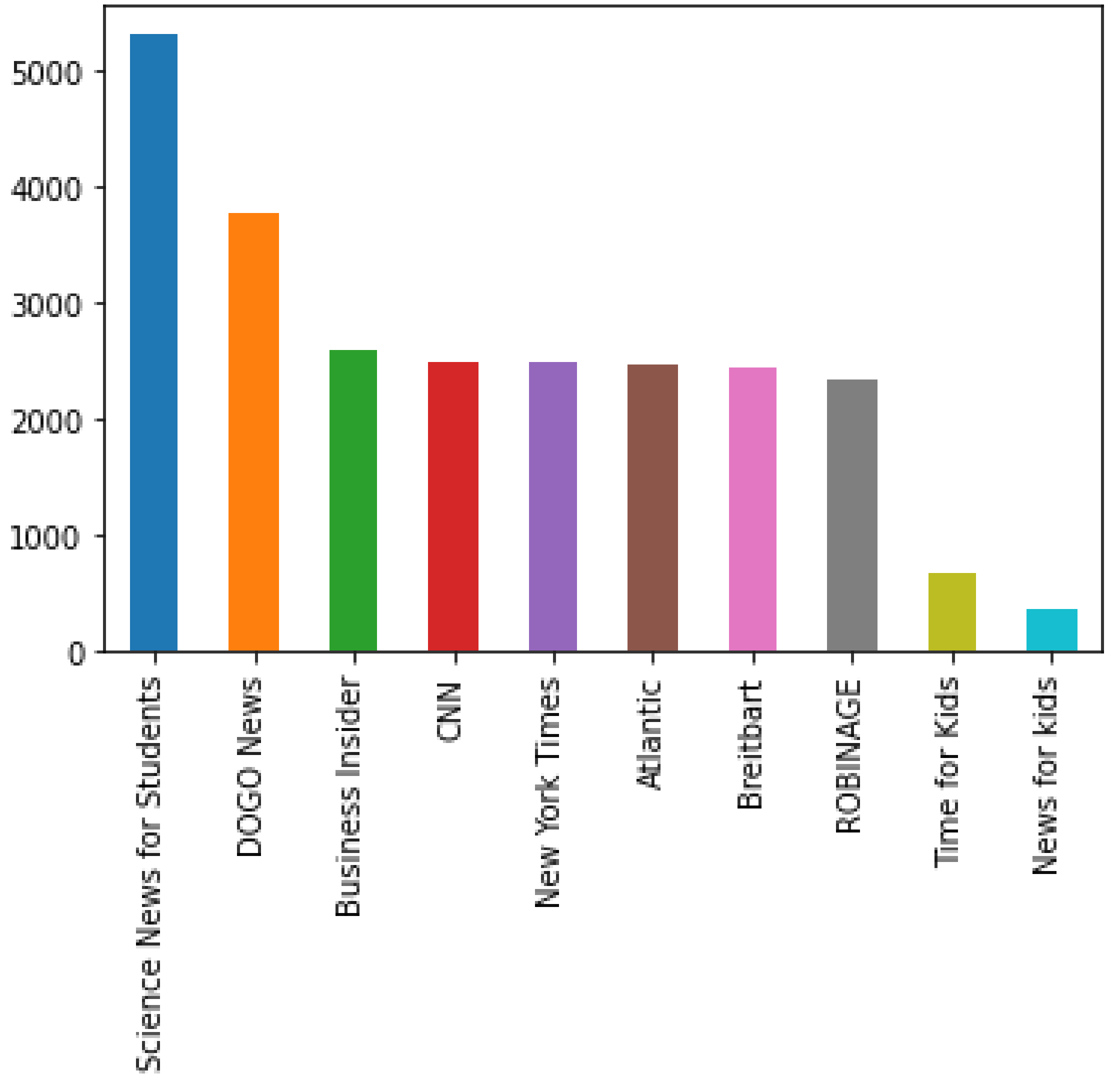


Figure 4.6: News article distribution based on Publishers

Chapter 5

Methodology

5.1 Overview

The posts of user's profile can be analyzed to detect user's age as eighteen plus or under eighteen and the incoming news posts can be categorized as kids or adults and displayed to users as per their age group. A framework of Social Media News Classification for underage user's have been displayed in Figure 5.1 the below:

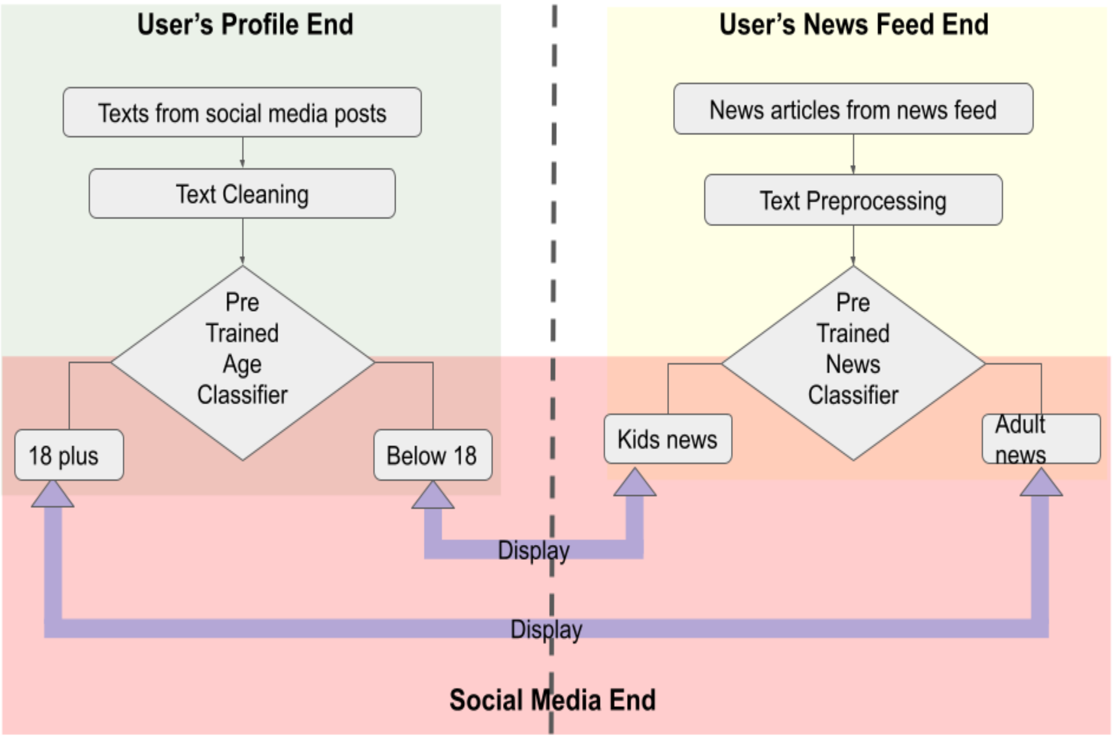


Figure 5.1: Methodology of Social Media News Classification for underage users

The three sectioned figure above gives a brief overview of our proposed methodology. The user's posts can be fetched from the database in the fragment named user's profile to classify the age and a scan of user's news feed database can be done to fetch the news related posts which can then be classified as kids or adults news in the user's news feed fragment. Finally the decision to display the news may be executed by the social media end fragment. Thus users will be able to view age appropriate news. This is the strategy we propose on our research to filter out social media news for underage users.

5.2 Pre-Trained Age Classifier

The pre-trained age classifier will be developed by following the below mentioned architecture in Figure 5.2

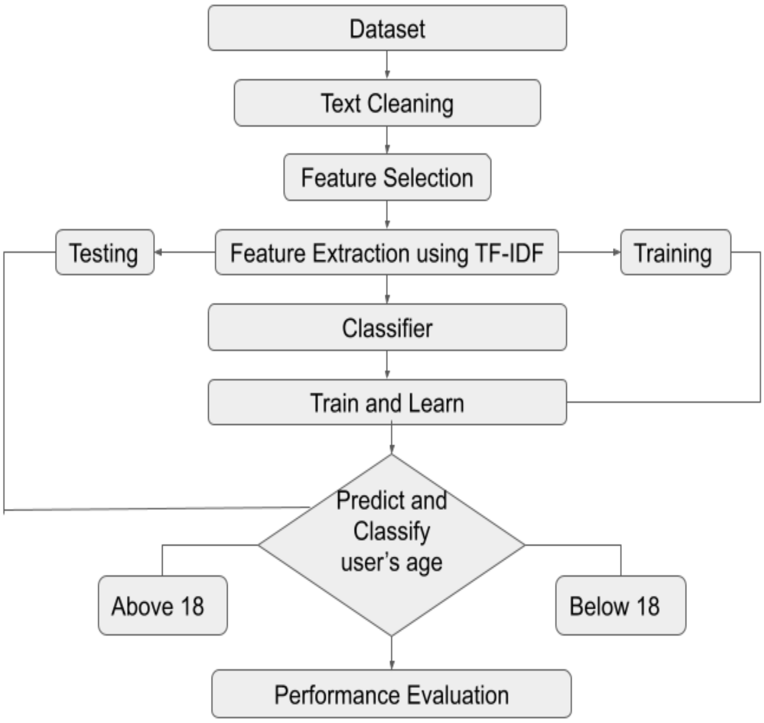


Figure 5.2: Work Flow of User Age Classification

At first dataset has to be collected .Then after cleaning and extracting linguistic features, feature selection has to be done. Then the dataset will be divided into training data and testing data. Training data are the data implemented to build up the age classification model and testing data are the data needed to validate the performance of the model. Then with the help of classifiers, the model will be able to predict and classify under eighteen and above eighteen users.

