DETECTION AND PREVENTION OF CYBERBULLYING FROM SOCIAL MEDIA BY EXPLORING MACHINE LEARNING ALGORITHMS

SHUTONU MITRA

TASFIA TASNIM

MD ARR RAFI ISLAM

B.Sc. ENGINEERING THESIS

DEPT. OF COMPUTER SCIENCE & ENGINEERING

MILITARY INSTITUTE OF SCIENCE AND TECHNOLOGY

DHAKA, BANGLADESH

MARCH, 2022

DETECTION AND PREVENTION OF CYBERBULLYING FROM SOCIAL MEDIA BY EXPLORING MACHINE LEARNING ALGORITHMS

SHUTONU MITRA (STUDENT NO.201814033)

TASFIA TASNIM (STUDENT NO.201814029)

MD ARR RAFI ISLAM (STUDENT NO.201814050)

A Thesis Submitted in Partial Fulfillment of the Requirements for the Degree of Bachelor of Science in Computer Science and Engineering

DEPARTMENT OF COMPUTER SCIENCE AND ENGINEERING

MILITARY INSTITUTE OF SCIENCE AND TECHNOLOGY

DHAKA, BANGLADESH

MARCH, 2022

DETECTION AND PREVENTION OF CYBERBULLYING FROM SOCIAL MEDIA BY EXPLORING MACHINE LEARNING ALGORITHMS

B.Sc. Engineering Thesis

By

SHUTONU MITRA (STUDENT NO.201814033)
TASFIA TASNIM (STUDENT NO.201814029)
MD ARR RAFI ISLAM (STUDENT NO.201814050)

Approved as to style and content by the Examiners in March 2022:

Mohammad Shajahan Majib	Supervisor
Professor of Computer Science & Engineering	
MIST, Dhaka	
Nafiz Imtiaz Khan	Member
Lecturer of Computer Science Engineering	
MIST, Dhaka	
Brig Gen Md Abdur Razzak	Head of the Department
Professor of Computer Science & Engineering	
MIST, Dhaka	

Department of Computer Science & Engineering, MIST, Dhaka

DETECTION AND PREVENTION OF CYBERBULLYING FROM SOCIAL MEDIA BY EXPLORING MACHINE LEARNING ALGORITHMS

DECLARATION

We hereby declare that the study reported in this thesis entitled as above is our original work and has not been submitted before anywhere for any degree or other purposes. Further, we certify that the intellectual content of this thesis is the product of our own work and that all the assistance received in preparing this thesis and sources have been acknowledged and cited in the reference section.

Shutonu Mitra

Student No. 201814033

Tasfia Tasnim

Student No. 201814029

Md Arr Rafi Islam

Student No. 201814050

Department of Computer Science & Engineering, MIST, Dhaka

ACKNOWLEDGEMENTS

We are thankful to Almighty Allah for his blessings for the successful completion of our thesis. Our heartiest gratitude, profound indebtedness and deep respect go to our supervisor, Col Shajahan Majib, Instructor Class-A, Department of Computer Science & Engineering, Military Institute of Science and Technology, for his constant supervision, affectionate guidance and great encouragement and motivation. His keen interest in the topic and valuable advice throughout the study were of great help in completing the thesis. We are especially grateful to the Department of Computer Science and Engineering (CSE) of Military Institute of Science and Technology (MIST) for providing their all-out support during the thesis work. Finally, we would like to thank our families and our coursemates for their appreciable assistance, patience and suggestions during the course of our thesis.

ABSTRACT

Social media is the most popular way to meet new people and interact with friends and associates nowadays. But unfortunately, users get subject to bully or harassment while surfing through social media. Over the last decade, cyberbullying surfaced as one of the most significant issues in the digital world. Although significant research has been carried out to identify cyberbullying through text-mining techniques on many online platforms, still there is a long way to have a concrete solution to remove cyberbullying from social media. This paper introduces a way to identification of cyberbullies from social media (Twitter only) texts through sentiment analysis and classifies them according to bullying characteristics depending on the proposed taxonomy. For the identification and classification process, machine learning algorithms were explored and word embedding techniques were applied to publicly available Twitter Data. In this context, a suitable framework consisting of three modules (e.g., user interaction, analytics, and decision making) is proposed to prevent cyberbullying from social media. The user interaction module contains user profiles from where posts and comments are taken to the analytics module, the analytics module generates results according to the type of bully and the decision-making module takes action finally. Temporary/permanent ban on posting or commenting, a bully badge is shown on the personal profile are the actions proposed.

সারসংক্ষেপ

সোশ্যাল মিডিয়া হল নতুন লোকেদের সাথে দেখা করার এবং বন্ধু এবং সহযোগীদের সাথে যোগাযোগ করার সবচেয়ে জনপ্রিয় উপায় এখনকার দিনে। কিন্তু দুর্ভাগ্যবশত, সোশ্যাল মিডিয়ার মাধ্যমে ব্যবহার করার সময় ব্যবহারকারীরা ধমক বা হয়রানির শিকার হন। গত এক দশকে, সাইবার বুলিং ডিজিটাল বিশ্বের সবচেয়ে উল্লেখযোগ্য সমস্যাগুলির মধ্যে একটি হিসাবে আবির্ভূত হয়েছে। যদিও অনেক অনলাইন প্ল্যাটফর্মে টেক্সট-মাইনিং কৌশলের মাধ্যমে সাইবার বুলিং শনাক্ত করার জন্য উল্লেখযোগ্য গবেষণা করা হয়েছে, তবুও সোশ্যাল মিডিয়া থেকে সাইবার বুলিং অপসারণের জন্য একটি সুনির্দিষ্ট সমাধানের দীর্ঘ পথ রয়েছে। এই পেপারটি সেন্টিমেন্ট বিশ্লেষণের মাধ্যমে সামাজিক মিডিয়া (শুধুমাত্র টুইটার) পাঠ্য থেকে সাইবারবুলিদের সনাক্তকরণের একটি উপায় প্রবর্তন করে এবং প্রস্তাবিত শ্রেণীবিন্যাসের উপর নির্ভর করে ধমকানোর বৈশিষ্ট্য অনুসারে তাদের শ্রেণীবদ্ধ করে। সনাক্তকরণ এবং শ্রেণিবিন্যাস প্রক্রিয়ার জন্য, মেশিন লার্নিং অ্যালগরিদমগুলি অন্বেষণ করা হয়েছিল এবং সর্বজনীনভাবে উপলব্ধ টুইটার ডেটাতে শব্দ এম্বেডিং কৌশল প্রয়োগ করা হয়েছিল। এই প্রসঙ্গে, সোশ্যাল মিডিয়া থেকে সাইবার বুলিং প্রতিরোধ করার জন্য তিনটি মডিউল (যেমন, ব্যবহারকারীর মিথস্ক্রিয়া, বিশ্লেষণ এবং সিদ্ধান্ত গ্রহণ) সমন্বিত একটি উপযুক্ত কাঠামোর প্রস্তাব করা হয়েছে। ব্যবহারকারীর ইন্টারঅ্যাকশন মডিউলটিতে ব্যবহারকারীর প্রোফাইল থাকে যেখান থেকে পোস্ট এবং মন্তব্য বিশ্লেষণ মডিউলে নিয়ে যাওয়া হয়, বিশ্লেষণ মডিউল বুলির ধরন অনুযায়ী ফলাফল তৈরি করে এবং সিদ্ধান্ত নেওয়ার মডিউল অবশেষে পদক্ষেপ নেয়। পোস্ট করা বা মন্তব্য করার উপর অস্থায়ী/স্থায়ী নিষেধাজ্ঞা, ব্যক্তিগত প্রোফাইলে দেখানো বুলি ব্যাজ হল প্রস্তাবিত পদক্ষেপ।

ABBREVIATIONS

AI- Artificial Intelligence

NLP- Natural Language Processing

ML- Machine Learning

SM- Social Media

LR- Logistic Regression

SVM- Support Vector Machine.

