Thesis Report

Identifying Hate Speech of Bangla Language Text using Natural Language Processing

by

Mushfiqur Rahman
18301121
Razia Sultana Jui
18301021
Chowdhury Nazmuz Sakib
18301109
Fahim Alavi Ridoy
19301071
Taskiea Tabassum Ananya
19301192

Supervisor:
Mr. Annajiat Alim Rasel
Co-supervisor:
Dr. Muhammad Iqbal Hossain
Mr. Dewan Ziaul Karim

Department of Computer Science and Engineering
School of Data and Sciences
Brac University
January 2024

Declaration

It is hereby declared that

1. The thesis submitted is our own original work while completing degree at Brac University.
2. The thesis does not contain material previously published or written by a third party, except where this is appropriately cited through full and accurate referencing.
3. We have acknowledged all main sources of help.

Student's Full Name & Signature:

mushfiqur

Mushfiqur Rahman

18301121

jui

Razia Sultana Jui

18301021

sakib

Chowdhury Nazmuz Sakib

18301109

alavi

Fahim Alavi Ridoy

19301071

Taskiea

Taskiea Tabassum Ananya

19301192

Approval

The thesis titled “Identifying hate speech of Bangla language text using natural language processing” submitted by

1. Mushfiqur Rahman (18301121)
2. Razia Sultana Jui (18301021)
3. Chowdhury Nazmuz Sakib (18301109)
4. Fahim Alavi Ridoy (19301071)
5. Taskiea Tabassum Ananya (19301192)

Of Fall 2023 has been accepted as satisfactory in partial fulfillment of the requirement for the degree of B.Sc. in Computer Science and Engineering.

Examining Committee:

Supervisor:
(Member)

Annajiat Alim Rasel

Senior Lecturer
CSE Department
Brac University

Program Coordinator:
(Member)

Dr. Md. Golam Rabiul Alam

Professor
CSE Department
Brac University

Head of Department:
(Chair)

Sadia Hamid Kazi

Chairperson
CSE Department
Brac University

Acknowledgement

Firstly, all praise to the Great Allah, for whom our thesis has been completed without any major interruption.
Secondly, to our co-supervisors and supervisor for their kind support and advice in our work. He helped us whenever we needed help.
And finally to our parents, without their throughout sup-port it may not be possible. With their kind support and prayer, we are now on the verge of our graduation.

Table of Contents

List of Figures

List of Tables

Chapter 1

Abstract

In this era of the internet, sharing information through social media has provided significant benefits to humans. People can easily access and observe others' lifestyles and work, as well as make comments or share thoughts about them. However, this practice also brings challenges, such as the spread of hate comments, abusive online criticism, spreading toxicity by giving hate comments etc. The internet's flexibility and anonymity have created a culture where users find it easy to express themselves aggressively in communication. As the amount of hate speech is increasing, there is a need for a method to automatically detect hate speech. To tackle this concern, recent research has utilized diverse feature engineering methods and machine learning algorithms to autonomously identify hate speech messages across various datasets. Since it is related to Natural Language Processing (NLP), our goal is to utilize NLP to detect hate speeches and demonstrate how Deep Learning and Machine learning can be used for this purpose. Since there are more than 7,100 languages spoken throughout the world [32], we have chosen the Bengali language as our dataset language. Additionally, with the help of machine learning and deep learning, we will train our model to automatically detect hate speech. We are utilizing Multinomial Naive Bayes, RNN, Random Forest, Logistic Regression, Decision Tree Classifier, CNN-LSTM Hybrid algorithm and Multi lingual Bidirectional Encoder Representations(mBert) for result comparison and optimal outcomes and accuracy. After employing all the above algorithms, we found the highest accuracy using the mBert for the binary classification, which is 90.00%. On the other hand, for Multilevel classifications, we have found the highest accuracy using CNN-LSTM Hybrid algorithm, which is 64% and the second highest is 62% using mBert. We are committed to further improving these results.

Chapter 2

Introduction

2.1 Introduction

In recent times, the incidence of hate speech in virtual environments has significantly increased. This escalation has led to physical violence, human rights violations, and crime. For instance, recent surveys have linked the rise in online hate speech content to hate crimes, including the election of Trump in the US [9] and the Manchester and London attacks in the UK [10]. As people become more engaged in social media, the risk of violence also rises. However, to address this situation, certain organizations have taken steps, such as the European Union enforcing social media platforms to sign a deal to remove all hate posts or comments within 24 hours [6]. However, this manual process has its limitations, including the potential for consequences and additional costs. This is where the concept of automatic hate speech detection comes into play. Social media platforms face criticism for not having robust prevention mechanisms against hate speech removal. However, automating the process presents a challenge due to language and cultural barriers. There are disagreements regarding different forms of hate speech, as what may be considered hate speech in one region may not be viewed as such in others. According to the United Nations, hate speech is defined as "any kind of communication in speech, writing, or behavior that attacks or uses pejorative or discriminatory language with reference to a person or group based on their religion, ethnicity, nationality, race, color, descent, gender, or other identity factors". Although there are differences in definitions, recent studies have shown that achieving favorable results is possible [5] [8] [2]. Therefore, feature engineering and classification machine learning algorithms are necessary. After analyzing the results of these approaches, we need to compare them with different feature engineering techniques and machine learning algorithms. In this study, we utilize Natural Language Processing (NLP) and compare various feature engineering and machine learning algorithms to achieve the best possible outcome. By applying the deep learning method, we aim to achieve high accuracy. This paper aims to demonstrate how to improve the solution to this problem. We analyzed a published dataset which contains 30,000 data samples from different sources on the internet, ensuring diversity. In the dataset, there are two columns which can be used as levels.We used one column for binary classification (hate speech or not hate speech) and the other column was used as multi-level classification as it has 7 unique categories such as (Sports, Entertainment, Crime, Religion, Politics, Celebrity and Meme TikTok) in it. We then applied Multinomial Naive Bayes, Recurrent Neural Networks(RNN), Random Forest, Logistic Regression, Decision Tree Classifier, CNN-LSTM Hybrid algorithm and mBert. Finally, we presented our findings and analyzed the

challenges associated with detecting hate speech.

2.2 Problem Statement

As social media use increases day by day, the issue of hate speech is also increasing. According to [26], numerous segments of the population have adopted social media nowadays. Because of that, hate speech has spread rapidly around the world. It is very easy to spread hate speech because to do this one only needs a phone or laptop, an internet connection and a corrupted mind. There are lots of social media platforms all over the world, like Facebook, Twitter and YouTube, where people can freely comment. Hate speech can affect social and political issues. Also, it can cause racism and nationalism. According to [39], Though hate speech always exists, during the COVID-19 pandemic it suddenly exacerbated. And this hate speech harms not only the targeted person but also the whole society. Hate speech suppresses the values of tolerance, diversity, and inclusion. It also violates human rights and harms society's development and peace. According to [30], there are 2.85 billion active users on the Facebook platform. Every day, 1.9 billion users access Facebook. This is a huge number. According to [13], the propagation of hate speech is increasing and to counter these governments, many companies give money and take necessary steps. This paper says that though they try to take steps against hate speech, there is a limitation to detecting these types of speech because of the lack of comparative evaluation. This is a huge problem, and the number of social media comments and posts is relatively large, so for any person or any organization, it is not possible to detect them manually. To decrease the problem of hate speech or take any necessary action against it, hate speech first needs to be detected. So, by using NLP and ML, we are going to detect hate speech with the help of some ML techniques and methods. Though there is some research work regarding these issues in [26] [11] [13] [20]. But our goal is to create a huge dataset, try to optimize more in the technique to increase the accuracy.