5.3 Pre-Trained News Classifier

The pre-trained news classifier will be developed by following the below mentioned architecture in Figure 5.3

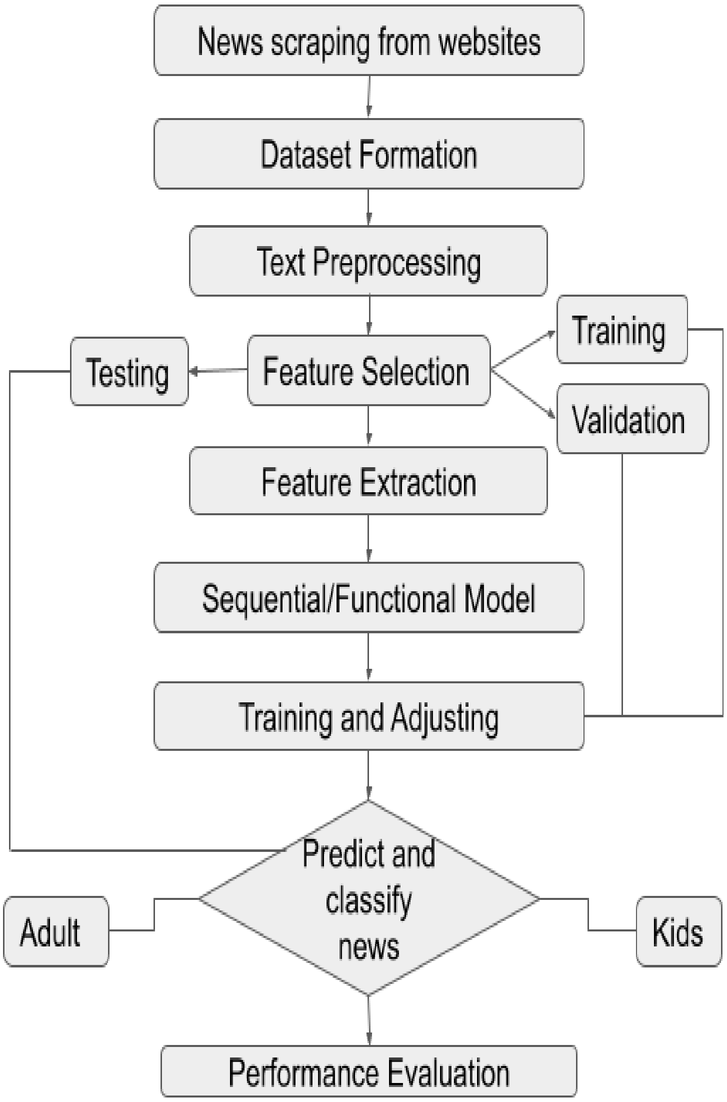


Figure 5.3: Work Flow of News Classification

Since no datasets for classifying news into Adult news and Kids news was available thus we decided on forming the dataset through web scraping and cleaning it. Then after text pre-processing, extraction of linguistic features, selecting features, preparing training and testing data, training sequential model, finally the test results will be predicted and the performance will be calculated. Here, the classification will give us two categories: adult news and kids news.

Chapter 6

Experiment and Results

6.1 Age Classification

6.1.1 Data Preparation

Only two columns: text and encoded_category was selected as features from our dataset. The encoded_category contained the labels and text column contained the posts. As the text was informal blogs or posts from users thus cleaning the data was a crucial step. This was performed by removing insignificant elements such as stopwords, punctuations, numbers, symbols, etc., and by turning all letters to lowercase. Finally the texts were extracted into vector forms using TF-IDF to make it ready for passing to the models. Eighty percent were kept for training and the rest Twenty percent for testing.

6.1.2 Models applied

The following classification models were applied to classify and predict our data.

- Logistic Regression [19]
- SVM (Support Vector Machine) [27]
- Random Forest [23]
- K-nn [21]
- Decision Tree [22]
- Naive Bayes [26]

6.1.3 Performance Metrics

The five performance metrics used to conclude the results were:

- Accuracy
- F1 score
- Precision
- Recall
- Area Under the Curve(AUC)

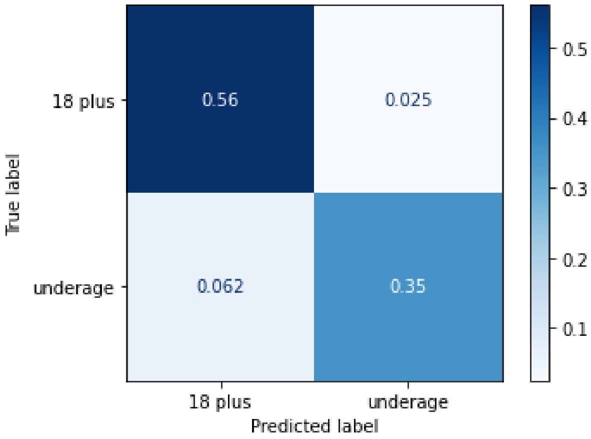
6.1.4 Performance Analysis

Logistic Regression, SVC, Multinomial Naive Bayes, Decision Tree, Random Forest and K-nn models were used to measure the results. To compare the results five performance metric scores were executed. These are Accuracy, F1 score, Precision, Recall and Area Under the Curve. The result comparison table is given below.

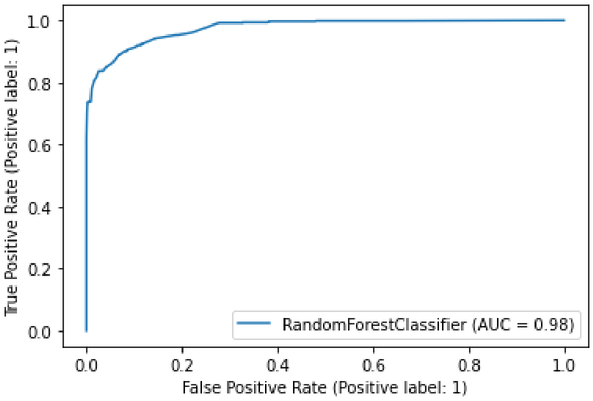
Table 6.1: Table showing result comparison of Age Classification Models

Models	Accuracy	F1 score	Precision	Recall	AUC
Random Forest	91.31%	89.02%	93.37%	85.06%	0.98
Decision Tree	91%	88.87%	91.12%	86.72%	0.92
Support Vector	89.625%	86.65%	92.77%	81.29%	0.95
K-nn	86.43%	81.96%	91.29%	74.36%	0.90
Naive Bayes	68.31%	47.67%	75.49%	34.84%	0.74
Logistic Regression	67.75%	57.21%	63.53%	52.03%	0.73

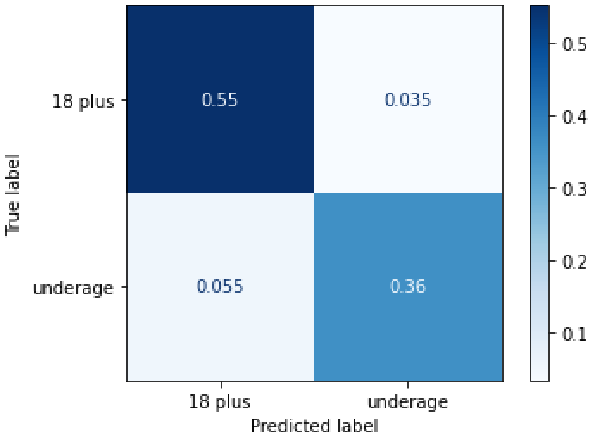
A brief overview of the classifier's performance may be evaluated from the confusion matrix and ROC curve demonstrated below:



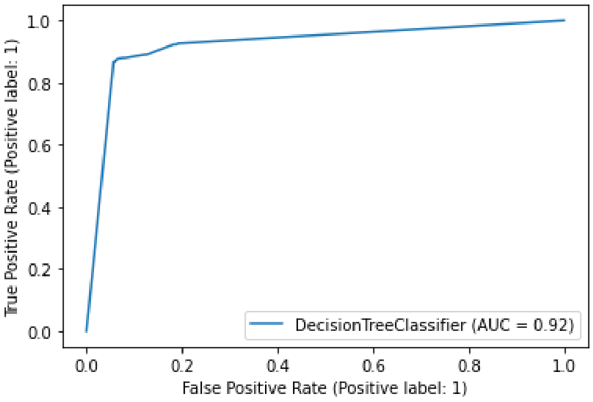
(a) Confusion Matrix of Random Forest



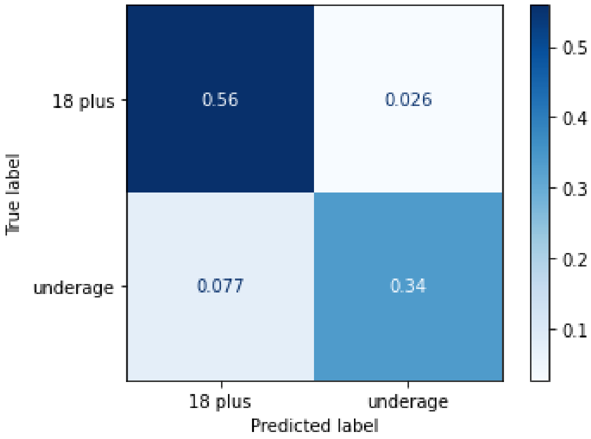
(b) ROC Curve of Random Forest



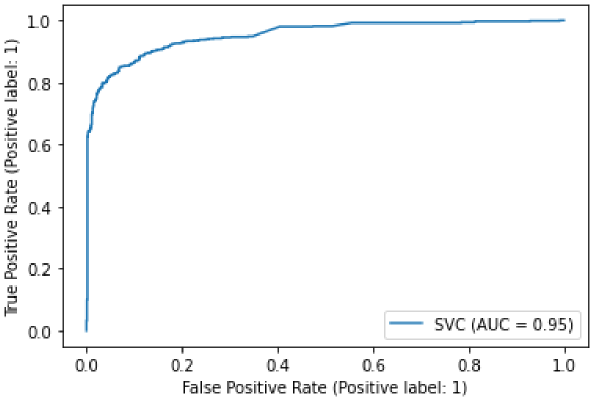
(a) Confusion Matrix of Decision Tree



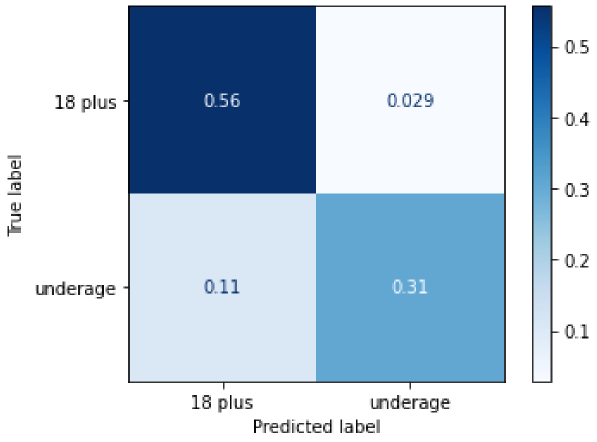
(b) ROC Curve of Decision Tree



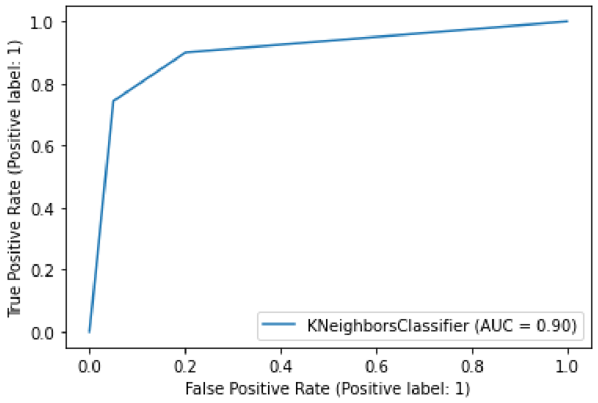
(a) Confusion Matrix of Support Vector Classifier



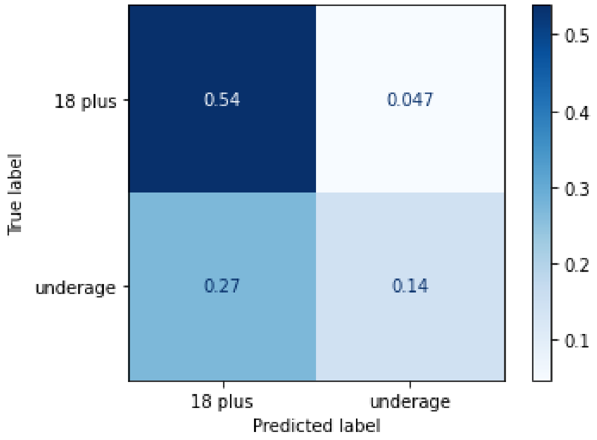
(b) ROC Curve of Support Vector Classifier



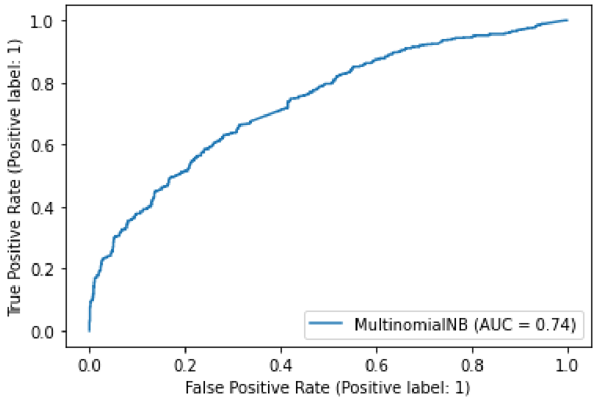
(a) Confusion Matrix of K-nn



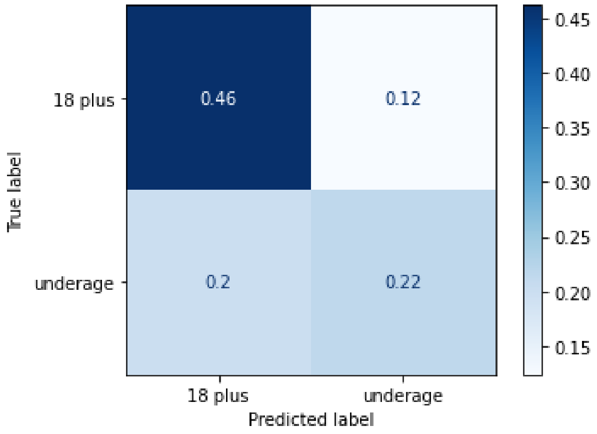
(b) ROC Curve of K-nn



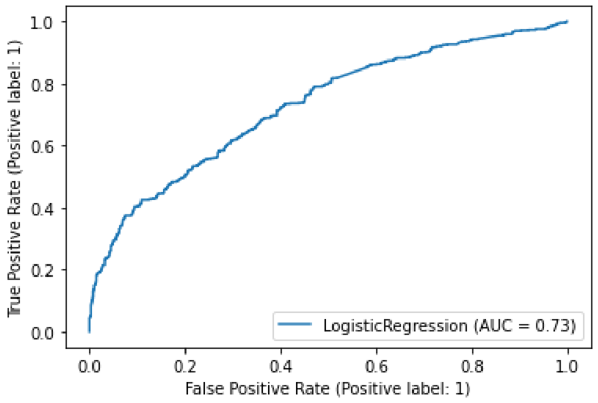
(a) Confusion Matrix of Multinomial Naive Bayes



(b) ROC Curve of Multinomial Naive Bayes



(a) Confusion Matrix of Logistic Regression



(b) ROC Curve of Logistic Regression

From the results we can see that Random Forest has gained the highest accuracy and that is 91.31%. The F1 score, Precision and AUC value of Random Forest is also the highest which is 89.02% and 93.37% and 0.9. The AUC value of Random Forest, Decision Tree, Support Vector and K-nn is above 0.90 which signifies that the category distinguishing ability of these models are quite good. In fact the AUC value of Random Forest is 0.98 that is very near to 1.00. However, Naive Bayes and Logistic Regression has showed very low performance in terms of all the metrics. Thus we cannot help but notice that all our non-linear models have performed exceptionally while the linear ones have lagged far behind from them. This reflects the non-linearity of our dataset. But it is well expected as we have worked with text data. Overall we could conclude that we are content with the results of age classifier so far.

6.2 News Classification

6.2.1 LSTM Sequential Model

Data Preparation

The following process was followed for data preparation.

- Cleaning : The text data was cleaned by removing stopwords, punctuations and digits to form a corpus to be later used by the model.
- Tokenization: The words were tokenized for passing to the model.
- Padding: Since we are working with text data, hence the input sequences turn out to be of different lengths based on article lengths. Thus all the texts were padded to ensure an even size of training variable for feeding the neural network model.
- Split: Among 25000 data, 80 percent was used to train the model. Another 10 percent was used to validate the model. Validation was mainly done to adjust our hyperparameters such as epoch, regularizers, callbacks, etc. And the rest 10 percent was utilized for testing purposes. It may be noted that other than the training dataset, both the validation and testing dataset was unseen to the model beforehand. An overview of the train, validation and test split maybe depicted from the fig Figure 6.7 below:

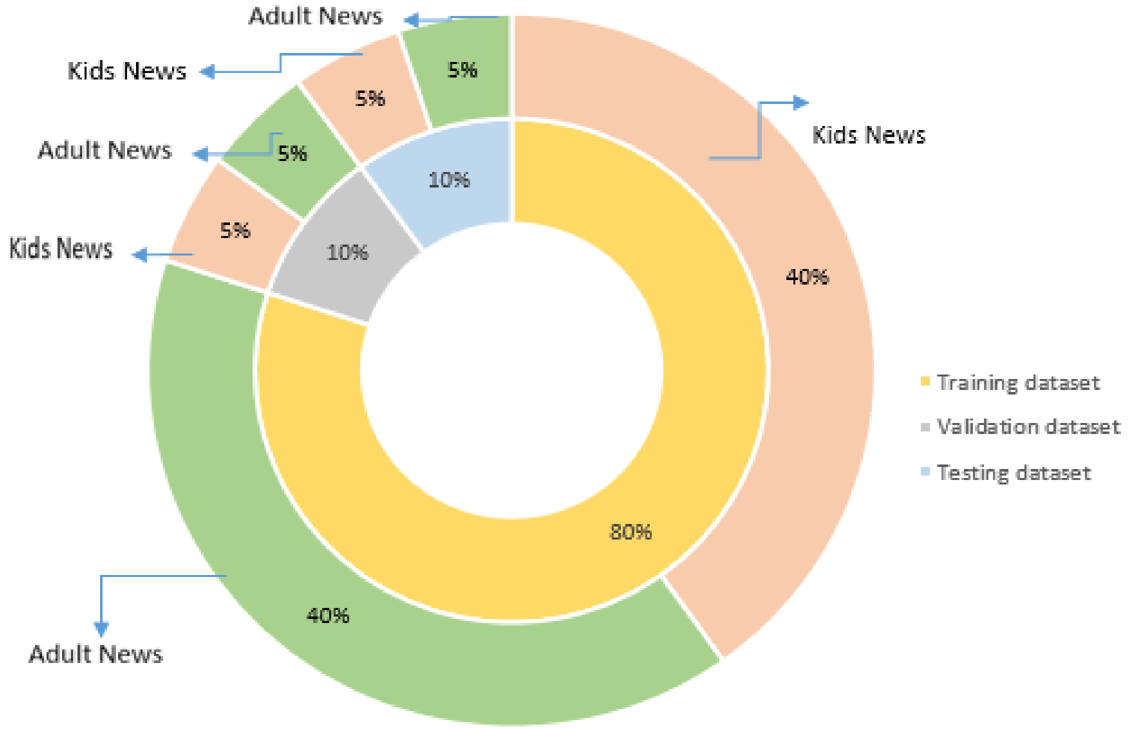


Figure 6.7: Train, Validation and Test Splitting

Sequential Model

The layers of our sequential model was developed as such:

- Embedding: An embedding layer was used to convert each words into a 16 bit vector.
- LSTM: An LSTM layer was added with unit size 16. Recurrent dropout was used to avoid overfitting and the return sequences was kept true to get output of all sequences and not only the last one.
- Flatten: This layer has been used to adjust the dimensional shape of the outputs of the LSTM layer.
- Dense: Lastly, the output is fed to the dense layer which will return the classification. For our activation function, we decided to stick to sigmoid. Also since our work is on binary classification, hence the unit of dense layer was set to 1.
- Optimizer: Adam optimizer with learning rate 3e-4 was used.
- Callback: A Callback function to monitor loss and restore weights accordingly was added.

A summary of input and output shapes, use of layers and the order of the layers our sequential model maybe observed below in fig Figure 6.8:

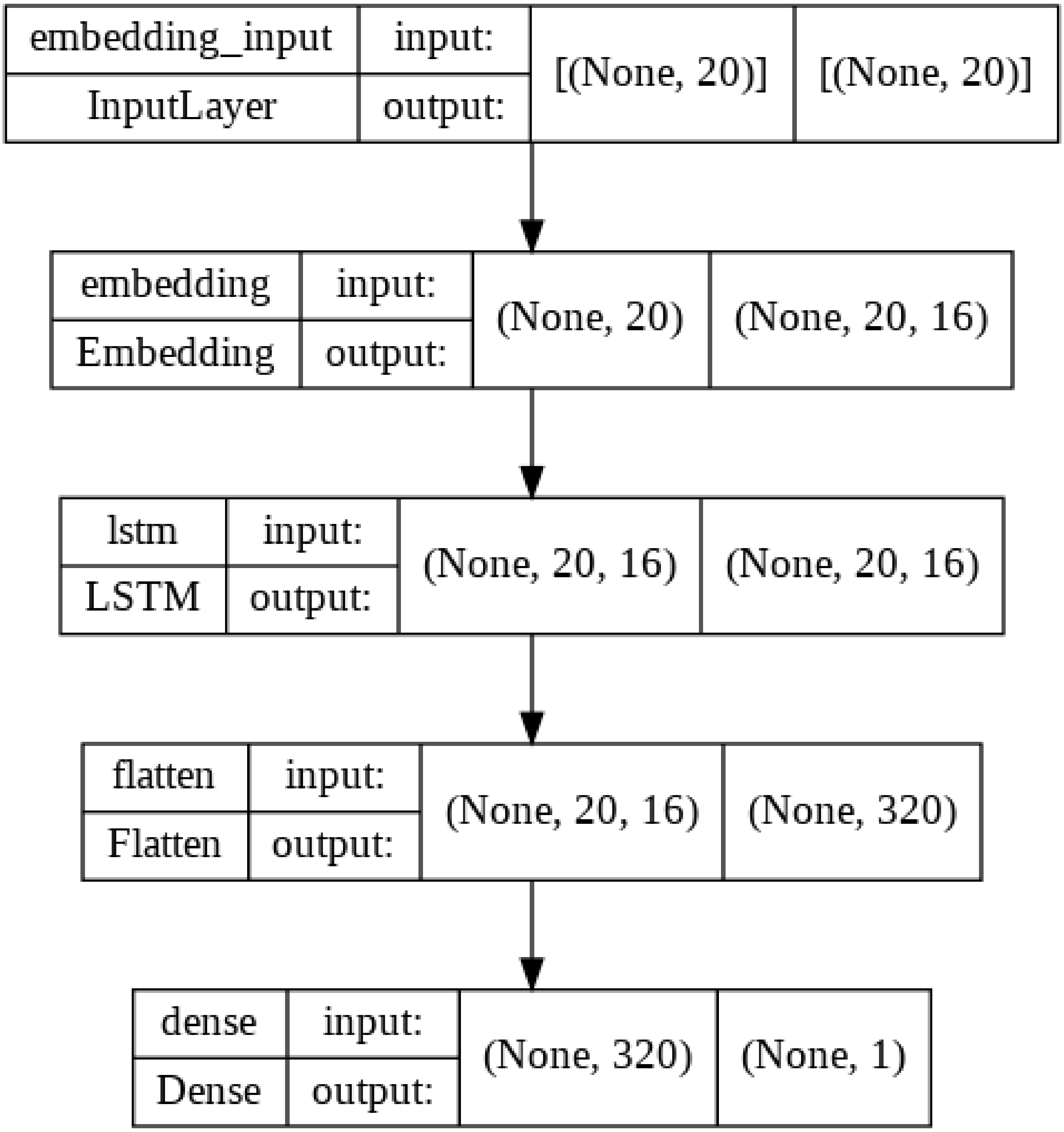
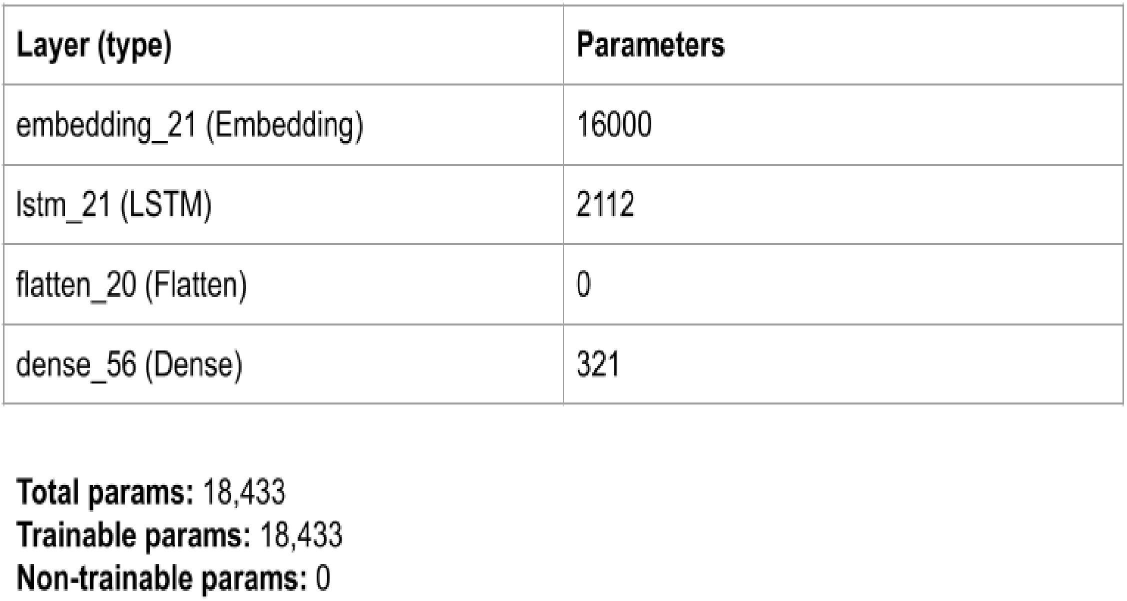


Figure 6.8: Sequential Model

The total number of training parameters of our sequential model is 18,433 which has been distributed throughout the layers. This maybe observed in figure below:



Layer (type)	**Parameters**
embedding_21 (Embedding)	16000
lstm_21 (LSTM)	2112
flatten_20 (Flatten)	0
dense_56 (Dense)	321

Total params: 18,433
Trainable params: 18,433
Non-trainable params: 0

Figure 6.9: Training parameters of the sequential Model

We note that all the parameters of the embedding, LSTM and Dense layer were trainable.

Training Performance

For the training the model, the previously allocated train dataset was used. And the validation dataset was used to get an idea of model's performance. The model was trained within 11 epochs and performance was monitored based on loss. The callback function for early stopping was used. The change in training accuracy, training loss, validation accuracy and validation loss may be elaborately noted from the table table 6.2 below:

Table 6.2: Table showing training and validation results

Epoch	Training Loss	Training Accuracy	Validation Loss	Validation Accuracy
1	0.3327	0.8911	0.1987	0.9364
2	0.1762	0.9468	0.1839	0.9412
3	0.1585	0.9530	0.1817	0.9424
4	0.1489	0.9575	0.1679	0.9460
5	0.1423	0.9585	0.1696	0.9488
6	0.1398	0.9585	0.1607	0.9524
7	0.1364	0.9593	0.1627	0.9492
8	0.1318	0.9614	0.1581	0.9488
9	0.1301	0.9614	0.1590	0.9500
10	0.1285	0.9617	0.1578	0.9452
11	0.1225	0.9633	0.1596	0.9484

Thus it may be concluded that as our model kept learning throughout it's whole training period thus the training performance of our model was good.