RF- Random Forest

DT- Decision Tree

NN- Neural Network

RNN- Recurrent Neural Network

BiLSTM- Bi-directional Long Short Term Memory

TFIDF- Term Frequency Inverse Document Frequency

BoW- Bag of Words

GloVe- Global Vectors

IRL- Inverse Reinforcement Learning

TP-True Positive

TN-True Negative

FP-False Positive

FN-False Negative

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

LIST OF EQUATIONS

CHAPTER 1

INTRODUCTION

This chapter discusses the research background, problem statement, thesis objectives, methodological overview, thesis scope and organization of the thesis. Firstly it describes the background of the research to introduce the problem statement and to establish the thesis objectives. Then the methodological overview is illustrated followed by the thesis scope and organization of the thesis.

1.1 Research Background

From The advent of the internet and information technologies have revolutionized our way of communication, social life management, thinking, reasoning, etc. Especially in the past few years, online communication has migrated towards interactive social networking sites, blogs, mobile chat applications, etc. transcending all regional and spatial limitations through the internet. According to the Data reportal Global Overview report published on 27th January 2021, there are now 4.20 billion Social Media (SM) users around the world, which has grown by 490 million over the past 12 months, delivering year-on-year growth of more than 13 percent. The number of SM users is now equivalent to more than 53 percent of the world's total population (Digital Report,2021).

The SM characteristics, such as accessibility, flexibility, being free and having well-connected social networks, provide users with liberty and flexibility to post and write on their platforms in the form of ideas, opinions, preferences, views, and discussions are spread among users rapidly through online social communication. This has resulted in not only positive exchange of ideas but has also lead to widespread dissemination of aggressive and potentially harmful content over the and has given these incidents an unprecedented power and influence to affect the lives of billions of people.

1.2 Problem Statement

Cyberbullying can be considered as a distinct phenomenon or as a sub-form of bullying with electronic devices (Smith, 2008). A qualitative research study showed that students aged 10 to 18 define cyberbullying as bullying through modern technological devices which are intended to hurt, part of a repetitive pattern of negative actions, and performed in a relationship characterized by a power imbalance . A 2018 research study found that a majority of teens (59%) experienced some form of cyberbullying (Anderson,2018). Asian countries with the most cyberbullying done are China (70%), Singapore (58%) and India (53%) (Digital Report, 2021). Girls are more likely than boys to be both victims and perpetrators of cyberbullying. Only 6% of boys reported being bullied online, compared to 15% of girls, particularly older girls aged 12-17. 41% of the older girls reported experiencing some form of online harassment (Djuraskovic, 2021). Many government bodies and non-profit-healthcare organizations have highlighted the harmful effects of cyberbullying on the victims which include: depression, anxiety, reduced self-worth,

difficulty sleeping, eating disorders, etc. based on psychological surveys across different countries (EFFECTS OF CYBERBULLYING). Although the NCHS report, released in April 2020, does not suggest a reason for the increase in suicides, cyberbullying is indeed be part of the equation (Cook, 2020). One 2018 study found that young adults under the age of 25 who were victimized by cyberbullying were twice as likely to commit suicide or self-harm in other ways (John, 2018). So research on cyberbully detection on SM sites for its efficient monitoring and prevention is a high necessity in recent times.

Twitter is a micro-blogging, social networking website wherein users can write short 140 character messages called Tweets. As of February 2021, there were over a 192million daily active users on Twitter. Half a billion tweets are sent every day which equates to 5,787 tweets per second (Lin, 2021). Because of Twitter's huge volumes of active users, cyberbullying on Twitter is a global phenomenon. For the benefit of abundance and availability of data, Natural Language Processing and Machine Learning can be used on datasets collected from Twitter for cyberbullying and troll detection. Machine learning is an evolving branch of computational algorithms that are designed to emulate human intelligence by learning from the surrounding environment (El Naqa & Murphy, 2015). Natural Language Processing (NLP) is a subfield of Machine Learning that explores how computers can be used to understand and manipulate natural language text or speech to do useful things (Chowdhury, 2003). Cyberbullying is done in human language, thus it is obvious that cyberbully detection is more accurate when it is done with NLP. Thus the task of cyberbullying detection can be broadly defined as the use of machine learning techniques to automatically classify text in messages on bullying content or infer characteristic features based on higher-order information, such as user features or social network attributes.

1.3 Objectives

The aim of this research shall focus on the exploration of machine learning, neural network, and ensemble technique-based algorithms and pretrained text analyzer models for identification and classification of cyberbullies in social media contents which shall be done by sentiment analysis of our collected twitter dataset through natural language processing. An in-depth analysis of existing research works shall be conducted to evaluate our developed model and create a benchmark for model performance and accuracy. Later on, a framework shall be proposed with the aim of preventing cyberbullying attempts from social media using the already suggested bully identification and bully classification model. This thesis will cover the fields of Natural Language Processing, Machine Learning, and Deep Learning on a larger scale.

The specific aims of the research were as follows:

- To explore how ML algorithms are applied to any text to detect and categorize bully content.
- To measure the sentiment of any speech or text by using NLP techniques.
- To compare the performance of existing ML algorithms applied on SM for the detection of cyberbullying.

● To propose a framework for the prevention of trolls or bullies using a model built on the best performing ML algorithm on the collected dataset.

1.4 Methodological Overview

To achieve the above objectives, a literature review was conducted to identify the definition and nature of cyberbullying and its detection and prevention methods. Following that, multiple prediction models based on random forest (RF), logistic regression (LR), support vector machine (SVM), multinomial Naive Bayes (MNV) and recurrent neural network (RNN) were developed in order to determine the most efficient cyberbully detection model in terms of precision, recall, and f1 score. Finally, a framework has been proposed with a view to preventing bullying attempt on social media.

1.5 Scope of Thesis

The tentative scope of this research shall focus on the exploration of machine learning, neural network-based algorithms for the identification and classification of cyberbullies in social media content which shall be done by sentiment analysis of collected twitter dataset through natural language processing. An in-depth analysis of existing research works shall be conducted to evaluate our developed model and create a benchmark for model performance and accuracy. Lastly, a framework shall be proposed for the prevention of cyberbullying from SM sites.

1.6 Organization of Thesis

The book has been structured as follows. The next section presents a brief overview of the background and related works in the area of cyberbully detection. Section III details the conducted research methodology of the Twitter data for analysis purposes. Section IV provides an in-depth analysis and discussion of the obtained results from different ML models exploration. Section V provides a discussion on the findings and results of cyberbully detection by the adopted methods in this paper. A proposed framework for implementing a cyberbully protocol is presented in Section VI. The final section concludes the paper with discussions and limitations.

CHAPTER 2
LITERATURE REVIEW

This chapter briefly discusses some of the key concepts of theoretical knowledge. It gives a basic overview of natural language processing, machine learning and neural network. Then it explains the algorithms and natural language processing techniques used in this thesis, as well as related work done in recent years. This chapter came to an end with a critical summary.

2.1 Natural Language Processing

Human language is filled with ambiguities that make it incredibly difficult to write software that accurately determines the intended meaning of text or voice data. Homonyms, homophones, sarcasm, idioms, metaphors, grammar and usage exceptions, variations in sentence structure these just a few of the irregularities of human language that take humans years to learn, but that programmers must teach natural language-driven applications to recognize and understand accurately from the start if those applications are going to be useful. Natural language processing (NLP) refers to the branch of computer science and more specifically, the branch of artificial intelligence or AI concerned with giving computers the ability to understand the text and spoken words in much the same way human beings can. NLP combines computational linguistics—rule-based modeling of human language with statistical, machine learning, and deep learning models. Together, these technologies enable computers to process human language in the form of text or voice data and to 'understand' its full meaning, complete with the speaker or writer's intent and sentiment.

2.2 Machine Learning

Some Machine learning is a branch of artificial intelligence (AI) and computer science which focuses on the use of data and algorithms to imitate the way that humans learn, gradually improving its accuracy. Here, the programmers don't need to code everything rather the machine can discover the rules by itself by exploring the training data. Machine learning can be approached in a variety of ways, but they are generally divided into three categories. Supervised learning is mainly used for binary classification, multi-class classification, Regression modeling and ensembling modeling. The unsupervised method of learning is used for finding unusual patterns or anomaly detection(identifying unusual data points in a data set), association mining (identifying sets of items in a data set that frequently occur together) and in many other cases. Reinforcement learning is most often used to teach a program how to complete a multi-step task with well-defined rules.