2.3 Research Objective

This research aims to develop a hate speech detection system to identify a block of text using machine learning and deep learning [34]. After applying detection methods, implementation details and performance are observed. Deep learning models such as multilingual BERT, Recurrent Neural Networks (RNN), and a LSTM-CNN hybrid model are used for improving accuracy in detecting hate speech [28].
The objectives of this research are:

1. To deeply understand machine learning, deep learning, and how they work.
2. To consider other approaches before selecting the best approach: datasets, detection methods, implementation details, and performance evaluation.
3. To deeply understand hate speech detection techniques.
4. To develop a model for identifying hate speech based on machine learning and natural learning processing.
5. To provide recommendations to enhance the model's performance.

Chapter 3

Related Work

The advancement of the internet has brought both advantages and disadvantages into perspective. One of them is using the internet to express negative thoughts and emotions to harm or slander someone or something. Identification of these types of expressions can help to be more cautious or take whatever necessary steps are needed to prevent them. In this section, the main goal is to review previous related works similar to the field we have chosen to research.

3.1 Literature Review

In the paper written by Yoon Kim, Convolutional Neural Network (CNN) is used for sentence classification which is a much better method than other existing ones [4]. According to the same paper, it is shown that a simple CNN method will assign the value from the dataset to different categories and based on that, it classifies the hate speeches [4]. This model has 91% accuracy, 90% recall, 90% F-measure, and 91% precision [11]. However, there is one underlying problem, that it misclassifies a lot of hate speeches [11].

In another paper published by the authors Z. Zhang, D. Robinson, and J. Tepper, Convolution-GRU based deep neural networks are used to identify hate speeches [13]. According to the paper, the conduct of the evaluation of the method of publicly available datasets shows that CNN+GRU has better learning accuracy as it captures word sequence and order information compared to other methods where only CNN is used [13]. At first, the authors introduced a method for automatically classifying hate speech using a model with a combination of GRU and CNN to improve the accuracy of the classification and did a comparative evaluation with public datasets which showed that their proposed method was better [13].

According to a recent paper, a pre-trained transfer learning protocol, BERT, is used for understanding hate speeches that are bidirectional [12]. The lack of adequately labeled data could be one of the reasons for using BERT. It is causing CNN and LSTM (Long Short Term Memory) to misclassify a lot of data. So, a transfer learning approach can be introduced based on English Wikipedia and the Book Corpus for hate speech detection on available benchmark datasets [19]. New parameters do not need to be learned, which is an advantage. The authors introduced fine-tuning strategies for exploring different layers of BERT detecting hate speech [19]. BERT-based fine-tuning layer, insert nonlinear layers,

Insert Bi-LSTM layer, and Insert CNN layer are the four fine-tuning processes found in the mentioned paper [12].

DistilBERT and BERT have the same architecture. The DistilBERT transfer method was used and the produced result was better than the BERT, attention-based, and some other neural network and transformer methods used to detect hate speech [20]. This method allows parallelization [20]. The authors analyzed the result in terms of six standard functional metrics of accuracy, precision, recall and F-measure, Mathews correlation coefficient and evaluation loss [20]. DistilBERT surpasses the baseline algorithms. It had a higher accuracy rate of 92%, 75% precision, 75% recall, a 75% MCC score, 28% evaluation loss, and 75% F-measure score [20]. The authors assessed DistilBERT on General Language Understanding Evaluation or GLUE benchmark [16]. This model is faster and smaller than the BERT model, which is why this is better in comparison with BERT. Because DistilBERT reduces 40% of a BERT model size and is 60% faster than BERT [16]. Also, the DistilBERT can maintain up to 97% of it's capabilities to understand it's language [16]. This model is cheaper to pre-train and fine-tuned with good performance. The authors avoided using a multi-tasking scheme for fine-tuning.

The next paper is based on the Bangla language where the authors used several deep learning models with pretrained Bengali words to predict hate speech [29]. They divided all sorts of comments into seven categories: sports, religion, crime, politics, entertainment, celebrity, and TikTok & memes [29]. Some of them are SVM, Word2Vec + LSTM, FastText + LSTM, BengFastText, etc. However, SVM has the best accuracy rate which is 87.5% [29].

According to the paper, the authors detected and analyzed hate speech through a Multichannel Convolutional-LSTM (MC-LSTM) network [18]. They used BengFastText, Word2Vec, GloVe, etc models which had F1 scores of 92.30%, 82.25%, and 90.45% respectively. BengFastText, which is one of the largest Bengali word embedding models, can catch the semantics of words without any problem [18].

Hate speech has become a common problem nowadays. In one of the papers, it is mentioned that a machine learning model is used to classify Bengali comments which is encoder-decoder based [22]. Here, for the encoder, a 1D convolutional layer is used, and for decoding, the ones that are LSTM, GRU, and attention based decoders are used that have a 77% accuracy rate [22]. This NLP tool is solely used to detect Bengali hate speeches. The authors, here, used both the Graph API and manual comments. Bangla Emot Module was used to detect emotions from the emojis [22]. TF-IDF and word embedding were used to perform better [22]. Two categories of abusive text classification which are binary and multi-class classifications are mentioned in the paper [22]. Multiclass had the highest accuracy.

According to the paper, there are six toxic categories in terms of hate speeches, which are determined by a binary classification model (LSTM with BERT) and a multi-label classifier (CNN-BiLSTM with attention). The binary classification model has an 89.42% accuracy rate and the multi-label classifier, on the other hand, has a 78.92% of accuracy rate [36]. If a comment is toxic or not, it can be identified through a pipeline by using a binary classification model and the toxicity type of the comment can be identified through

multi-class classification [36]. A framework, LIME, is used to identify deep learning models [36]. LSTM with BERT Embedding had the best performance in binary classification compared to others like MConv-LSTM with BERT Embedding or Bangla BERT fine tuning [36].

In the next paper, machine learning and deep learning based approaches are used. Only multimodal datasets were used in this paper. Conv-LSTM and XLM-RoBERTa have good performance with F1 scores of 78% and 82% respectively [31]. ResNet-152 and DenseNet-161 models are used for memes [31].

In another paper, it is mentioned that using GRU based RNN on n-gram dataset, a new language model can be created which will predict the words using provided inputs [15]. The average accuracy rate for 5-gram, 4-gram and Tri-gram are 99.70%, 99.24%, and 95. 84% respectively.

To detect hate speech of Bengali language, the texts are classified into different categories using transformer-based neural architecture methods (such as monolingual Bangla BERT-base, multilingual BERT-cased, etc.) after preprocessing [24]. Machine learning and neural network baselines are better than ML and DNN in terms of performance, as the accuracy score of F1 is 91%, 89%, 84%, and 78% respectively for personal, geographical, religious and political hates [24]. XML-RoBERTa model is the best-fitted model in terms of performance that has the 87% of F1 score [24].

The authors introduced HateXplain, the first benchmark hate speech dataset which captures human rationales for the labeling in this paper [25]. MTurk is used to collect dataset. CNN-GRU, BiRNN, BiRNN-Attention, and BERT models are used in this paper. BiRNN-HateXplain and BERT-HateXplain have a better performance [25]. BERT-HateXplain has the best bias but the worst score for sufficiency than the others [25]. CNN-GRU has the best score in terms of sufficiency.

Using abusive languages has become pretty common nowadays. Authors Viktor Hangya and Alexander Fraser mentioned a two steps approach, training models in a multitask fashion and carrying out adaptation to the target requirements [37]. The MLD approach is used for adaptation. Some datasets used are AMI, GermEval, LSA, SRW, ToLD-Br, etc.

In the paper, machine learning and deep learning based algorithms such as LinearSVC, ANN, RF, RNN with a LSTM can detect multi-type abusive Bengali text [14]. Some new stemming rules are introduced to gain better algorithm performance. Deep learning based RNN algorithm has an accuracy rate of 72.20%[14].

According to the paper, three groups were mentioned: traditional (shallow) classification methods, word embedding-based deep methods, and transformers-based deep methods [45]. Some popular deep neural network architectures include CNN, LSTM, and Bi-LSTM [33]. The authors of the paper used three datasets, Davidson, Founta and Twitter Sentiment Analysis (TSA) [33]. TF-IDF embeddings with XGB have the highest F1 score using Davidson datasets. Glove embedding with CNN and Bi-LSTM is similar as well.