A comparison between training loss and validation loss throughout the training period may be observed in the plot of Figure 6.10 below:

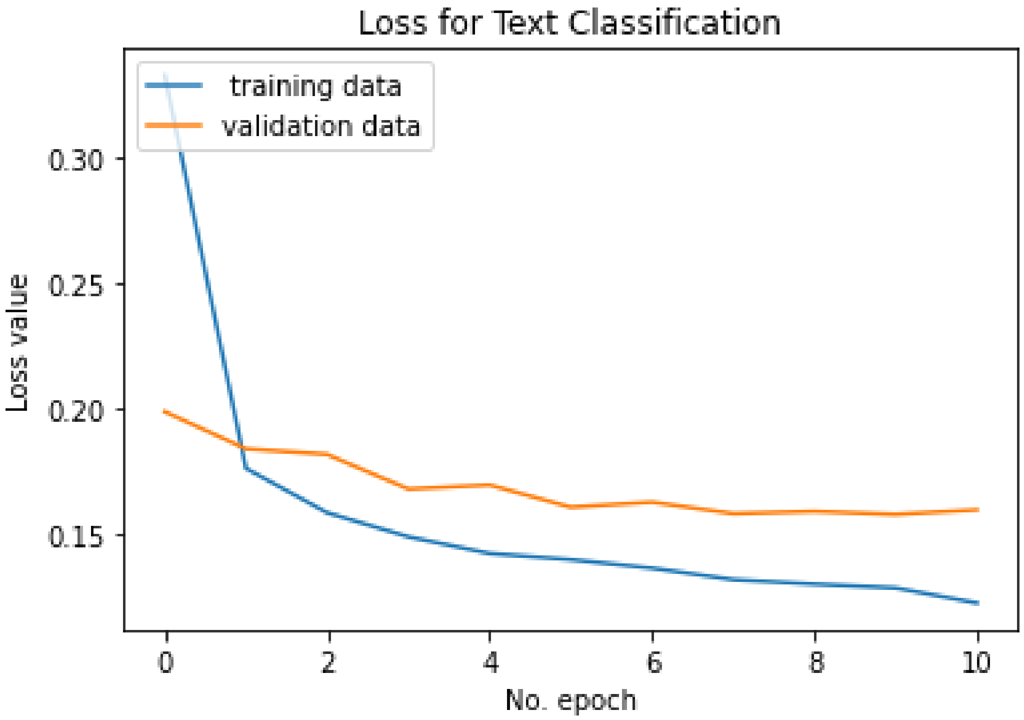


Figure 6.10: Training loss versus Validation loss

The training loss starts out higher than the validation loss in first epoch. But drops below the validation loss in the very next epoch. From then the training loss reduces to almost 0.1225. While the validation loss can also be seen decreasing throughout the training time. The lowest training loss is in the last epoch that is 0.1225 and the lowest validation loss 0.1578 was noted during the tenth epoch.

A plot describing the training accuracy versus validation accuracy may be observed in Figure 6.16 below:

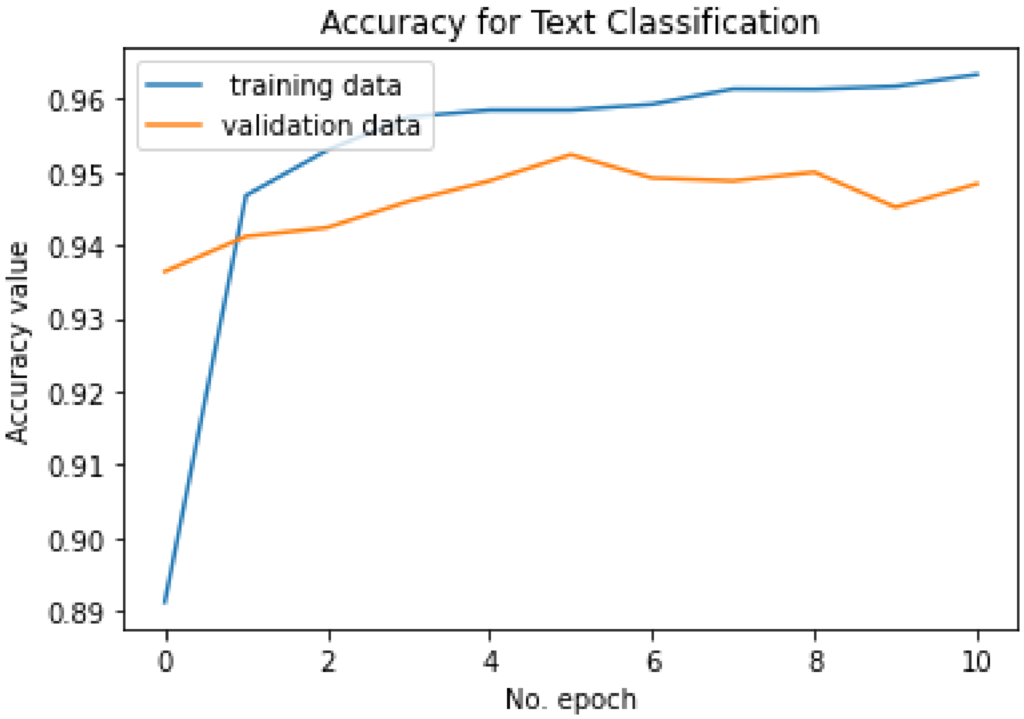


Figure 6.11: Training accuracy versus Validation accuracy

Similar to the scenario to the loss, we can observe that the validation accuracy is higher than training accuracy in first epoch. But then rises above validation accuracy in the exact next epoch and keeps increasing until the end. The validation accuracy also rises following the training accuracy. The highest training accuracy can be reported as 96.33% during the last epoch and the highest validation accurcay can be reported as 95.00% in the ninth epoch.

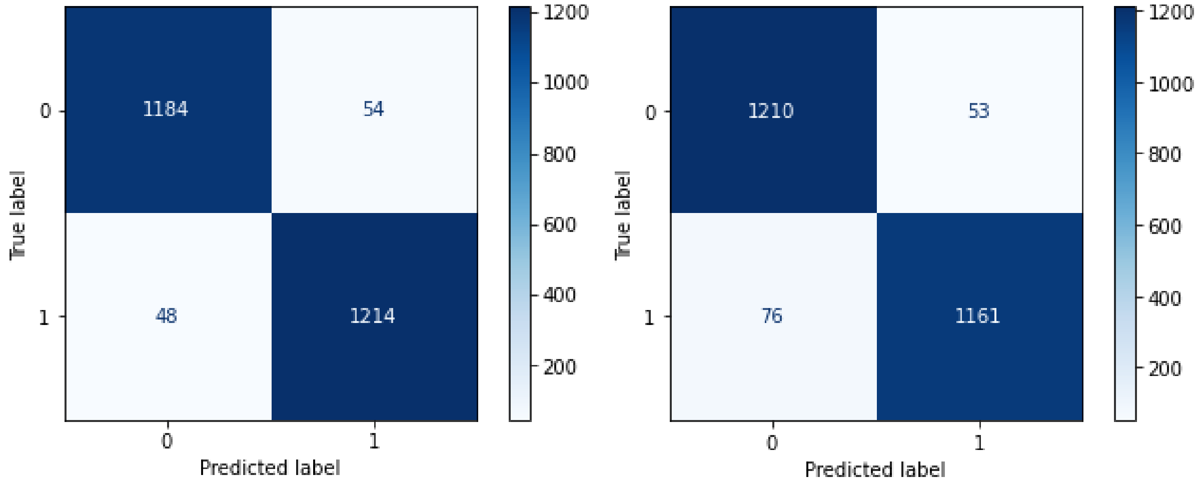
Evaluation and Results

The ten percent of dataset pre-defined for testing purpose was used to access the perfromance of the model. After evaluating the model using our train, test and validation data, we noted these metrics in table 6.7

Table 6.3: Table showing result comparison

Metric	Training	Testing	Validation
Accuracy	97.40%	94.84%	95.92%
Loss	0.1000	0.1596	0.1365
Steps	625	79	79
Time	3 secs	0 secs	0 secs

The confusion matrix of our validation and testing dataset is shown in the Figure 6.12a and Figure 6.12b below:



(a) Confusion Matrix of Validation Dataset

(b) Confusion Matrix of Test Dataset

Here the label class 0 denotes kids news category and 1 denotes the adult news category. We can note that the true positive of both the matrices are much more compared to the false values.