2.3 Neural Network

A neural network is a series of algorithms that endeavors to recognize underlying relationships in a set of data through a process that mimics the way the human brain operates. In this sense, neural networks refer to systems of neurons, either organic or artificial in nature. Neural networks can adapt to changing input; so the network generates the best possible result without needing to redesign the output criteria. There are three main components: an input layer, a processing layer, and an output layer. The inputs may be weighted based on various criteria. Within the processing layer, which is hidden from view, there are nodes and connections between these nodes, meant to be analogous to the neurons and synapses in an animal brain. The architecture of a neural network is depicted in Fig. 2.1.

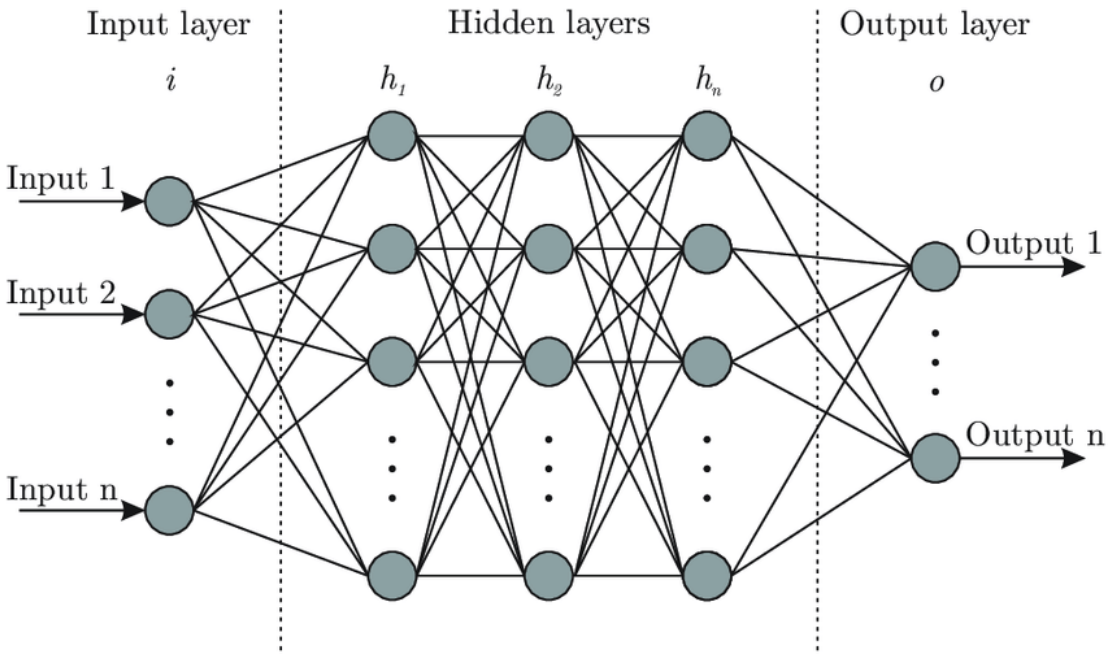


Fig. 2.1: Architecture of Neural Network

2.4 Machine Learning Algorithms

A machine learning algorithm is a set of instructions that tells the computer how to accomplish a specific task. There are a variety of machine learning algorithms available, each with a different level of accuracy based on the data set's characteristics. Some algorithms are for supervised learning, and others are for unsupervised learning. We'll briefly discuss some of the machine learning algorithms that were used to train our data set in this segment.

2.4.1 Liblinear Based Logistic Regression

Logistic regression is the appropriate regression analysis to conduct when the dependent variable is dichotomous (binary). Like all regression analyses, logistic regression is a predictive analysis. Logistic regression is used to describe data and to explain the relationship between one dependent binary variable and one or more nominal, ordinal, interval or ratio-level independent variables. Logistic regression essentially uses a logistic function defined below to model a binary output variable. In the logistic function equation, x is the input variable.

$$Logistic\ function\ = \frac{1}{1+e^{-x}} \qquad (2.1)$$

The logistic regression model has its basis in the odds of a 2-level outcome of interest. It takes the natural logarithm of the odds as a regression function of the predictors. With 1 predictor, X, this takes the form $ln[(Y = 1)] = \beta_0 + \beta_1 X$ where ln stands for the natural logarithm, Y is the outcome and Y=1 when the event happens (versus Y=0 when it does not), β_0 is the intercept term, and β_1 represents the regression coefficient, the change in the logarithm of the odds of the event with a 1-unit change in the predictor X. The difference in the logarithms of 2 values is equal to the logarithm of the ratio of the 2 values, so by taking the exponential of β_1, we obtain the ratio of the odds (the odds ratio) corresponding to a 1-unit change in (X Wright, 1995). LIBLINEAR supports two popular binary linear classifiers: Linear Regression (LR) and Linear Support Vector Machine (SVM). Given a set of instance-label pairs $(xi, yi), I = 1, l, xi\ \varepsilon\ Rn, yi\ \varepsilon\ -1, +1$, both methods solve the following unconstrained optimization problem with different loss functions ξ(w; xi, yi):

$$min(w)\ =\ \frac{1}{2}\ w^T w\ +\ C\ \sum_{i=1}^{l} \xi(w;\ xi,\ yi) \qquad (2.2)$$

where C > 0 is a penalty parameter. For SVM, the two common loss functions $max\ (1 + yi\ w^T\ xi,\ 0)$ and $max(1 - yi\ w^T\ xi,\ 0)$. The former is referred to as L1-SVM, while the latter is L2-SVM. For LR, the loss function is $log(1 + e^k yi\ w^T\ xi\ )$, which is derived from a probabilistic model (Fan et al., 2008). For multi-class problems, we implement the one-vs-the-rest strategy.

2.4.2 Naïve Bayes

Naive Bayes is a classification technique based on Bayes' Theorem with an assumption of independence among predictors. In simple terms, a Naive Bayes classifier assumes that the presence of a particular feature in a class is unrelated to the presence of any other feature. Bayesian classifiers assign the most likely class to a given example described by its feature vector. Learning such classifiers can be greatly simplified by assuming that features are independent given class.The following equation is the formula for Bayes' Theorem:

$$P\,(C) = \sum_{i=0}^{n} P\,(Xi|C) \qquad (2.3)$$

Here Xi = (X1, ..., Xn) is a feature vector and C is a class. Despite this unrealistic assumption, the resulting classifier known as Naive Bayes is remarkably successful in practice, often competing with much more sophisticated techniques. Naive Bayes has proven effective in many practical applications, including text classification, medical diagnosis, and systems performance management (Rish et al.,2001). Multinomial Naive Bayes Classifier is a supervised learning method that uses probability and is focused on text classification cases. This method follows the principle of multinomial distribution in conditional probability. Although using multinomial distributions, this algorithm can be applied to test cases by converting to a nominal form that can be computed with an integer value. The probability calculation is described in the equation:

$$P\,(d) \propto P\,(C) \prod_{1<k<nd} P(c) \qquad (2.4)$$

where P (tk|c) the conditional probability of the word tk at appears in the document having class c. In the equation P (tk|c) is the likelihood probability of tk in class c. While P (c) is the prior probability of the document appearing in class c. The class determination is to compare the posterior probability results obtained, then the class with the largest posterior probability is the class chosen as the predicted result (Kibriya et al., 2004).

2.4.2 Random Forest

RF is a supervised learning algorithm. The forest it builds is an ensemble of decision trees, usually trained with the bagging method. The general idea of the bagging method is that a combination of learning models increases the overall result (Niklas, 2019). RF consists of a large number of individual decision trees that help get a more accurate and stable prediction. Significant improvements in classification accuracy have resulted from growing an ensemble of trees and letting them vote for the most popular class. A graphical illustration of majority voting by RF is shown in Fig. 2.2. In order to grow these ensembles, often random vectors are generated that govern the growth of each tree in the ensemble (Breiman, 2001). Each individual tree in the random forest spits out a class prediction and the class with the most votes becomes the model's prediction. An important prerequisite needed for ensuring better accuracy in RF is that the trees of the forest and more importantly their predictions need to be uncorrelated (or at least have low correlations with each other). RF ensures that the behavior of each individual tree is not too correlated with the behavior of any of the other trees in the model by following two methods- bagging, feature randomness (Tony, 2019).