The TF-iDF-based MLP and SVM models also have good performance[33].

The authors, Aneri Rana and Sonali Jha, talked about how offensiveness can be detected in multimedia via three modalities: visual, acoustic, and verbal [35]. A new dataset Hate Speech Detection Video Dataset (HSDVD) is introduced. In the published dataset of 24k tweets by Davidson (2017), 5.77% hate speech, 77.43% offensive or neither[35].

In a paper by Md. Saroar Jahan and Mourad Oussalah [38], the authors provided a systematic review where they focused on NLP and deep learning architectures & technologies using the PRISMA framework. This approach analyzes theoretical aspects and practical resources [38]. They mentioned that the output of the machine learning model could be either multiclass where the model distinguishes if it is a hate speech or not, or it could be a binary decision. They had collected multilingual data from two different databases and excluding the unrelated, unnecessary, & duplicate records, they were left with 463 articles [38] to use for their work. Only 51% of all works [38] were in English language and the other 49% were in various languages such as Arabic (13%), Turkish (6%), Hindi (4%), Bengali (1%), Korean (1%), French (1%), etc [38] . Among the machine learning approaches, they distinguished the datasets using supervised (73%), semi-supervised, and unsupervised approaches. Even though there is better accuracy in performance using the unsupervised approach, the supervised approach is used widely due to the multiplication of benchmarking datasets [38]. Their paper showed that TF-IDF-based features cover 29%, word embedding models cover 33%, PoStag covers 3%, topic modeling cover 3%, and sentiment cover 3% of the entire records. They used three strategies of annotation scheme (binary scheme, non-binary scheme, multi-level annotation) that were based on binary classification and ternary class level [38]. They concluded that BERT-based models (90% F1 score) perform better than other models like FastText, Word2Vec, GloVe, etc [38].

In another paper [11] written by Shanita Biere and Prof. Dr. Sandjai Bhulai, they used CNN architecture using 24,783 English tweets as datasets that had been classified into three classes which are hate (5%), offensive (77%), and neither (18%) [11] to detect hate speech. The authors mentioned that their final model had an accuracy of 91%, precision of 91%, recall of 90%, an F-measure of 90%, and a loss of 36% [11]. In their model, almost 80% of the hate class [11] is not classified correctly resulting in making the model biased towards the offensive class.

Chapter 4

Methodology

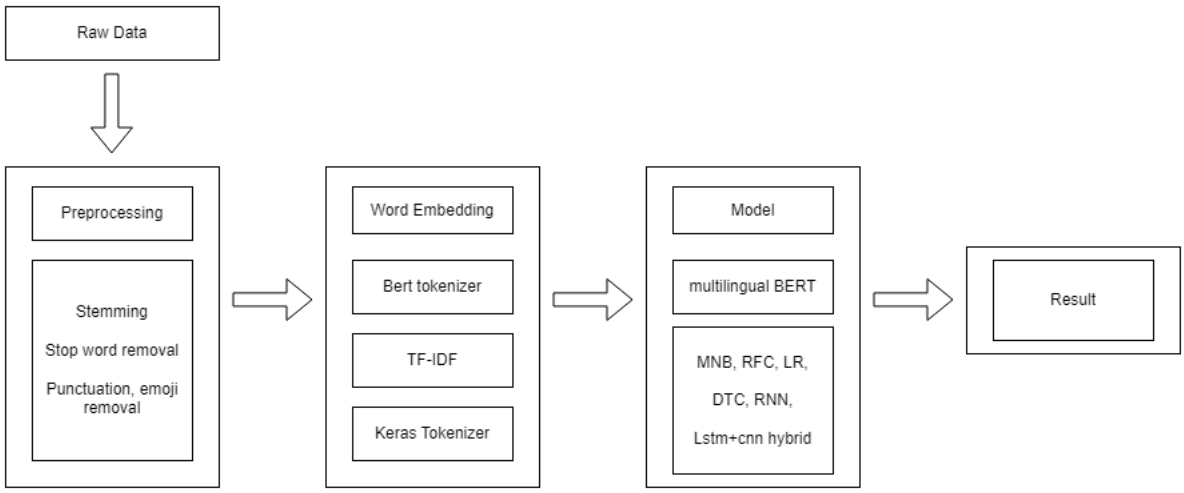

Figure 4.1: Work Flow

The above graph shows the general workflow of our studies.

4.1 Dataset Collection

We collected the dataset from [29]. This dataset contains 30,000 Bangla language data from various social media platforms. This dataset contains different categories of hate speech. From 2017 to 2020, Bangladeshi Facebook and YouTube comments on divisive issues were collected for the data collection. Sports, entertainment, crime, religion, politics, celebrity memes, and other categories were used to group the comments. The feedback was also gathered from public Facebook sites, well-known writers, scientists, and celebrities. YouTube became the primary source of data as a result of Facebook's graph API's limited accessibility. The extraction of comments was completed after choosing divisive subjects and popular videos. Below are shown some data examples and statistics about the dataset:

Sentence	Hate Speech	Category
যত্তসব পাপন শালার ফাজলামী!!!!!	1	1
সালার বাবা কয়টা??? মারালি বাবা জারালি বাবা মুতখোর বাবা বোকাচুদা বাবা যেগুলা বসে আছে সবার পাছা দিয়া গরম আগন ভরতে হবে। সালার ভন্ডর বাচ্চারা।	1	2
জনাব জাফর ষার,অতি অসম্যানের সাথে জানাচ্ছি, অাপনি একটা বানচোদ।	1	3
সৌম্য সরকার কে বাদ দেওয়া হোক	0	3

Table 4.1: Sample Dataset

Hate Speech	Not Hate Speech	Total
10000	20000	30000

Table 4.2: Total Data

4.2 Data Preprocessing

The dataset we are using contains raw data and for that reason doing proper data preprocessing is important for better results. The process of converting unprocessed data into a format that can be used, viewed, and understood during further in-depth study is known as data preparation or preprocessing [23]. It involves a series of actions to clean, transform, and prepare raw data for analysis and model training. Data preprocessing helps improve the data's quality by removing noise, handling missing values, and reducing dimensionality through feature selection and extraction. This step is crucial as the data quality directly influences the effectiveness of machine learning algorithms. Effective data preprocessing can improve the efficiency and accuracy of models, making it a critical component in the data analysis pipeline.

4.2.1 Data Cleaning

Data cleaning, a pivotal stage in data preprocessing, entails pinpointing and rectifying errors, inconsistencies, and inaccuracies within datasets. The primary goal is to ensure that the data is accurate, reliable, and ready for analysis [23]. When we acquire raw data, it often contains a lot of extraneous information. For instance, if someone leaves a remark, it frequently includes emoji, null data, punctuation, and different signs as well. Therefore, data cleaning is required in order to sanitize the data.

Null Data Remover

A null data remover is a data preprocessing technique used to eliminate or handle missing or null values in a dataset. Our primary purpose is to ensure data completeness and integrity before performing data analysis or training models. This process involves identifying missing values in the dataset and applying strategies like deletion, imputation, or interpolation to either remove or replace the missing data points [23]. The choice of strategy depends on the nature of the data and the specific analysis or modeling goals. Removing

null data helps prevent bias and errors in subsequent analyses and ensures that data-driven insights are based on as much complete and reliable information as possible.

Emoji and Punctuation Removal

Emoji and punctuation data removal is another data preprocessing step that involves cleaning text data by removing emojis (such as smiley faces or symbols) and punctuation marks (like commas, periods, or exclamation points). This procedure involves analyzing text and simplifying and standardizing it, making it easier for algorithms to understand and process. Removing emojis and punctuation helps reduce noise and focuses the analysis on the essential text content, improving the accuracy and effectiveness of our model and text-based applications.