Finally precision and recall were calculated using sklearn's Classification Report. A brief overview of it on the validation dataset maybe concluded in the table table 6.4 below:

Table 6.4: Table showing result of Validation Dataset

Support	Precision	Recall	F1-Score
Kids	0.96	0.95	0.96
Adults	0.95	0.95	0.95
Total	0.95	0.95	0.95

And a brief overview of the results of testing dataset maybe concluded in the table table 6.5 below:

Table 6.5: Table showing result of Test Dataset

Support	Precision	Recall	F1-Score
Kids	0.96	0.96	0.96
Adults	0.96	0.96	0.96
Total	0.96	0.96	0.96

Thus it may be concluded that along with accuracy, precision and recall, F1-Score of our model was also good. Also we have achieved high accuracies on all the training, testing and validation dataset. But the highest accuracy and lowest loss both was noted on the training data. And a little less accurcay and little more loss on our validation and testing dataset. But that is understandable and well expected. Since the model was already familiar to the training dataset, it resulted in the highest accuracy. And as validation and testing dataset were unknown to the model hence the lower accuracies. This also indicates that our model was not overfitted. And the hyper-parameters were properly set which resulted in such a better performance.

6.2.2 Functional Model with BERT

Data Formatting

The following process was followed for data preparation.

- Pre-processing : The text data was pre-processed by a BERT preprocessor.
- Encoding: The data was then encoded using a BERT encoder for passing to the model.
- Split: Similar to LSTM sequential model, 80 percent was used to train, 10 percent to validate and the rest 10 percent for testing purposes.

Functional Model

Our functional model was developed as such:

- Input: The text were fed to the input layer.
- Keras Layer: Pre-processing with the help of BERT was performed in this layer.
- Keras Layer-1: The pre-processed outputs were encoded with BERT in this layer.
- Dropout: Dropout was added to avoid overfitting.
- Hidden-1: This layer consisted of a Dense layer of 128 neurons and a sigmoid activation function.
- Hidden-2: Another Dense layer with 32 neurons and a sigmoid activation function like the previous one was used.
- Output: Lastly, the output layer is a dense layer with 1 neuron which will return the classification. For our activation function, we decided to stick to sigmoid as it is evidently good for binary classification.
- Optimizer: Similar to the sequential one, Adam optimizer with learning rate 3e-4 was used here too.
- Callback: Callback function was used to monitor loss and initiate early stopping if necessary.

A summary of the functional model with BERT pre-processing maybe noted below in fig Figure 6.13:

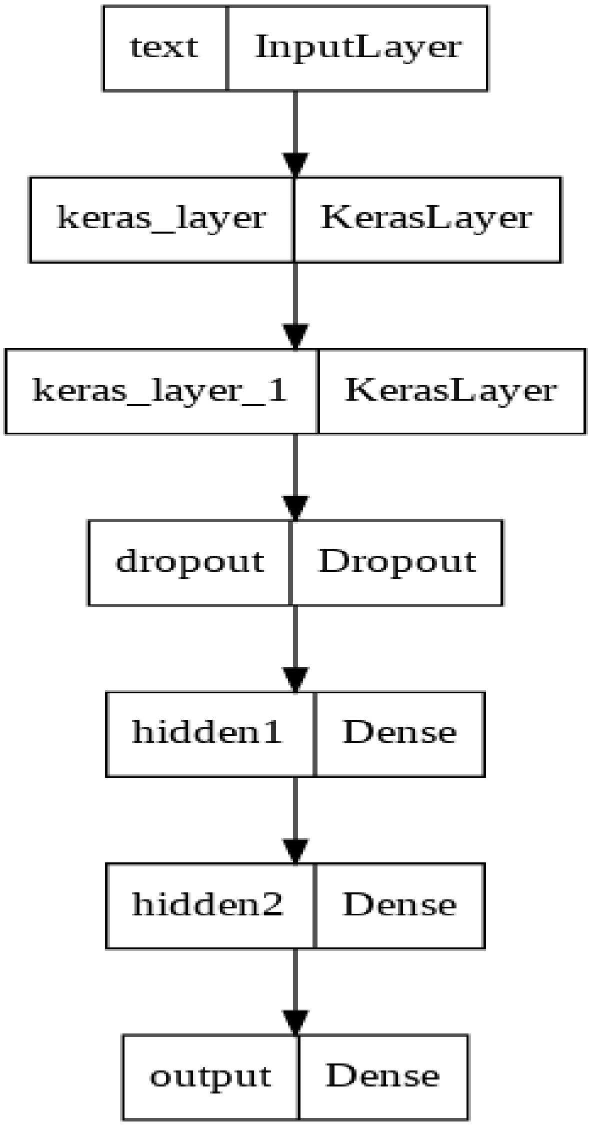
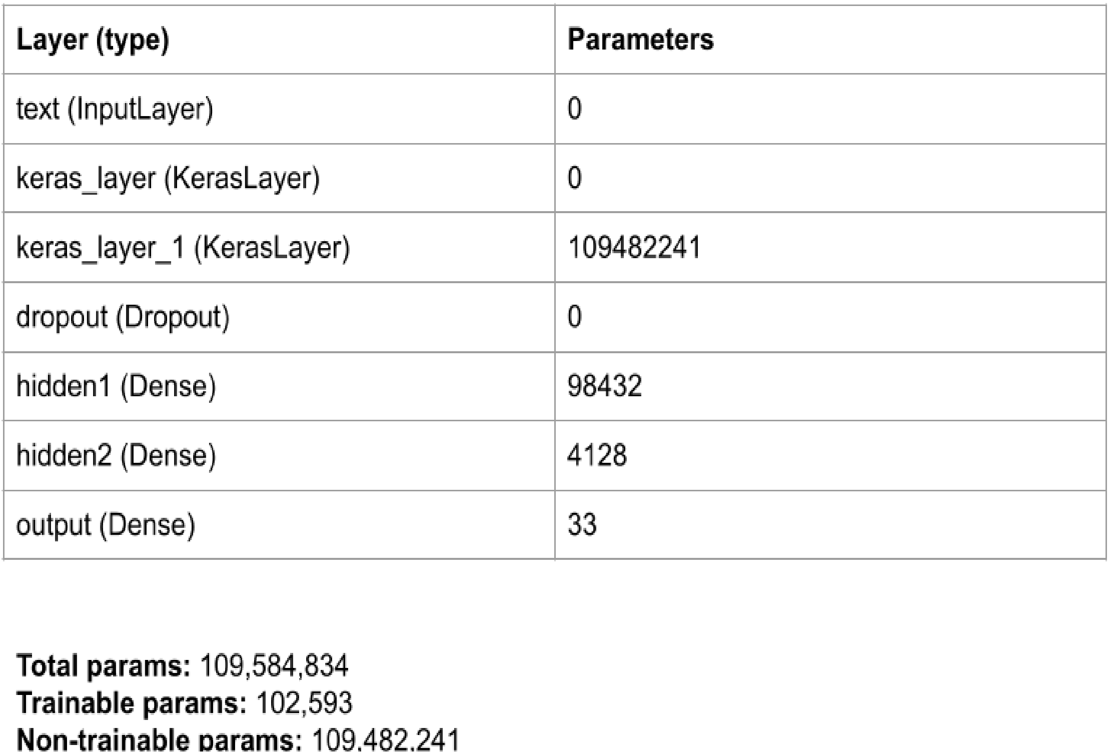


Figure 6.13: Functional Model

The parameters of the layers of our functional model may elaborately be concluded from the figure below:



Layer (type)	**Parameters**
text (InputLayer)	0
keras_layer (KerasLayer)	0
keras_layer_1 (KerasLayer)	109482241
dropout (Dropout)	0
hidden1 (Dense)	98432
hidden2 (Dense)	4128
output (Dense)	33

Total params: 109,584,834
Trainable params: 102,593
Non-trainable params: 109.482.241

Figure 6.14: Training parameters of the sequential Model

Out of total 109,584,834 parameters, only 102,593 were trainable. These are actually the parameters of the two hidden dense layers and the output dense layer. The rest 109,482,241 non trainable parameters are from the keras_layer_1. The BERT encoding was concluded in this layer. And since BERT is a pre-trained model, hence the parameters of this layer was not used in training the model.

Training Performance

For the training the functional model, the previously allocated train dataset was used. And the validation dataset was used to get an idea of model's performance. The model was trained within 11 epochs and performance was monitored based on loss. The callback function for early stopping was used. The change in training accuracy, training loss, validation accuracy and validation loss may be elaborately noted from the table table 6.6 below:

Table 6.6: Table showing training and validation results

Epoch	Training Loss	Training Accuracy	Validation Loss	Validation Accuracy
1	0.3612	0.8552	0.2116	0.9292
2	0.1836	0.9315	0.1751	0.9320
3	0.1619	0.9383	0.1478	0.9440
4	0.1490	0.9444	0.1420	0.9456
5	0.1438	0.9461	0.1495	0.9416
6	0.1367	0.9486	0.1299	0.9532
7	0.1383	0.9473	0.1367	0.9500
8	0.1314	0.9508	0.1229	0.9544
9	0.1305	0.9505	0.1291	0.9504
10	0.1282	0.9513	0.1208	0.9536
11	0.1252	0.9524	0.1214	0.9528

A comparison between training loss and validation loss of our functional model throughout the training period may be observed in the plot of Figure 6.15 below:

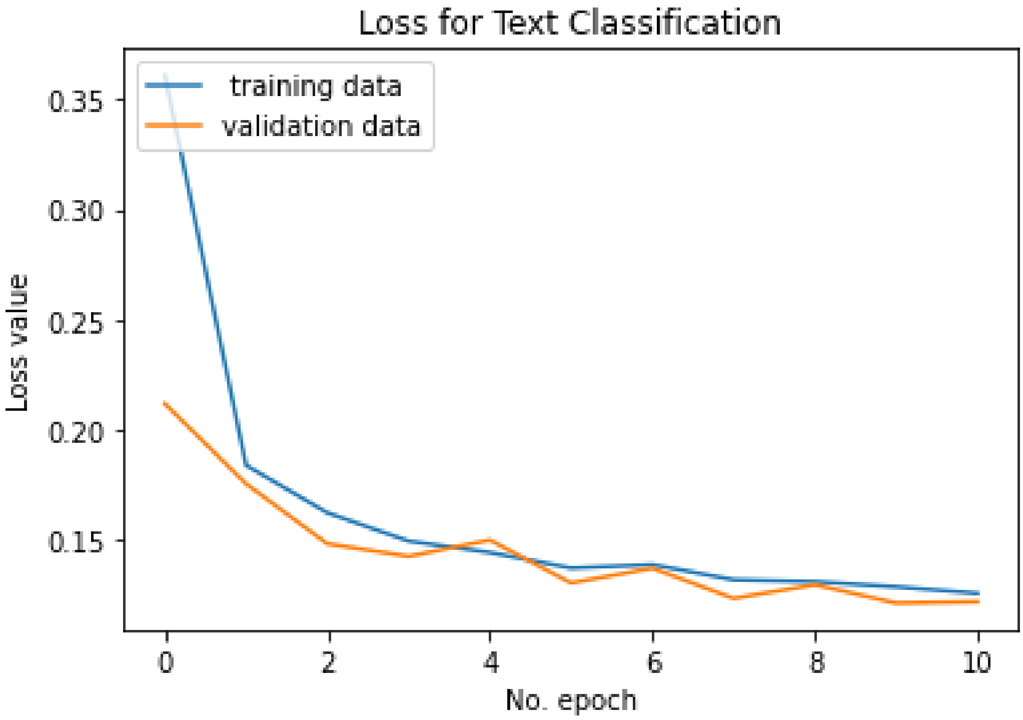


Figure 6.15: Training loss versus Validation loss

The training loss and validation loss maybe seen increasing, decreasing and intersecting each other throughout the time period. However, the trend of the graph seems to be in the descending order.

A plot describing the training accuracy versus validation accuracy may be observed in Figure 6.16 below:

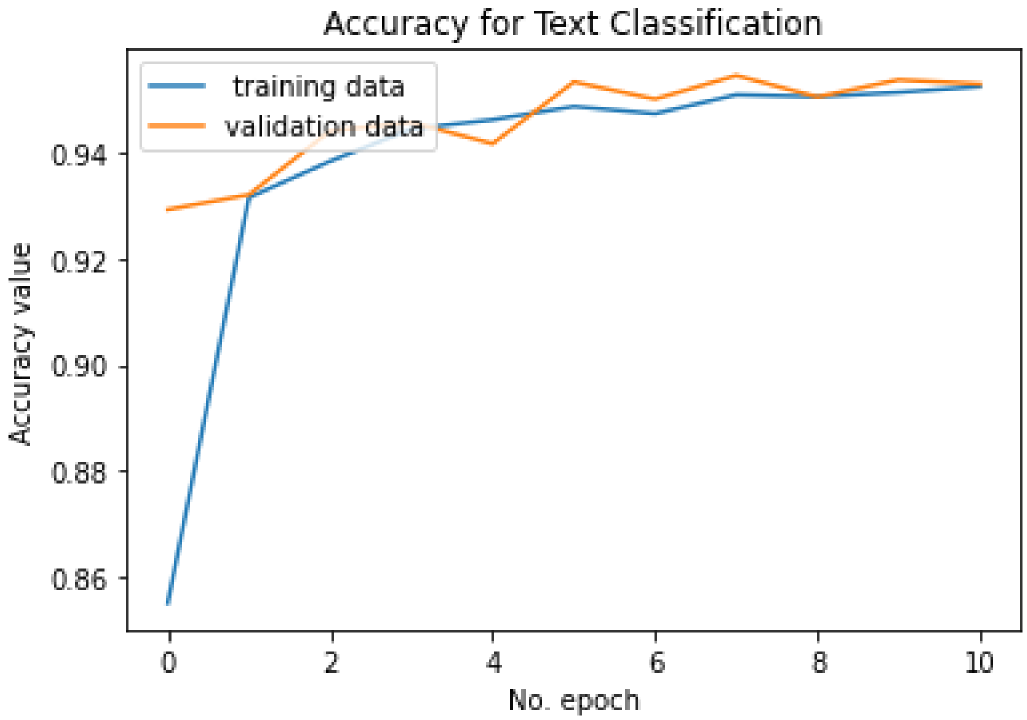


Figure 6.16: Training accuracy versus Validation accuracy

The validation accuracy is increasing along the epochs. But validation accuracy increases and decreases. But both training and validation accuracy values remain within similar range of each other.

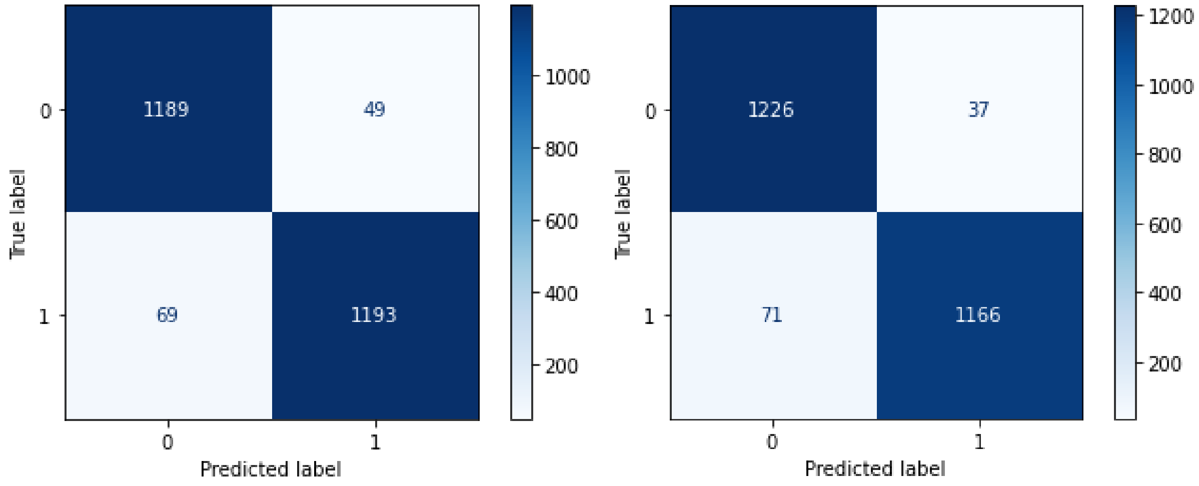
Evaluation and Results

The ten percent of dataset pre-defined for testing purpose was used to access the perfromance of the model. After evaluating the model using our train, test and validation data, we noted these metrics in table 6.7

Table 6.7: Table showing result comparison

Metric	Training	Testing	Validation
Accuracy	95.66%	95.68%	95.28%
Loss	0.1212	0.1221	0.1224
Steps	625	79	79
Time	240 secs	29 secs	29 secs

The confusion matrix of our validation and testing dataset is shown in the Figure 6.17a and Figure 6.17b below:



(a) Confusion Matrix of Validation Dataset

(b) Confusion Matrix of Test Dataset

The confusion matrices certify the predicting ability of our classifier.

Precision and recall of validation and test dataset were also calculated using sklearn's Classification Report in addition to accuracy which may be summarized in table 6.8 and table 6.9below:

Table 6.8: Table showing result of Validation Dataset

Support	Precision	Recall	F1-Score
Kids	0.95	0.96	0.95
Adults	0.96	0.95	0.95
Total	0.95	0.95	0.95

Table 6.9: Table showing result of Test Dataset

Support	Precision	Recall	F1-Score
Kids	0.95	0.97	0.96
Adults	0.97	0.94	0.96
Total	0.96	0.96	0.96

From the analysis of this subsection, we may note that the accuracy, precision and recall, F1-Score of the functional model was good like the sequential model's.

6.2.3 Comparison between the Functional and Sequential Model

Among the two models, the LSTM Sequential Model displayed a little bit better performance than the Functional Model with BERT pre-processing. But metrics value of both the models were well satisfactory. The main advantage of using BERT was the pre-processing and encoding simplicity with just two lines of code. As BERT is already a pre-trained model hence preprocessing and encoding can be done using pre made BERT preprocessors and encoders. Although, the training and evaluation time taken for the model with BERT was noticeably more than that of the LSTM sequential model, the pre-processing and feature extraction for LSTM model took much longer than BERT's preprocessing and encoding. However, the BERT model required a high power GPU, but the LSTM model trained decently even without a GPU. Thus weighing the pros and cons of the two models, we may state that both did well.

Chapter 7

Conclusion and Future Work

7.1 Conclusion

The main goal of our research was to propose a solution for displaying age appropriate news to users on social media platform. For reaching that goal we have built a machine learning model which classifies the age of a person based on the posts he or she writes and at the same time another artificial neural network model which classifies news as adult or kids news. So far we have reached an accuracy of 91.31% for age classification and 96.16% for news classification. For the future we wish to extend our dataset and try age based news classification based on the news headlines and not the articles.