•**Bagging:** DTs are very sensitive to the data they are trained on. Small changes to the training set can result in significantly different tree structures. RF takes advantage of this by allowing each individual tree to randomly sample from the data set with replacement, resulting in different trees. This process is known as bagging.

•**Feature Randomness:** In a normal DT, when it is time to split a node, every possible feature is considered and the one that produces the most separation between the observations in the left node vs. those in the right node is picked. In contrast, each tree in

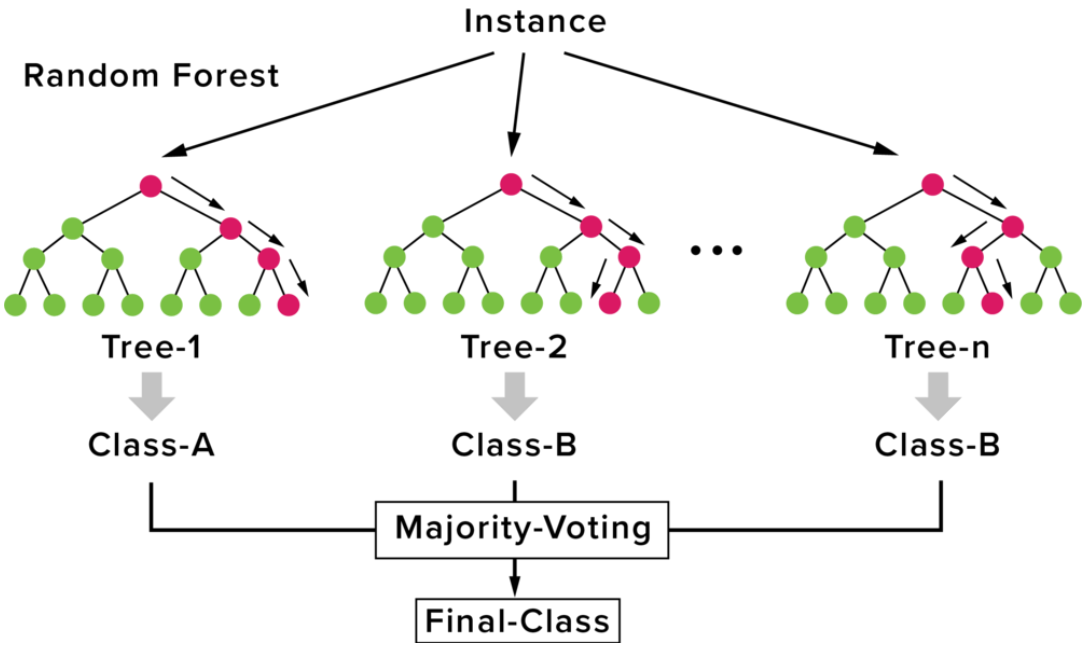


Fig. 2.2: Graphical illustration of majority voting by RF

RF can pick only from a random subset of features. This forces even more variation amongst the trees in the model and ultimately results in lower correlation across trees and more diversification.

This diversity or less correlation in the trees of RF helps to get better performance.

2.4.4 Linear Support Vector Machine

SVM (Support Vector Machine) is a supervised machine learning algorithm that can be used to solve classification and regression problems. It is, however, mostly used to solve classification problems. In the SVM algorithm, we plot each data item as a point in n-dimensional space (where n is a number of features you have) with the value of each feature being the value of a particular coordinate. A graphical presentation of the one-dimensional SVM model is shown in Fig. 2.3. Then, classification is performed by finding the hyper-plane that differentiates the two classes very well (Christmann & Steinwart, 2008).

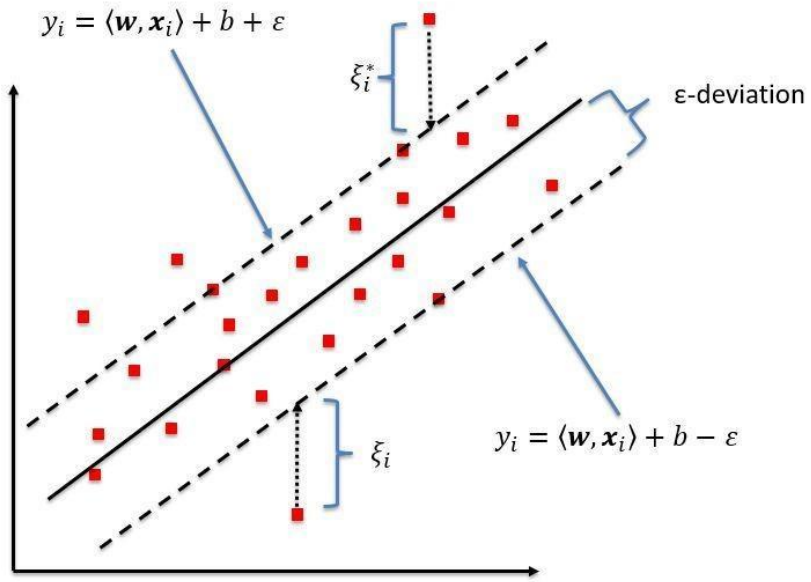


Fig. 2.3: Graphical presentation of one-dimensional SVM model

2.4.5 Bidirectional LSTM Recurrent Neural Network

A Recurrent Neural Network is a generalization of a feedforward neural network having an internal memory for processing sequence. RNNs are recurrent in nature as it performs the same function for every input of data while the output of the current input depends on the past computation. A variation of RNN is LSTM also known as the Long Short Term Memory is an RNN architecture with feedback connections, resolving the vanishing gradient problem of RNN. Long Short Term Memory networks – usually just called "LSTMs"– are a special kind of RNN, capable of learning long-term dependencies. LSTMs are explicitly designed to avoid the long-term dependency problem. Remembering information for long periods of time is practically their default behavior, not something they struggle to learn. The architecture of RNN is shown in Fig. 2.4. However, the bidirectional recurrent neural networks (RNN) are really just putting two independent RNNs together. The input sequence is fed in normal time order for one network, and in reverse time order for another. The outputs of the two networks are usually concatenated at each time step.