4.2.2 Data Transformation

Data transformation is the process of converting data from one system's format—the source system's—to the format needed by a destination system. Data transformation is often involved in a variety of data integration and management tasks, including data warehousing and data wrangling [23] .

Stemming

Stemming is a method of natural language processing that generates morphological variations of a base or root word. Stemming algorithms, often referred to as stemmers, transform words like "retrieval" into their base form, "retrieve". Text is normalized throughout this step to make it easier to process during the pipelining stage of natural language processing. In text preliminary processing, retrieval of data, and text mining applications, stemming is a crucial step. However, it additionally makes writing harder to understand and cannot always provide the right word's root form. Tokenized words are created by dissecting texts into individual words using tokenization. For tasks like text categorization, retrieval of data, and text summarization, stemming is helpful. People these days prefer to utilize short forms; thus they frequently use the short version of a term to post any comments. In these situations, data stemming aids in understanding or locating any word's root.

4.2.3 Feature Extraction

Feature extraction is a method that transforms raw data into meaningful numerical features, maintaining the essential content of the original data set. Feature extraction involves the creation of novel features by utilizing linear or nonlinear combinations of the existing variables [23].

TF-IDF Embedding

Inverse Document Frequency (IDF) is a complement to Term Frequency (TF). It basically entails determining a word's importance within a text's corpus or word order. A word's relevance is modified by its frequency in the dataset, which takes into account how frequently it appears in the text.

BERT Tokenizer

Sub word-based tokenization is a technique used by BERT tokenizer. Unknown words are broken down into shorter words or characters by sub word tokenization so that the model can make sense of the tokens. 'Boys', for instance, is separated into 'Boy' and's'. The vocabulary for BERT is generated using the word piece method [12]. This approach enables BERT to handle out of vocabulary words also helps capture more fine-tuned linguistic information [12].

4.3 Baseline models and setup

Though there are a lot of Machine learning algorithms for analyzing the numerical data, there are very few algorithms for the text data. As we are working with a text data set we use some of the text analyzer data algorithms for our work. For example we use Multinomial Naive Bayes, Linear SVC, Random Forest Regression, Logistic regression, DecisionTreeClassifier, lstm+cnn and mBert.

4.3.1 Models

Multinomial Naive Bayes

This algorithm is one of the most useful supervised learning algorithms which is used to analyze the categorical text data.[35] This algorithm is very popular nowadays. This is a probabilistic learning method and this algorithm is mostly used in the NLP(Natural Language Processing).
From the name we can see that this algorithm works based on the bayes theorem. But unlike naive bayes it can predict the tag from text. This is a collection of many algorithms and shares common principles and this algorithm is not affected by one's absence [7].

$$P(x|C_k) = \frac{\sum_i *x_i!}{\prod_i *x_i!} \prod_i (Pk_i)^{x_i} \tag{4.1}$$

Naive bayes mainly predict the probability of an event given by a probability of known event. This algorithm is very popular for analyzing the text data of multiple classes. This algorithm is very compact and powerful as this only calculates the probability so this algorithm is easy to implement. Also works on both discrete and continuous data[7].

Recurrent Neural Network (RNN)

Recurrent neural networks, or RNNs, are a type of artificial neural network designed to process data sequences. They do exceptionally well in problems involving speech recognition, natural language processing, time series data, and related fields. Unlike feedforward neural networks, which use a linear strategy, RNNs have looping connections that allow them to maintain a hidden state and record information about previous inputs in the sequence [21].

In RNN the weight of the matrics remain same across the network. For each input x it measure the hidden state h[21].

$$h = \sigma(UX + Wh_{-1} + B) \tag{4.2}$$

$$h_t = f(h_{t-1}, x_t) \tag{4.3}$$

Random Forest Regression

This model is also a supervised learning algorithm and for regression this model uses the ensemble learning method[17]. ensemble learning means that this model takes the combined prediction from different algorithms which makes the prediction more accurate than one model.
This model is also powerful and accurate and non linear relationships can be featured by this algorithm.
First of all take a random data from the training dataset then based on that data make a decision tree then choose as many trees as we want and repeat the first two process and to feed new data we need to make decision trees and predict the value of y[17].

Logistic Regression

This model is also a supervised learning algorithm. This algorithm uses some independent variable to predict the categorical dependent variable[3]. There are different kinds of Logistic regression.
This algorithm takes the value between 0 and 1 and then by voting this algorithm tries to predict the output and solve the classification problem.

Decision Tree Classifier

This model is also a supervised learning algorithm and can perform in both regression and classification problems[1].
In this algorithm the decision tree has two nodes one is leaf node and the other is decision node. Leaf nodes are the output of that decision tree and decision nodes are taking decisions. There are multiple branches of that decision tree. There is a root node and then based on 0 and 1 or yes no the decision node becomes either leaf node or in the sub tree.

CNN+LSTM

This model is also a supervised learning algorithm. LSTM means long short term memory which means this type of RNN is used to solve time series problems, long text recognized problem etc. In RNN there is no memory but to solve vanishing gradient problem and memory problem it comes the LSTM. This is very powerful algorithm as it can backup some memory.
CNN means convolution neural network which is also used to solve the text recognization problem. So this LSTM and CNN together make the hybrid algorithm whish is very accurate and powerful.

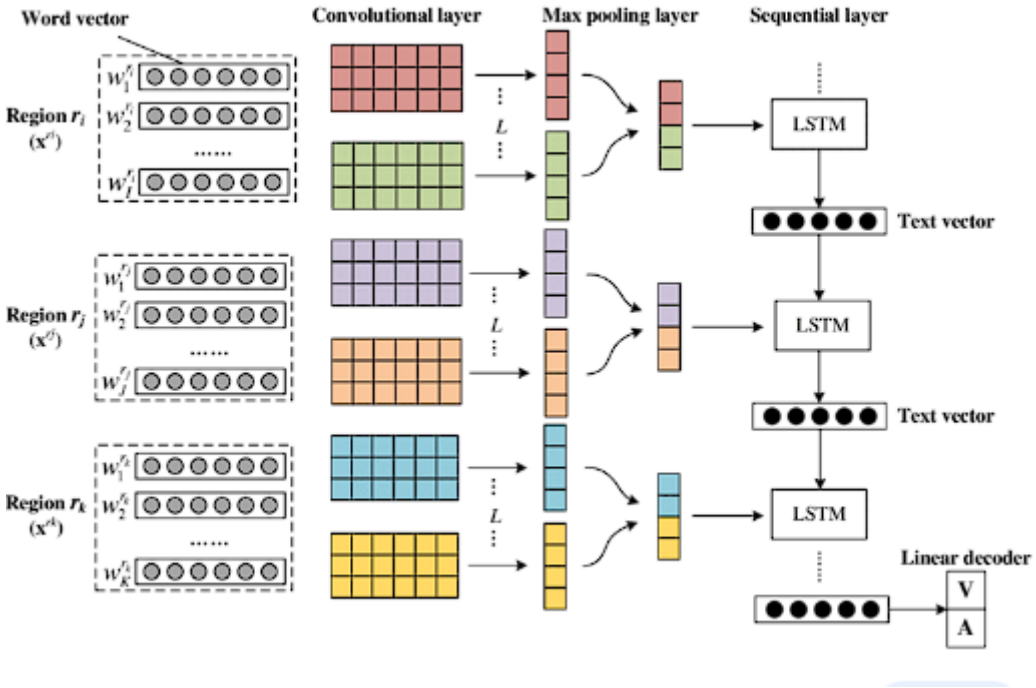

Figure 4.2: CNN+LSTM

Multilingual BERT

mBERT (Multilingual BERT) is a natural language processing (NLP) model developed by Google Research. It is an extension of the original BERT (Bidirectional Encoder Representations from Transformers) model specifically designed to handle multiple languages. BERT itself is a deep learning model that has achieved state-of-the-art performance on various NLP tasks by pretraining on a massive corpus of text and then fine-tuning for specific tasks[27].