7.2 Future Work

So far, we have applied the Machine Learning Models only on 8000 rows out of 681,288 rows of the dataset for classifying the underage users and 18 plus users. Our next approach is to apply the ML models on the whole dataset. For getting better accuracy in that case we need to do further preprocessing. We will also apply the Artificial Neural Network Models to compare the performance with that of the ML models. For the kids friendly news and adult news classification it is our next target to prepare the dataset based on the news headlines not the news articles and apply many more advanced Neural Network models on that. The advanced version of LSTM,i.e, Gated Recurrent Units (GRU) [48] can be applied and also many more to have a comparison with the previous result. Then we have the target to make an application where underage users will be detected from person's social media profiles,such as,facebook,twitter,instagram,etc. It can also help to indentify fake profiles. Then,the social media news with adult content will be filtered out for under 18 users. Only the kids friendly news which are preferable for the kids will be shown to their newsfeed.

References

[1] Global social media statistics research summary 2022 https://www.smartinsights.com/social-media-marketing/social-media-strategy/new-global-social-media-research (Accessed: 22 June 2022).

[2] Social Media Age Guide for Parents and Guardians https://www.beechfield.herts.sch.uk/social-media-age-guide-for-parents-and-guardians (Accessed: 22 June 2022).

[3] Age of Majority https://www.law.cornell.edu/wex/age_of_majority (Accessed: 22 June 2022).

[4] Age of majority. https://en.wikipedia.org/wiki/Age_of_majority(Accessed: 22 June 2022).

[5] Mason Walker and Katerina Eva Matsa. News consumption across social media in 2021. 2021.

[6] Shlomo Argamon, Moshe Koppel, James W Pennebaker, and Jonathan Schler. Mining the blogosphere: Age, gender and the varieties of self-expression. *First Monday*, 2007.

[7] Claudia Peersman, Walter Daelemans, and Leona Van Vaerenbergh. Predicting age and gender in online social networks. In *Proceedings of the 3rd international workshop on Search and mining user-generated contents*, pages 37–44, 2011.

[8] James Marquardt, Golnoosh Farnadi, Gayathri Vasudevan, Marie-Francine Moens, Sergio Davalos, Ankur Teredesai, and Martine De Cock. Age and gender identification in social media. *Proceedings of CLEF 2014 Evaluation Labs*, 1180:1129–1136, 2014.

[9] Avar Pentel. Automatic age detection using text readability features. In *EDM (Workshops)*, 2015.

[10] Jonathan Schler, Moshe Koppel, Shlomo Argamon, and James W Pennebaker. Effects of age and gender on blogging. In *AAAI spring symposium: Computational approaches to analyzing weblogs*, volume 6, pages 199–205, 2006.

[11] Inoshika Dilrukshi, Kasun De Zoysa, and Amitha Caldera. Twitter news classification using svm. In *2013 8th International Conference on Computer Science & Education*, pages 287–291. IEEE, 2013.

[12] Rongbo Du, Reihaneh Safavi-Naini, and Willy Susilo. Web filtering using text classification. pages 325–330, 2003.

[13] Carsten Eickhoff, Pavel Serdyukov, and Arjen P De Vries. A combined topical/non-topical approach to identifying web sites for children. In *Proceedings of the fourth ACM international conference on Web search and data mining*, pages 505–514, 2011.

[14] Tamara Polajnar, Richard Glassey, and Leif Azzopardi. Detection of news feeds items appropriate for children. In *European Conference on Information Retrieval*, pages 63–72. Springer, 2012.

[15] Menghan Zhang. Applications of deep learning in news text classification. *Scientific Programming*, 2021, 2021.

[16] Hasibe Busra Dogru, Sahra Tilki, Akhtar Jamil, and Alaa Ali Hameed. Deep learning-based classification of news texts using doc2vec model. In *2021 1st International Conference on Artificial Intelligence and Data Analytics (CAIDA)*, pages 91–96. IEEE, 2021.

[17] Sandeep Kaur and Navdeep Kaur Khiva. Online news classification using deep learning technique. *International Research Journal of Engineering and Technology (IRJET)*, 3(10):558–563, 2016.

[18] Md Ahmed Foysal, Syed Tangim Pasha, Sheikh Abujar, Syed Akhter Hossain, et al. Bengali news classification using long short-term memory. In *Emerging Technologies in Data Mining and Information Security*, pages 329–338. Springer, 2021.

[19] Raymond E Wright. Logistic regression. 1995.

[20] Statistics Solutions. What is logistic regression. *Retrieved from*, 2016.

[21] Brijeshkumar Y Panchal. Book genre categorization using machine learning algorithms (k-nearest neighbor, support vector machine and logistic regression) using customized dataset. *Book Genre Categorization Using Machine Learning Algorithms (K-Nearest Neighbor, Support Vector Machine and Logistic Regression) using Customized Dataset*, 2021.

[22] Shilamkoti Sowmy and K Ramesh Babu. A decision tree based recommended system for tourism. In *21st International Conference on Automation and Computing (ICAC)*, 2015.

[23] Vladimir Svetnik, Andy Liaw, Christopher Tong, J Christopher Culberson, Robert P Sheridan, and Bradley P Feuston. Random forest: a classification and regression tool for compound classification and qsar modeling. *Journal of chemical information and computer sciences*, 43(6):1947–1958, 2003.

[24] Ashraf M Kibriya, Eibe Frank, Bernhard Pfahringer, and Geoffrey Holmes. Multinomial naive bayes for text categorization revisited. In *Australasian Joint Conference on Artificial Intelligence*, pages 488–499. Springer, 2004.

[25] Richard Swinburne. Bayes' theorem. *Revue Philosophique de la France Et de l*, 194(2), 2004.

[26] Irina Rish et al. An empirical study of the naive bayes classifier. In *IJCAI 2001 workshop on empirical methods in artificial intelligence*, volume 3, pages 41–46, 2001.

[27] Shan Suthaharan. Support vector machine. In *Machine learning models and algorithms for big data classification*, pages 207–235. Springer, 2016.

[28] Akiko Aizawa. An information-theoretic perspective of tf–idf measures. *Information Processing & Management*, 39(1):45–65, 2003.

[29] Sahar Ghannay, Benoit Favre, Yannick Esteve, and Nathalie Camelin. Word embedding evaluation and combination. In *Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC'16)*, pages 300–305, 2016.

[30] AM Hay. The derivation of global estimates from a confusion matrix. *International Journal of Remote Sensing*, 9(8):1395–1398, 1988.

[31] Jinming Zou, Yi Han, and Sung-Sau So. Overview of artificial neural networks. *Artificial Neural Networks*, pages 14–22, 2008.

[32] Softmax function. https://en.wikipedia.org/wiki/Softmax_function(Accessed: 22 June 2022).

[33] Daniel Svozil, Vladimir Kvasnicka, and Jiri Pospichal. Introduction to multi-layer feed-forward neural networks. *Chemometrics and intelligent laboratory systems*, 39(1):43–62, 1997.

[34] Eric A Wan. Temporal backpropagation for fir neural networks. In *1990 IJCNN International Joint Conference on Neural Networks*, pages 575–580. IEEE, 1990.

[35] Forest Agostinelli, Matthew Hoffman, Peter Sadowski, and Pierre Baldi. Learning activation functions to improve deep neural networks. *arXiv preprint arXiv:1412.6830*, 2014.

[36] Xiaojie Jin, Chunyan Xu, Jiashi Feng, Yunchao Wei, Junjun Xiong, and Shuicheng Yan. Deep learning with s-shaped rectified linear activation units. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 30, 2016.

[37] Hidenori Ide and Takio Kurita. Improvement of learning for cnn with relu activation by sparse regularization. In *2017 International Joint Conference on Neural Networks (IJCNN)*, pages 2684–2691. IEEE, 2017.

[38] Kamel Abdelouahab, Maxime Pelcat, and François Berry. Why tanh is a hardware friendly activation function for cnns. In *Proceedings of the 11th international conference on distributed smart cameras*, pages 199–201, 2017.

[39] Sridhar Narayan. The generalized sigmoid activation function: Competitive supervised learning. *Information sciences*, 99(1-2):69–82, 1997.

[40] Kozo Takayama, Mikito Fujikawa, and Tsuneji Nagai. Artificial neural network as a novel method to optimize pharmaceutical formulations. *Pharmaceutical research*, 16(1):1–6, 1999.

[41] Diederik P Kingma and Jimmy Ba. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*, 2014.

[42] Agnes Lydia and Sagayaraj Francis. Adagrad—an optimizer for stochastic gradient descent. *Int. J. Inf. Comput. Sci*, 6(5):566–568, 2019.

[43] Jarek Duda. Sgd momentum optimizer with step estimation by online parabola model. *arXiv preprint arXiv:1907.07063*, 2019.

[44] Tomas Mikolov, Martin Karafiát, Lukas Burget, Jan Cernockỳ, and Sanjeev Khudanpur. Recurrent neural network based language model. In *Interspeech*, volume 2, pages 1045–1048. Makuhari, 2010.

[45] Sepp Hochreiter and Jürgen Schmidhuber. Long short-term memory. *Neural computation*, 9(8):1735–1780, 1997.

[46] Blog Authorship Corpus. https://www.kaggle.com/datasets/rtatman/blog-authorship-corpus(Accessed: 21 June 2022).

[47] Andrew Thompson. All the news, Aug 2017.

[48] Junyoung Chung, Caglar Gulcehre, KyungHyun Cho, and Yoshua Bengio. Empirical evaluation of gated recurrent neural networks on sequence modeling. *arXiv preprint arXiv:1412.3555*, 2014.

Generated using Undegraduate Thesis LaTeX Template, Version 1.4. Department of Computer Science and Engineering, Ahsanullah University of Science and Technology, Dhaka, Bangladesh.

This thesis was generated on Monday 22nd August, 2022 at 9:50am.