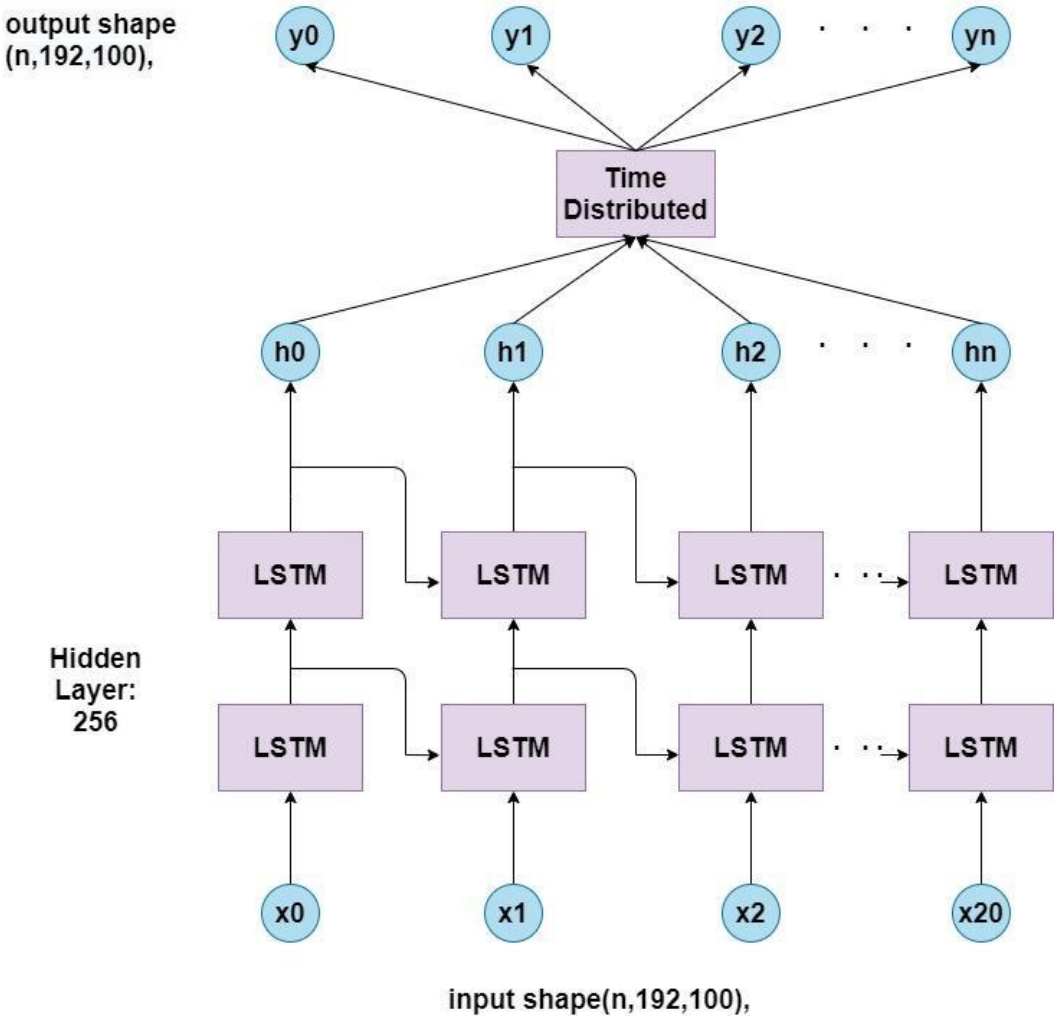


Fig. 2.4: Architecture of RNN

2.5 Natural Language Processing Techniques

Feature extraction and pre-processing are crucial NLP techniques used for text classification applications. In this section, we introduce methods for cleaning text data sets, thus removing implicit noise and allowing information for informative featurization. Furthermore, we discuss two common methods of text feature extraction: Weighted word and word embedding techniques.

2.5.1 Text Preprocessing

This most text and document data sets contain many unnecessary words such as stop words, misspellings, punctuation, etc. In many algorithms, especially statistical and probabilistic learning algorithms, noise and unnecessary features can have adverse effects on system performance. The text pre-processing used for our collected dataset (Tweets) is described in brief below:

•**Tokenization:** Tokenization is a pre-processing method that breaks a stream of text into words, phrases, symbols, or other meaningful elements called tokens. The main goal of this step is the investigation of the words in a sentence. Both text classification and text mining require a parser that processes the tokenization of the documents.

•**Lower-casing:** Since documents consist of many sentences, diverse capitalization can be hugely problematic when classifying large documents. The most common approach for dealing with inconsistent capitalization is to reduce every letter to lower case.

•**Contraction and Abbreviation handling:** Contraction and abbreviation are other forms of text anomalies that are handled in the pre-processing step. An abbreviation is a shortened form of a word or phrase which contains mostly first letters from the words, such as SVM which stands for Support Vector Machine. And a contraction is a shortened form of a word (or group of words) that omits certain letters or sounds. In most contractions, an apostrophe represents the missing letters. The most common contractions are made up of verbs, auxiliaries, or modals attached to other words: He would=He'd. I have=I've.

•**Stemming:** In NLP, one word could appear in different forms (i.e., singular and plural noun forms) while the semantic meaning of each form is the same. One method for consolidating different forms of a word into the same feature space is stemming. Text stemming modifies words to obtain variant word forms using different linguistic processes such as affixation (addition of affixes). For example, the stem of the word "studying" is "study".

•**Noise Reduction:** Most of the text and document data sets contain many unnecessary characters such as punctuation, special characters, on-English words, Html characters. Critical punctuation and special characters are important for the human understanding of documents, but they can be detrimental to classification algorithms.

2.5.2 Text Feature Extraction

Text feature extraction is the process of taking out a list of words from the text data and then transforming them into a feature set that is usable by a classifier. The two popular types of feature extraction used in the study are described below:

The most basic form of weighted word feature extraction is Term Frequency, where each word is mapped to a number corresponding to the number of occurrences of that word in the whole corpora. Methods that extend the results of TF generally use word frequency as a Boolean or logarithmically scaled weighting. In all weight words methods, each document is translated to a vector (with length equal to that of the document) containing the frequency of the words in that document. Although this approach is intuitive, it is limited by the fact that particular words that are commonly used in the language may dominate such representations. The weighted words techniques used in the study are described below:

•**Bag of Words**: The bag-of-words model (BoW model) is a reduced and simplified representation of a text document from selected parts of the text, based on specific criteria, such as word frequency. In a BoW, a body of text, such as a document or a sentence, is thought of as a bag of words. Lists of words are created in the BoW process. These words in a matrix are not sentences that structure sentences and grammar, and the semantic relationship between these words is ignored in their collection and construction. The words are often representative of the content of a sentence. While grammar and order of appearance are ignored, multiplicity is counted and may be used later to determine the focus points of the documents.

•**Term Frequency- Inverse Document Frequency:** Inverse Document Frequency (IDF) assigns a higher weight to words with either high or low frequencies terms in the document. This combination of TF and IDF is well known as Term Frequency-Inverse document frequency (TF-IDF). The mathematical representation of the weight of a term in a document by TF-IDF is given in the equation:

$$W(d, t) = TF(d, t) * log(N/df(t)) \quad (2.5)$$

Here N is the number of documents and df (t) is the number of documents containing the term t in the corpus. The first term in the equation improves the recall while the second term improves the precision of the word embedding Tokunaga & Makoto (1994).

Word embedding is a feature learning technique in which each word or phrase from the vocabulary is mapped to an N dimension vector of real numbers. Various word embedding methods have been proposed to translate unigrams into understandable input for machine learning algorithms. A powerful word embedding technique that has been used in this study is Global Vectors (GloVe). The approach is to present each word by a high dimension vector and train based on the surrounding words over a huge corpus. The

pre-trained word embedding used in many works is based on 400,000 vocabularies trained over Wikipedia 2014 and Gigaword 5 as the corpus and 50 dimensions for word presentation. GloVe also provides other pre-trained word vectorizations with 100, 200, 300 dimensions which are trained over even bigger corpora, including Twitter content.

2.6 Related Works

The detection of online cyberbullying has seen an increase in societal importance, popularity in research, and available open data. Researchers have been trying to address the issue of cyberbullying using text-mining techniques for over a decade. The most common means of constructing cyberbullying detection models is to use a text classification approach that involves the construction of machine learning classifiers from labeled text instances collected from social media posts, comments, tweets, etc. Reviewing some of the recent research works shall shed some light on this context.

2.6.1 Related Works Concerning Cyberbully Detection

Hee et al (Van Hee et al., 2018) conducted a study on the automatic detection of cyberbullying in SM text where linear support vector machines (SVM) were exploited. The classifier yielded an F1 score of 64% and 61% for English and Dutch cyberbully-related posts, respectively. Luceri et al (Luceri et al. ,2020) engaged in to study on Inverse Reinforcement Learning (IRL) which is a case study of Russian trolls in the 2016 US election done in June 2020 where IRL was used to capture troll behavior and troll accounts with Area under the Curve (AUC) of 89.1%. Additionally, a study was done by Saravanaraj et al (Saravanaraj et al., 2016) on the automatic detection of cyberbullying from Twitter where Naive Bayes and Random Forest classifiers were used to detect cyberbullying related posts and rumors were detected by using type and topic-specific classification and Twitter speech-act classifier. Moreover, Mihaylov and Nakov (Mihaylov & Nakov,2019) studied hunting for troll comments in news forums. In this study, an L2-regularized logistic regression with LIBLINEAR is trained to build two classifiers that can distinguish paid trolls and non-trolls with 81-82% accuracy. Weller and Woo (Weller & Woo, 2019) did a study on Identifying Russian trolls on Reddit with deep learning and BERT word embedding. The study proposes a three-layer neural network architecture; their best model contains a Recurrent Convolutional Neural Network (RCNN) that outperforms current ML practices with an AUC of 84.6%.