The key innovation of mBERT is its ability to understand and generate representations of text in multiple languages. This is achieved by training the model on a multilingual dataset containing text from a wide range of languages. As a result, mBERT can provide contextual embeddings for words and sentences in multiple languages, making it a valuable tool for cross-lingual NLP tasks such as translation, sentiment analysis, and named entity recognition[27].

In summary, mBERT (Multilingual BERT) is a multilingual extension of the BERT model that can understand and work with text in multiple languages, offering significant benefits for cross-lingual NLP applications.

4.3.2 Experimental Setup

In this experiment, we used machine learning model as well as deep learning model to find the best accuracy in detecting hate speech. The dataset is divided into 80:20 split ratio for training and testing for all the models used. There is four traditional machine learning model and three deep learning model used in this experiment. The models were trained for 10 epochs using the Adam optimizer, with a batch size of 16 and a learning rate of 2e-5. Below we mentioned all the models that we evaluated on our dataset.

- Multilingual BERT with BERT tokenizer
- Recurrent Neural Networks with Keras Tokenizer
- CNN-LSTM Hybrid with Keras Tokenizer
- Multinomial NB with TF-IDF Embedding
- Logistic Regression with TF-IDF Embedding
- Decision Tree Classifier with TF-IDF Embedding
- Random Forest classifier with TF-IDF Embedding

4.3.3 Evaluation metric

We choose to use accuracy and F1-score as our main performance measures in our study. Their complementing nature in evaluating a model's categorization abilities is the main reason for this decision. The F1-score delivers a balanced evaluation of precision and recall, making it well-suited for imbalanced datasets or scenarios where class distribution matters. Accuracy provides a clear measure of overall correctness in predictions. We think that using a dual assessment strategy will provide us a thorough picture of how well our model classifies occurrences. We have used bold for the best-performing score and underline for the second-best score.

Chapter 5

Result & Analysis

5.1 Result

In this section we are going to discuss the findings of our study. We used all the above mention models to find the best accuracy and performance. We used binary classification such as 'hate speech' and 'not hate speech' as output. After evaluating all the model, we can observe that multilingual BERT provides us with the highest performance. We also used multilevel classification And after performing all the above mentioned machine learning and deep learning model from CNN-LSTM Hybrid model we get the heighest accuracy.

5.2 Model performance

Binary Classification:
As mentioned before, we used traditional machine learning as well as deep learning model to find the best accuracy. For the traditional ml models, we used Multinomial Naive Bayes, Logistic Regression, Decision Tree Classifier, and the Random Forest Classifier.

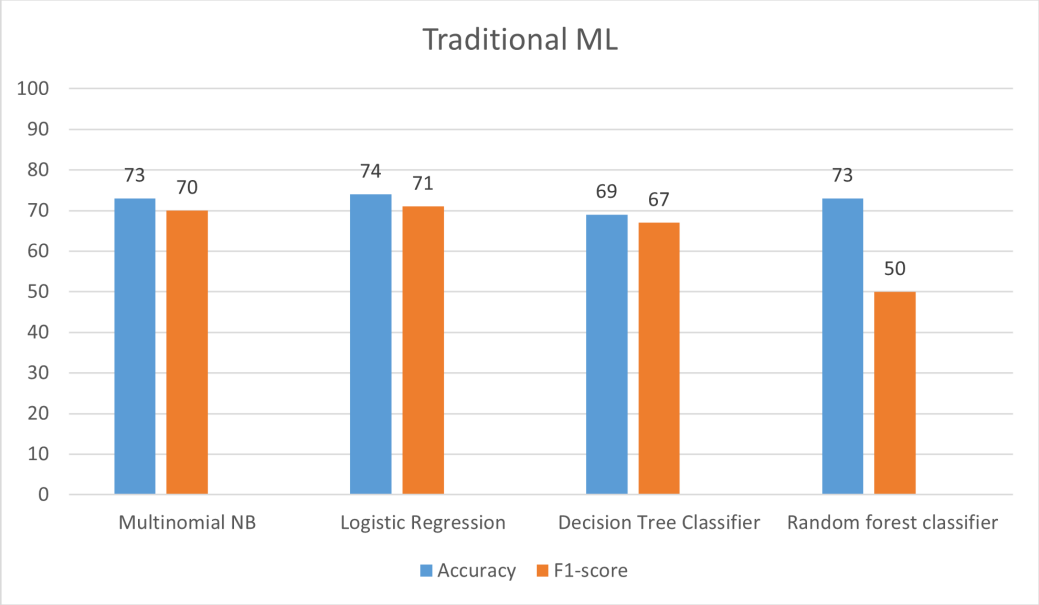

Figure 5.1: Machine Learning for Binary Classification

	Accuracy	F1-Score
Multinomial NB with TF-IDF Embedding	0.73	0.7
Logistic Regression with TF-IDF Embedding	0.74	0.71
Decision Tree Classifier with TF-IDF Embedding	0.69	0.67
Random forest Classifier with TF-IDF Embedding	0.73	0.5

Table 5.1: Accuracy Table of ML for Binary Classification

Now for the deep learning models we used multilingual BERT, Recurrent Neural Networks (RNN), and a hybrid model combining Convolutional Neural Networks (CNN) and Long Short-Term Memory (LSTM) networks.

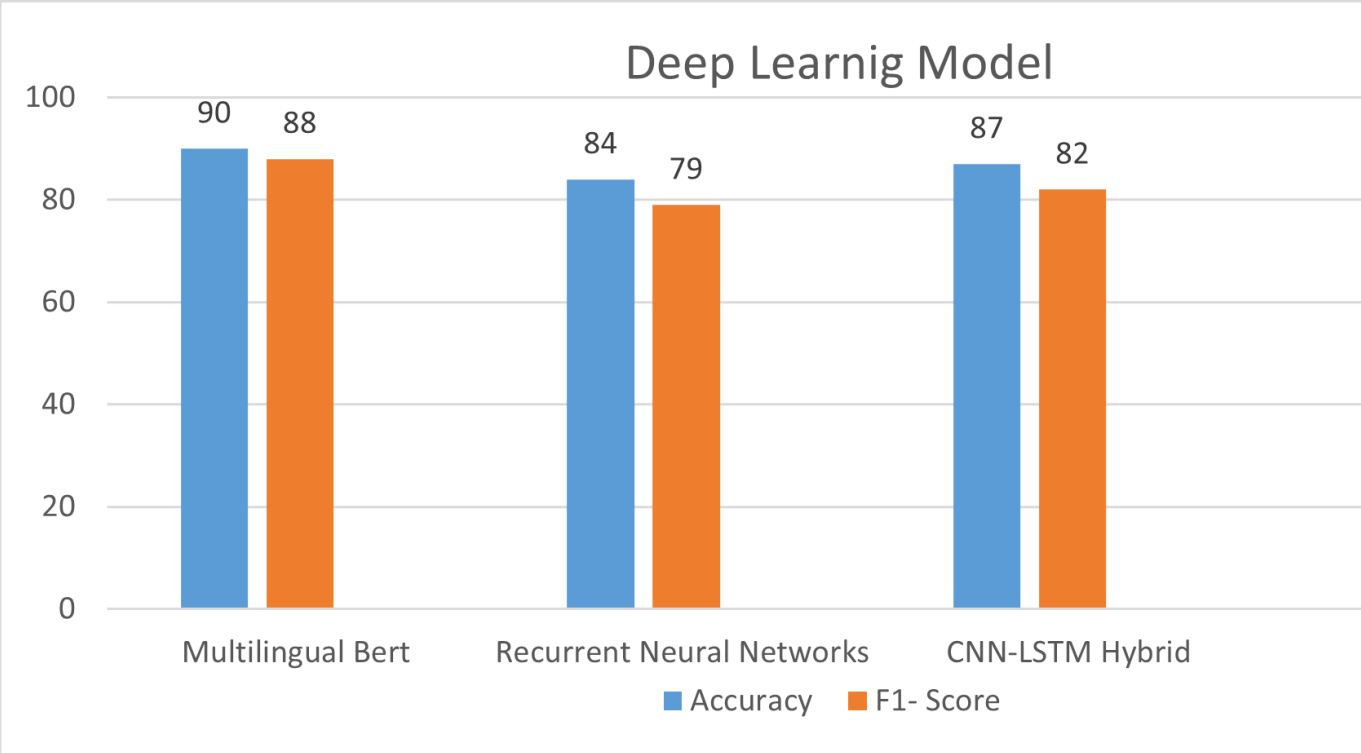

Figure 5.2: Deep learning for Binary Classification

	Accuracy	F1-Score
Multilingual Bert with Bert tokenizer	**0.9**	**0.88**
Recurrent Neural Networks with Keras tokenizer	0.84	0.79
CNN-LSTM Hybrid with Keras tokenizer	0.87	0.82

Table 5.2: Accuracy Table of DL for Binary Classification

From the above section we can observe that deep learning models like multilingual BERT and the hybrid of CNN-LSTM displayed better hate speech recognition performance then their traditional machine learning models. As deep learning models can capture the semantic meaning and context of words and phrases as well as utilize pre-trained word embedding to better understand the relationship between words, it can detect better hate speech pattern. Also deep learning models benefits from large datasets and the dataset we are using is also large it helps in hate speech detection by enabling them generalize better. Deep learning models, especially transformer-based architectures like BERT has better performance due to their pre-trained knowledge and we can see that in our study as it has 90% accuracy and 88% F1 score.

Multilevel Classification:

For Multilevel Classification we also use machine learning and deep learning model for best accuracy.
Same as binary classification here for machine learning model we use Multinomial Naive Bayes, Logistic Regression, Decision Tree Classifier, and the Random Forest Classifier.

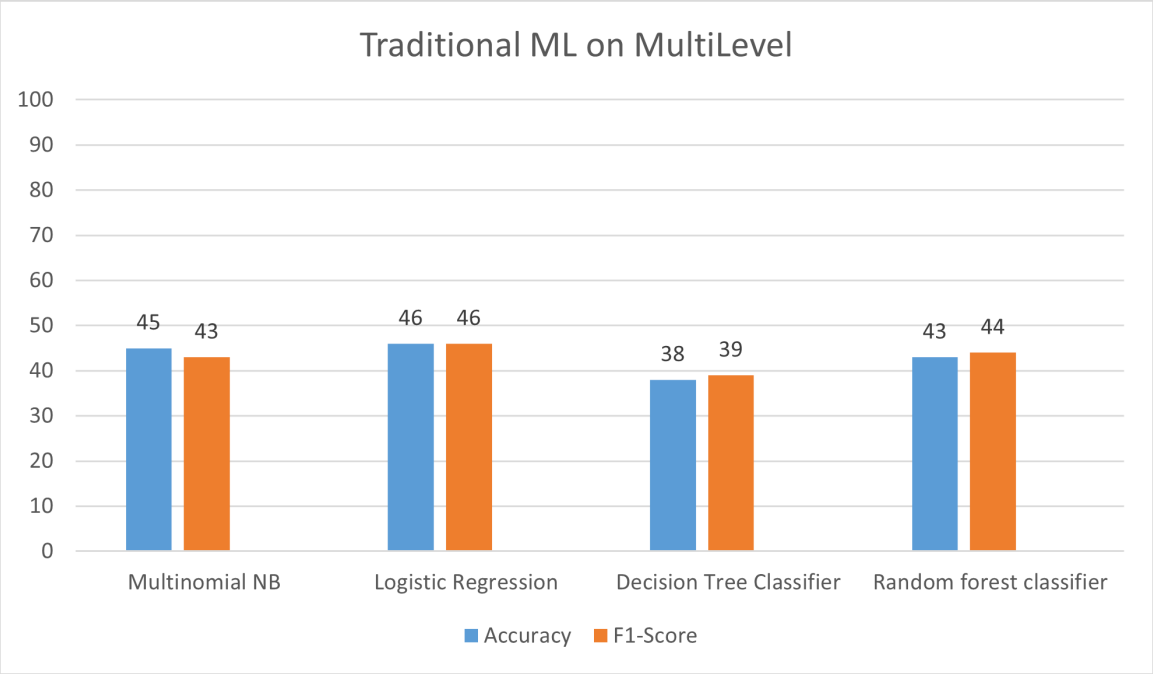

Figure 5.3: Machine Learning for Multilevel Classification

	Accuracy	F1-Score
Multinomial NB with TF-IDF Embedding	0.45	0.43
Logistic Regression with TF-IDF Embedding	0.46	0.46
Decision Tree Classifier with TF-IDF Embedding	0.38	0.39
Random forest classifier with TF-IDF Embedding	0.43	0.44

Table 5.3: Accuracy Table of ML for Multilevel Classification

And for Multilevel Classification in the Deep learning model we used multilingual BERT, Recurrent Neural Networks (RNN), and a hybrid model combining Convolutional Neural Networks (CNN) and Long Short-Term Memory (LSTM) networks.

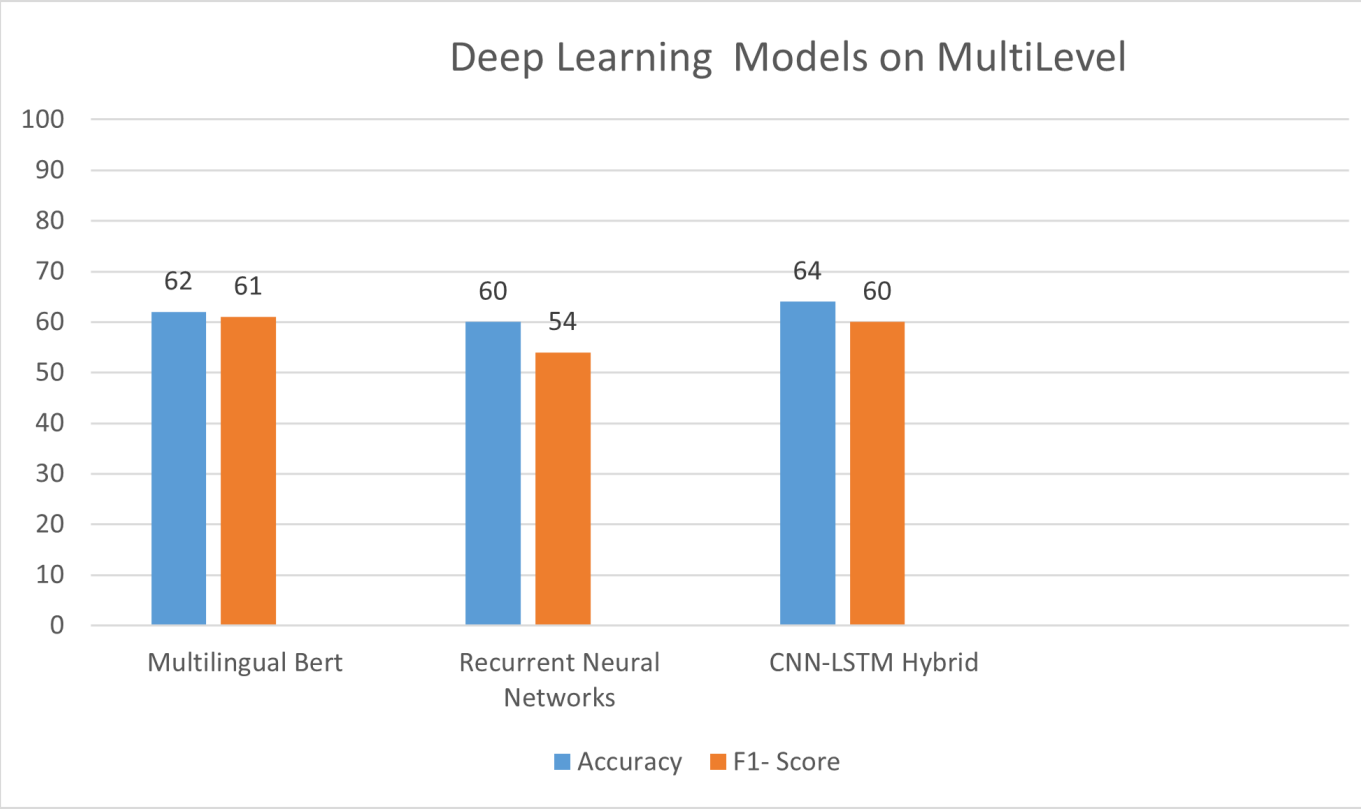

Figure 5.4: Deep learning for Binary Classification

	Accuracy	F1-Score
Multilingual Bert with Bert tokenizer	**0.62**	**0.61**
Recurrent Neural Networks with Keras tokenizer	0.6	0.54
CNN-LSTM Hybrid with Keras tokenizer	0.64	0.6