Another study has also been carried out by Cheng et al (Cheng et al.,2019b) on the novel problem of cyberbullying detection within a multi-modal context (e.g., image, video, user profile, time, and location) by exploiting SM data in a collaborative way rather than depending only on textual data. The proposed framework XBully, first reformulated multi-modal social media data as a heterogeneous network and then aimed to learn node embed-ding representations upon it, which outperformed effectiveness, timeliness, and scalability over existing traditional methods through experimentation on the Instagram dataset. Cheng et al (Cheng et al. 2019a) proposed a cyberbully framework named

PiBully which is able to model peer influence in a collaborative environment and Taylor cyberbullying prediction for each individual user. (Ge et al Ge et al.,2021) proposed a principled graph-based approach to model the temporal dynamics and topic coherence throughout user interaction by constructing a unified temporal graph for each social media session. Sheeba and Devaneyan (Sheeba & Pradeep Devaneyan, 2016) proposed a framework to detect cyberbullying content using the GenLeven algorithm. to classify the type of bully as harassment, insult, terrorism, or flaming fuzzy rule base was used in this framework.

In a study by Anzovino ,Fersini and Rosso (Anzovino et al., 2018) automatic identification and classification of misogynistic language on Twitter proposed a taxonomy for modeling the misogyny phenomenon in online social environments with a Twitter dataset manually labeled. According to them, the misogyny detection might be considered as a special case of abusive language and therefore, they chose representative features such as. . . N-grams, linguistic, syntactic, embedding and Linear Support Vector Machine (SVM), Random Forest (RF), Naive Bayes (NB) and Multilayer Perceptron Neural Network (MPNN) were most effective in text categorization. Additionally, Maruf, Ajwad and Ashrafi (Al Marouf et al., 2019) proposed a framework for detecting character assassination via troll comments on social media using psycholinguistic tools depicting the online behavior from one's activities performed online by tracing digital footprints. The novel framework includes identifying the troll behind the mask by detecting troll comments from public posts utilizing psycholinguistic tools such as LIWC and SlangNet for extracting psycholinguistic features and slang words, respectively. The works on cyberbully detection are shown in tabular format in appendix A.

2.6.2 Related Works Concerning Cyberbully Prevention

Some studies have also been carried out with the aim of preventing and eliminating cyberbullying by applying the proposed detection method of bullying. A study was done by Sugandhi et al (Sugandhi et al., 2016) on automatic prevention and monitoring of cyber-bullying designed a response grading system that maps an appropriate response for each cyberbully class taking into account the various parameters like the present social and political scenario, the severity, the overall sentiment against a particular issue etc. The sentiment analyzer and a bully classifier were built on multi-class SVM classifiers which use the Lin-ear SVC kernel with an accuracy of 61.88%. A study conducted by Prime and Suri (Parime & Suri, 2014) on cyberbullying detection and prevention from a data mining and psychological perspective stated the implementation of linear SVM (Support Vector Machine) with word vector and stemming process for sentiment analysis and bully measurement along with prevention measures such alerting SM site admin. Another study conducted by Yao et al (Yao et al., 2019) with the aim of eliminating of cyberbully focused on robust detection through a novel two-stage online approach CONCISE, designed for timely and accurate Cyberbullying detection on Instagram media Sessions along with a sequential hypothesis testing formulation that seeks to drastically reduce the number of features used in classifying each comment while maintaining high classification accuracy. CONCISE raises an alert only after a certain number of detections

have been made. The works on cyberbully prevention are shown in tabular format in appendix B.

2.7 Critical Summary

Much of the recent research uses small, heterogeneous datasets, without a thorough evaluation of applicability. For our research works a dataset shall be used to detect the bully, using the most recent tweets from the internet. Although there is huge work on bully detection, it was not done from our country's perspective. We aim to target events in our country along with international contexts and work with the most recent dataset collected from tweets. Most cyberbully detection models can only detect cyberbullying; they cannot classify the type of bully. The reason behind this could be the annotation scheme of data.

CHAPTER 3

METHODOLOGY

The research was carried out in two phases. After defining objectives, the first step was devoted to finding the best performing cyberbully detection model using collected twitter data. Several machine learning models were analyzed and their performances were evaluated in this phase. A framework for the prevention of cyberbullying from social media using the detection model had been proposed in the later phase. An overview of the research methodology is depicted in Fig. 3.1.

In Phase I, related articles were selected after a brief literature review described in Chapter 2. Following this, the dataset was collected from Twitter and annotated by using defined taxonomy. Then machine learning models were developed on pre-processed data and their performances were compared in terms of precision, recall and f1 score. The results of this step revealed which algorithm performed better at detecting cyberbullying.

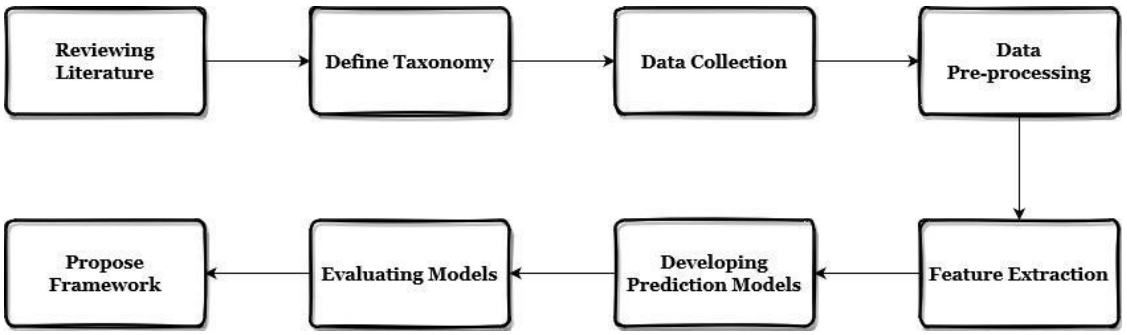


Fig-3.1: Research Methodology

CHAPTER 4

DEVELOPING MACHINE LEARNING MODELS

The ML models were built in the following stages: Data Collection, Data Preprocessing, Feature Extraction, Developing the Prediction Models and Evaluating the Prediction Models. The phases are briefly discussed in the following subsections:

4.1 Data acquisition

The study is done only on English tweets. The public tweets related to cyberbullying are collected using a publicly available application programming interface (API) worldwide. The collected cyberbully-related tweets were posted on Twitter between January 1, 2020, and May 31, 2021. This study used tweets collected from the internet with a limitation of 10000. The data labeling subsection contains the annotation process. The API we used works as a search machine in the tweeter and collects the search result in a CSV form. The tweets collected for the dataset contain the time of posting the tweet and the tweet. The API used here considers tweet ID as a primary key so that one tweet does not occur multiple times.

4.2 Data Pre-processing

Data preprocessing is a data mining technique that involves performing the necessary data transformation and cleaning the collected dataset to make the raw dataset in an understandable format. The extracted tweets are in unstructured format due to URLs, mentions, hashtags, emojis, smileys, signs, grammatical errors, and abbreviations used by Twitter users. Thus, data preprocessing is required to eliminate all special characters and unstructured forms before training the ML classifiers. The acquired data is pre-processed through the cleaning of text data by removing username, URL, emojis, etc. using a python library tweet-preprocessor, filtering non-English words and Html markups, conversion of tweets to lowercase characters, and removing tabs, and spaces. The stop words like “a”, “an”, “the” are removed using python’s gensim library as they carry little meaning in a sentence. The stemming and tokenization have also been done to get them to a normalized state. Thus, the study is conducted on preprocessed English tweets to decrease the size of the feature set to make it suitable for efficient learning for the algorithms used in cyberbully detection purposes. An overview of the steps is given in Fig. 4.1.

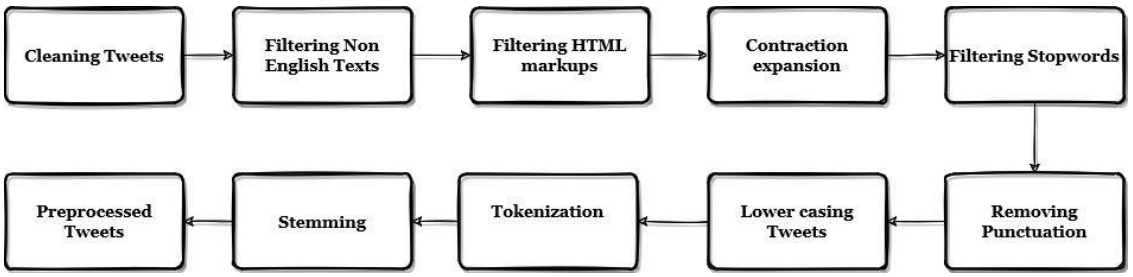


Fig. 4.1: Tweet Preprocessing Steps

4.3 Data Annotation

From the collected dataset two corpora have been annotated for cyberbully detection and classification. For the detection of cyberbullying, the corpora have been annotated using the Vader lexicon. The sentiment analysis tool Vader (Valance Aware Dictionary and Sentiment Reasoner) is a rule-based lexicon especially attuned to sentiments expressed as texts in social media sites in which the list of lexical features such as words, punctuation marks emoticons, etc. labeled according to their semantic orientation as either positive or negative sentiment (Hutto & Gilbert, 2014). This tool analyses textual data and returns a dictionary of scores referring to the intensity of positivity, negativity, neutralism, and the compound score of the three. Since a bully in social media mostly expresses negative sentiment towards its victims, sensing negativity from a text posted on social media sites would be enough to identify the bullying intention of the user. Thus, for the bully detection model, the dataset corpora have been labeled as positive sentimental tweets and negative sentimental tweets through analyzing the compound score of the Vader sentiment analyzer.

Judging on the negativity from the text, identification of the bully wouldn't be so accurate, hence comes the need for a bully classification model which would be able to identify a bully according to its bullying characteristics. Hence a taxonomy has been designed which describes some specific textual categories related to cyberbullying, which include racial, offensive, misogynistic, and attacking comments targeted against the victims. All of these forms were inspired by social studies on cyberbullying and manual inspection of cyberbullying examples. The designed taxonomy is stated below:

- Racism: The tweets show the sentiment of attacking a particular race or showing disrespect by name-callings such as "nigga" and similar words.
- Offensive: The tweets containing" fuck you", "badass" and so on words that attack someone is labeled as offensive.
- Misogyny: These are the tweets focusing on attacking women. The commonly occurring words are "hoe", "feminism" with negative sentiment, "bitch", and so on.
- Attacking: Some tweets indirectly attack a community. Those will be labeled as attacking. This type of tweet usually contains words like lame, useless, or words that criticize a community.

In addition to the tweets with negative sentiments annotated as racism, attacking, etc. labels some tweets with neutral and positive sentiments also remained in the corpora for which the taxonomy designed has been extended to the following labels:

- Positive: The tweets with positive sentiments containing “good”, “very good”, “wonderful”, “thank you”, “well done” and similar positive words are marked as positive.
- Neutral: These tweets show neutral sentiment. Some words that contain are similar to “good”, “bad” or the tweets which represent news that has no positive or negative or other keywords that are going to be defined in the following part.

For the bully classification model, two anonymous annotators have annotated the same corpora independently according to the designed taxonomy. To measure how well the annotators made the same annotation decision for a certain category, the inter-annotator agreement has been measured using Cohen's kappa score for pairwise annotators. The computed kappa score is around 0.83 which states the agreement between the annotators is almost perfect.

4.4 Feature Extraction

Numerical representations are a prerequisite for most machine learning models as algorithms learn to approximate functions that map inputs to outputs for developing the model. Embedding methods encode symbolic representations of words, emojis, categorical items, dates, times, other features, etc into meaningful numbers in order to extract the features from textual data through vectorizing or encoding the texts to be fed into the model. For this study embeddings like Term Frequency Inverse Document Frequency (TFIDF), Bag of Words (BoW), GloVe pre-trained word embedding on the Twitter dataset have been used for extracting feature vectors from the preprocessed tweets.

The bully identification model shall be able to classify text in two classes: positive sentiment and negative sentiment. In order to find the best detector model of bullying text different machine learning algorithms were chosen based on recent work relating to the detection of cyberbullying from textual data. On the contrary for bully classification, models have been developed using machine learning algorithms, neural networks, and pre-trained models for text analyses. The classification model classifies tweets in six classes: racism, offensive, attacking, misogyny, neutral, positive. In both cases, a random train test split of 70-30 was done, where 70% data was considered as the training dataset, while 30\% data was considered as the test dataset. The models were developed using Scikit-learn, Tensorflow, and Pytorch python modules. Models of conventional machine learning algorithms such as Logistic Regression, Multinomial Naive Bayes, Random Forest, Linear SVM have been built on the vectorized dataset using TFIDF, BoW. GloVe embedding layer has been used on the RNN model.

4.5 Developing Models

A brief description of the algorithms, neural networks, considered in this study are provided as follows:

4.5.1 Liblinear based Logistic regression

Logistic regression (LR) is a statistical method similar to linear regression since LR finds an equation that predicts an outcome for a binary variable, Y, from one or more response variables, X (DiGangi & Moore, 2012). It is mostly used for binary classification where it 24 extracts features with real values from the input and passes them as a sum to the sigmoid function after multiplying each by its weight to generate a probability. The multiclass LR uses the softmax function instead of sigmoid Jurafsky & Martin (2018). A Logistic Regression classifier with a Liblinear solver can minimize significant cost function regarding textual data classification and thus it has wide acceptance in the field of cyberbully identification.

4.5.2 Multinomial Naive Bayes

The Naive Bayes is simply an algorithm that classifies assuming the features are independent of its given class. It is a learning algorithm that is frequently used in text classification problems. The multinomial Naive Bayes classifier is suitable for classification with discrete features like word counts for text classification. The classifier with multinomial distribution can work with both integer and fractional feature counts and thus TFIDF and Bag of Words embedding methods can be useful with this algorithm (Kibriya et al., 2004b).

4.5.3 Random Forest

The Random forest algorithm is a popular example of the ensemble technique's bagging algorithm that uses averaging to ensemble a number of individual decision trees trained on a subset of the training dataset. It constructs many decision trees that will be used to classify a new instance by the majority vote where each decision tree node uses a subset of attributes randomly selected from the whole original set of attributes. Additionally, each tree uses a different bootstrap sample data in the same manner as bagging (Oshiro et al., 2012). Due to its high computational ability, a rigorous amount of work related to text classification has used the algorithm. In this study, the max-length of the tree has been pruned to 50 in developing both models.

4.5.4 Regularized Linear SVM

The SVM algorithm finds the optimal hyperplane in classification searching for similar instances of classes which are called support vectors. An advantage of SVMs over other learning algorithms is that it can achieve good performance when applied to real-world

problems like text classification. SVM classifier can be regularized by optimizing various regularization parameters such as C, gamma etc (Hearst et al., 1998). The classifier with a linear kernel has been said to perform better than others in multi-dimensional spaces like textual data.

4.5.5 BiLSTM RNN with GloVe Embedding

Bidirectional LSTM is a recurrent neural architecture consisting of forwarding and backward LSTM's. The forward LSTM reads the input text in forwarding order and uses the contextual information from the past. The backward LSTM reads the text description in the reverse order and preserves the contextual information from the future. These two LSTM's generate two independent sequence output vectors. We obtained an output vector for each word by concatenating these forward and backward vectors. Fig. 4.2 shows the architecture of BiLSTM RNN with GloVe embeddings. The BiLSTM model has only been built for bully classification purposes with four classes attacking, misogyny, neutral and racism on the same dataset.