Table 5.4: Accuracy Table of DL for Multilevel Classification

Same as binary classification in the multilevel classification we can notice that deep learning models like multilingual BERT and the hybrid of CNN-LSTM displayed better hate speech recognition performance then their traditional machine learning models. As deep learning models can capture the semantic meaning and context of words and phrases as well as utilize pre-trained word embedding to better understand the relationship between words, it can detect better hate speech pattern. Also deep learning models benefits from large datasets and the dataset we are using is also large it helps in hate speech detection by enabling them generalize better. But unlike binary classification in this multilevel classification CNN-LSTM Hybrid gives us better accuracy due to it's memory support. As our dataset is large and aslo we are categorized them in multiple level here CNN helps to solve the text recognizing problem and LSTM sloves the memory problem. So CNN-LSTM model has better performance and we can see that in our study as it has 64% accuracy and 60% F1 score.

5.3 Performance Comparison

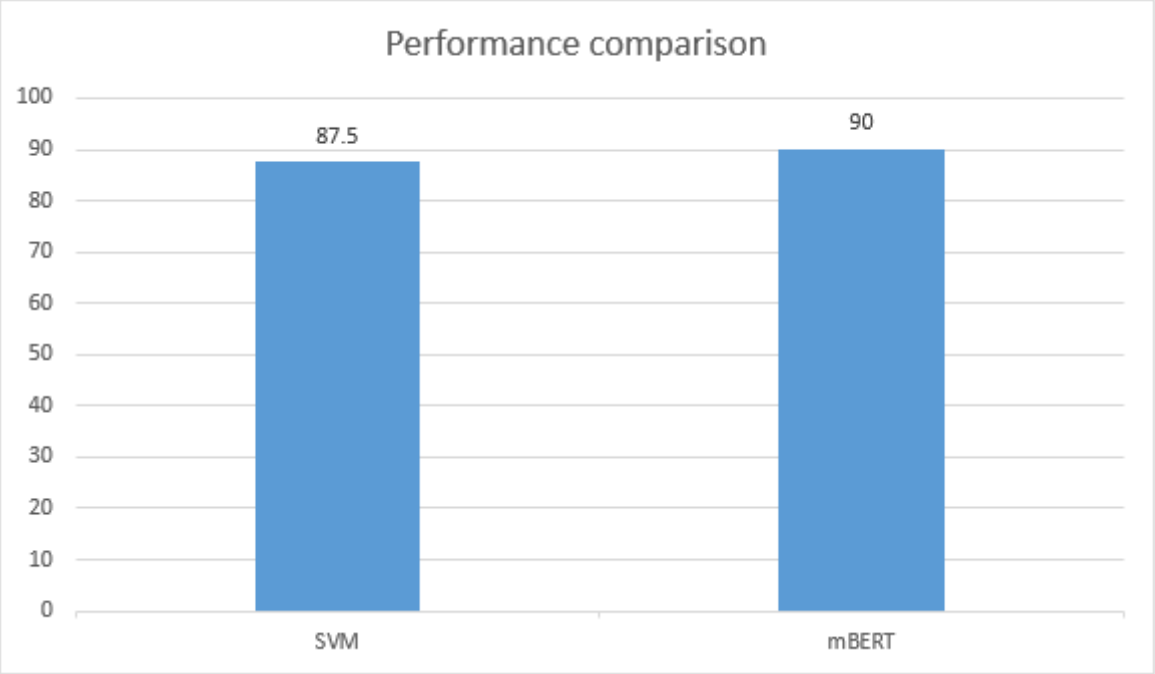

Figure 5.5: Comparison

For binary classification, the dataset we collected from [29] also ran some model like SVC, LSTM and Bi-LSTM on this dataset. So, this is just a basic comparison to check the performance between these two. [29] achieved the highest accuracy using the supervised model support vector machine (SVM).
In our study we got the highest accuracy using transformer-based model multilingual BERT with Bert tokenizer.

Multilevel Classification:

Model	Accuracy
CNN-LSTM Hybrid with Keras tokenizer	64.0

Table 5.5: Accuracy Table

Due to its greater complexity and difficulties differentiating between several hate speech categories, multilevel hate comment detection is more likely to be misclassified and frequently has worse accuracy than binary classification. For this reason, though not as high as binary classification, we got the highest accuracy in case of multilevel classification using a hybrid model of CNN and LSTM.

Chapter 6

Conclusion And Future Work

6.1 Conclusion

In this paper, we utilize Natural Language Processing to detect hate speech used by different social media users. We categorize it into two different levels, The first one is Binary classification (Hate speech and Not hate speech) and the other one is Multi level classification (Sports, Entertainment, Crime, Religion, Politics, Celebrity and Meme & TikTok). We chose Bengali as the language for our dataset. Although finding the dataset posed some challenges, as Bengali hate speech datasets are not widely available on dataset sites, we searched for alternative resources to gather more data. We then applied natural language processing and classification algorithms to classify our data. To determine accuracy, we ran our datasets through different algorithms. Specifically, we employed Multinomial Naive Bayes, Recurrent Neural Networks(RNN), Random Forest, Logistic Regression, Decision Tree Classifier, and CNN-LSTM Hybrid algorithms. We found the highest accuracy using the Multilingual Bert(mBert) for the binary classification and for the multi level classification we found the highest accuracy using CNN-LSTM Hybrid Model. However, as mentioned earlier, the definition of hate speech varies across different regions, making it challenging to detect the classes consistently. Nevertheless, the Multilingual Bert(mBert) model and CNN-LSTM model proves to be effective methods for automatically detecting hate speech. We will focus on further improving the accuracy of these models to identify hate speech more accurately.

6.2 Future Work

From the work we did, we achieved a quite high performance in detecting hate comments in binary classification but in multilevel classification we can see an accuracy drop. For future work regarding this study, exploring Large Language Model (LLMs) could be beneficial. Leveraging this kind of advance language model might enhance the systems understanding of context and subtle linguistic nuances in Bangla hate speech. Also, using a more extensive and diverse dataset and fine tuning model for cultural variations might result in more accurate hate speech detection. Furthermore, exploring multi-modal approaches to integrate textual and visual information might offer more understanding of hate speech.

Bibliography

[1] P. H. Swain and H. Hauska, "The decision tree classifier: Design and potential," *IEEE Transactions on Geoscience Electronics*, vol. 15, no. 3, pp. 142–147, 1977.

[2] E. Greevy and A. F. Smeaton, "Classifying racist texts using a support vector machine," in *Proceedings of the 27th annual international ACM SIGIR conference on Research and development in information retrieval*, 2004, pp. 468–469.

[3] M. P. LaValley, "Logistic regression," *Circulation*, vol. 117, no. 18, pp. 2395–2399, 2008.

[4] Y. Kim, "Convolutional neural networks for sentence classification," *arXiv preprint arXiv:1408.5882*, 2014.

[5] N. D. Gitari, Z. Zuping, H. Damien, and J. Long, "A lexicon-based approach for hate speech detection," *International Journal of Multimedia and Ubiquitous Engineering*, vol. 10, no. 4, pp. 215–230, 2015.

[6] A. Hern, "Facebook, youtube, twitter and microsoft sign eu hate speech code," *The Guardian*, vol. 31, 2016.

[7] L. Jiang, S. Wang, C. Li, and L. Zhang, "Structure extended multinomial naive bayes," *Information Sciences*, vol. 329, pp. 346–356, 2016.

[8] S. Tulkens, L. Hilte, E. Lodewyckx, B. Verhoeven, and W. Daelemans, "A dictionary-based approach to racism detection in dutch social media," *arXiv preprint arXiv:1608.08738*, 2016.

[9] J. Rosa and Y. Bonilla, "Deprovincializing trump, decolonizing diversity, and unsettling anthropology," *American Ethnologist*, vol. 44, no. 2, pp. 201–208, 2017.

[10] A. Travis, "Anti-muslim hate crime surges after manchester and london bridge attacks," *The Guardian*, vol. 20, 2017.

[11] S. Biere, S. Bhulai, and M. B. Analytics, "Hate speech detection using natural language processing techniques," *Master Business AnalyticsDepartment of Mathematics Faculty of Science*, 2018.

[12] J. Devlin, M.-W. Chang, K. Lee, and K. Toutanova, "Bert: Pre-training of deep bidirectional transformers for language understanding," *arXiv preprint arXiv:1810.04805*, 2018.

[13] Z. Zhang, D. Robinson, and J. Tepper, "Detecting hate speech on twitter using a convolution-gru based deep neural network," in *The Semantic Web: 15th International Conference, ESWC 2018, Heraklion, Crete, Greece, June 3–7, 2018, Proceedings 15*, Springer, 2018, pp. 745–760.

[14] E. A. Emon, S. Rahman, J. Banarjee, A. K. Das, and T. Mittra, "A deep learning approach to detect abusive bengali text," in *2019 7th International Conference on Smart Computing & Communications (ICSCC)*, IEEE, 2019, pp. 1–5.

[15] O. F. Rakib, S. Akter, M. A. Khan, A. K. Das, and K. M. Habibullah, "Bangla word prediction and sentence completion using gru: An extended version of rnn on n-gram language model," in *2019 International Conference on Sustainable Technologies for Industry 4.0 (STI)*, IEEE, 2019, pp. 1–6.

[16] V. Sanh, L. Debut, J. Chaumond, and T. Wolf, "Distilbert, a distilled version of bert: Smaller, faster, cheaper and lighter," *arXiv preprint arXiv:1910.01108*, 2019.

[17] C. Bakshi, *Random forest regression*, Medium, Jun. 2020. [Online]. Available: https://levelup.gitconnected.com/random-forest-regression-209c0f354c84.

[18] M. R. Karim, B. R. Chakravarthi, J. P. McCrae, and M. Cochez, "Classification benchmarks for under-resourced bengali language based on multichannel convolutional-lstm network," in *2020 IEEE 7th International Conference on Data Science and Advanced Analytics (DSAA)*, IEEE, 2020, pp. 390–399.

[19] M. Mozafari, R. Farahbakhsh, and N. Crespi, "A bert-based transfer learning approach for hate speech detection in online social media," in *Complex Networks and Their Applications VIII: Volume 1 Proceedings of the Eighth International Conference on Complex Networks and Their Applications COMPLEX NETWORKS 2019 8*, Springer, 2020, pp. 928–940.

[20] R. T. Mutanga, N. Naicker, and O. O. Olugbara, "Hate speech detection in twitter using transformer methods," *International Journal of Advanced Computer Science and Applications*, vol. 11, no. 9, 2020.

[21] A. Sherstinsky, "Fundamentals of recurrent neural network (rnn) and long short-term memory (lstm) network," *Physica D: Nonlinear Phenomena*, vol. 404, p. 132 306, 2020.

[22] A. K. Das, A. Al Asif, A. Paul, and M. N. Hossain, "Bangla hate speech detection on social media using attention-based recurrent neural network," *Journal of Intelligent Systems*, vol. 30, no. 1, pp. 578–591, 2021.

[23] C. Fan, M. Chen, X. Wang, J. Wang, and B. Huang, "A review on data preprocessing techniques toward efficient and reliable knowledge discovery from building operational data," *Frontiers in Energy Research*, vol. 9, p. 652 801, 2021.

[24] M. R. Karim, S. K. Dey, T. Islam, *et al.*, "Deephateexplainer: Explainable hate speech detection in under-resourced bengali language," in *2021 IEEE 8th International Conference on Data Science and Advanced Analytics (DSAA)*, IEEE, 2021, pp. 1–10.

[25] B. Mathew, P. Saha, S. M. Yimam, C. Biemann, P. Goyal, and A. Mukherjee, "Hatexplain: A benchmark dataset for explainable hate speech detection," in *Proceedings of the AAAI conference on artificial intelligence*, vol. 35, 2021, pp. 14 867–14 875.

[26] N. S. Mullah and W. M. N. W. Zainon, "Advances in machine learning algorithms for hate speech detection in social media: A review," *IEEE Access*, vol. 9, pp. 88 364–88 376, 2021.

[27] I. Papadimitriou, E. A. Chi, R. Futrell, and K. Mahowald, "Deep subjecthood: Higher-order grammatical features in multilingual bert," *arXiv preprint arXiv:2101.11043*, 2021.

[28] C. Paul and P. Bora, "Detecting hate speech using deep learning techniques," *International Journal of Advanced Computer Science and Applications (IJACSA)*, vol. 12, no. 2, pp. 619–623, 2021.

[29] N. Romim, M. Ahmed, H. Talukder, and M. Saiful Islam, "Hate speech detection in the bengali language: A dataset and its baseline evaluation," in *Proceedings of International Joint Conference on Advances in Computational Intelligence: IJCACI 2020*, Springer, 2021, pp. 457–468.

[30] B. Dean, *How many people use facebook in 2022?* Backlinko, Jan. 2022. [Online]. Available: https://backlinko.com/facebook-users.

[31] M. R. Karim, S. K. Dey, T. Islam, M. Shajalal, and B. R. Chakravarthi, "Multimodal hate speech detection from bengali memes and texts," in *International Conference on Speech and Language Technologies for Low-resource Languages*, Springer, 2022, pp. 293–308.

[32] A. Klappenbach, "The 12 most spoken languages in the world," *Retrieved Jan*, vol. 7, p. 2022, 2022.

[33] J. S. Malik, G. Pang, and A. v. d. Hengel, "Deep learning for hate speech detection: A comparative study," *arXiv preprint arXiv:2202.09517*, 2022.

[34] G. Pang, *Deep learning for hate speech detection: A large-scale empirical evaluation*, Medium, Mar. 2022. [Online]. Available: https://towardsdatascience.com/deep-learning-for-hate-speech-detection-a-large-scale-empirical-evaluation-92831ded6bb6.

[35] A. Rana and S. Jha, "Emotion based hate speech detection using multimodal learning," *arXiv preprint arXiv:2202.06218*, 2022.

[36] T. A. Belal, G. Shahariar, and M. H. Kabir, "Interpretable multi labeled bengali toxic comments classification using deep learning," in *2023 International Conference on Electrical, Computer and Communication Engineering (ECCE)*, IEEE, 2023, pp. 1–6.

[37] V. Hangya and A. Fraser, "How to solve few-shot abusive content detection using the data we actually have," *arXiv preprint arXiv:2305.14081*, 2023.

[38] M. S. Jahan and M. Oussalah, "A systematic review of hate speech automatic detection using natural language processing.," *Neurocomputing*, p. 126 232, 2023.

[39] U. Nations, *Why tackle hate speech?* United Nations. [Online]. Available: https://www.un.org/en/hate-speech/impact-and-prevention/why-tackle-hate-speech.