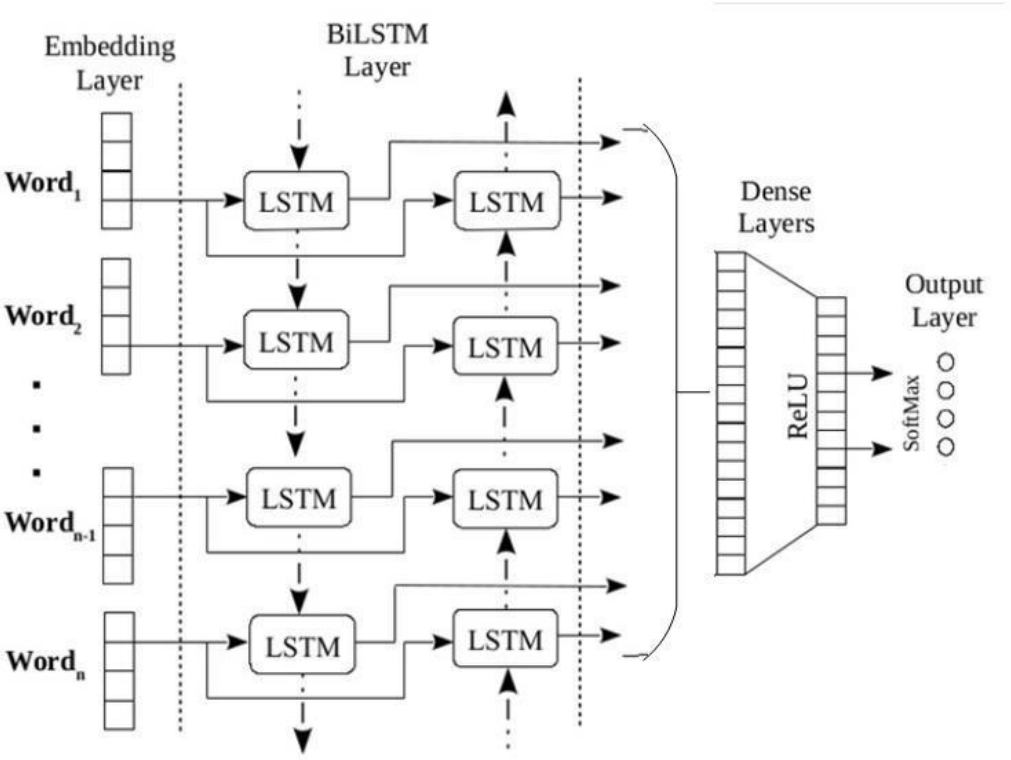


Fig. 4.2: BiLSTM RNN with Glove Embedding

CHAPTER 5
ANALYZING AND COMPARING ML ALGORITHMS

Performance measures generally evaluate specific aspects of classification task performance and thus do not always present identical information. In this chapter, we discuss evaluation metrics and performance measures and highlight ways in which the performance of the classifier models was compared.

5.1 Evaluating Performance through Precision, Recall and F1 Scores

Precision, Recall and F1 scores are good performance metrics for any classification task. These metrics are based on a "confusion matrix" that comprises true positives (TP), false positives (FP), false negatives (FN), and true negatives (TN). The significance of these four elements may vary based on the classification application. The fraction of known positives that are correctly predicted is called sensitivity i.e., true positive rate or recall (Equation 5.1). The proportion of correctly predicted positives to all positives is called precision, i.e., positive predictive value (Equation 5.2). The F1 score is derived from the precision and recall score (Equation 5.3).

$$precision = \frac{TP}{TP} + FP \tag{5.1}$$

$$recall = \frac{TP}{TP} + FN \tag{5.2}$$

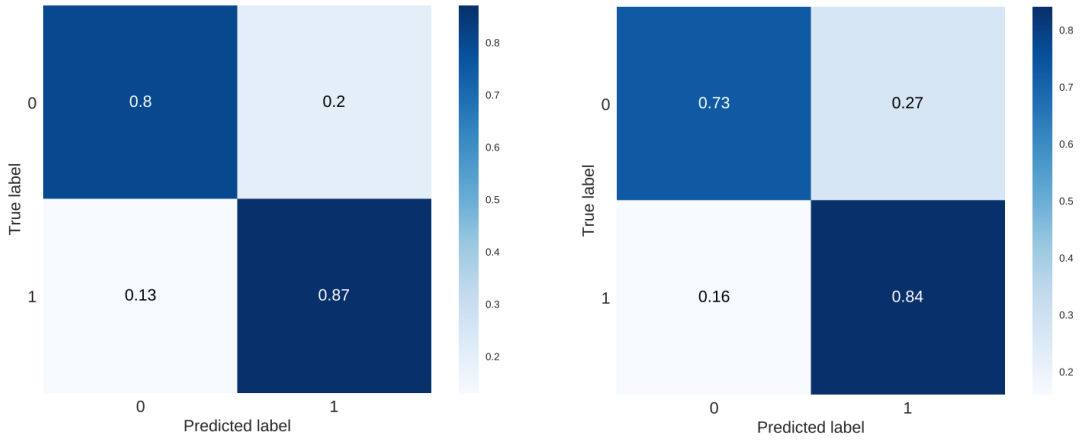
$$F1 - score = 2 * \left(precision * \frac{recall}{precison} + recall \right) \tag{5.3}$$

The prediction models are generated using the training dataset whereas the performance of prediction models is evaluated for the unknown dataset (test set). The test set is 30% of the total dataset. The performance of bully identification models for the training and testing instances was measured in terms of precision, recall, and f1 score, and the results are shown in table 5.1.

It is evident from table 5.1 that the models with TFIDF embedding work better than BoW embedding. In the BoW embedding technique Logistic Regression performed better whereas in TFIDF embedding Random Forest showed better accuracy. The confusion matrices of the Logistic Regression model are provided in Fig. 5.1 and Fig. 5.2 respectively for train and test data. In both Figures class "non-bully" or "positive-sentiment" is denoted as 0 and class "bully" or "negative-sentiment" is denoted as 1.

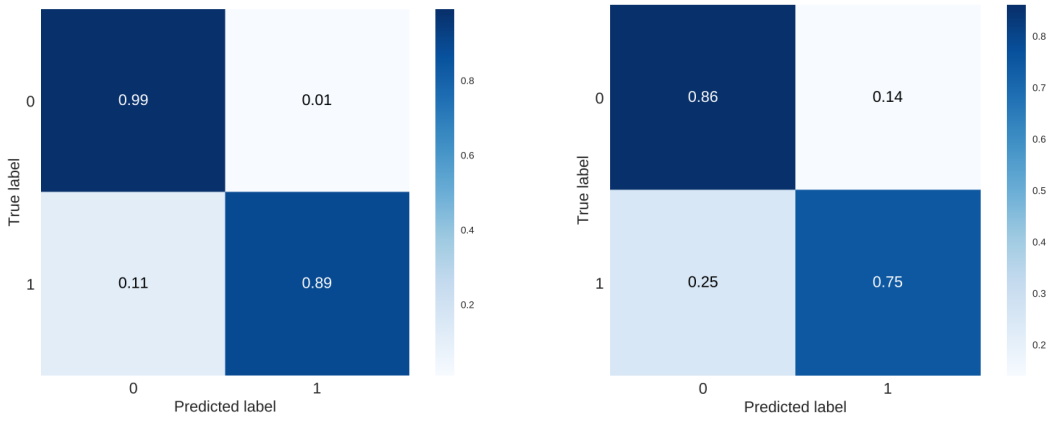
Table-5.1: Performance of Bully Identification Model

Models		Train Set			Test Set		
Embedding	**Algorithm**	**Precision**	**Recall**	**F1**	**Precision**	**Recall**	**F1**
BOW	Log Regression	0.832	0.830	0.830	0.782	0.779	0.778
	Naive Bayes	0.765	0.765	0.765	0.723	0.723	0.723
	SVM	0.837	0.834	0.834	0.774	0.769	0.768
	RF	0.893	0.883	0.882	0.750	0.747	0.746
TFIDF	Log Regression	0.892	0.892	0.892	0.794	0.794	0.794
	Naive Bayes	0.873	0.873	0.873	0.742	0.742	0.742
	SVM	0.910	0.909	0.909	0.781	0.781	0.781
	RF	0.943	0.937	0.937	0.813	0.809	0.808



(a) (b)

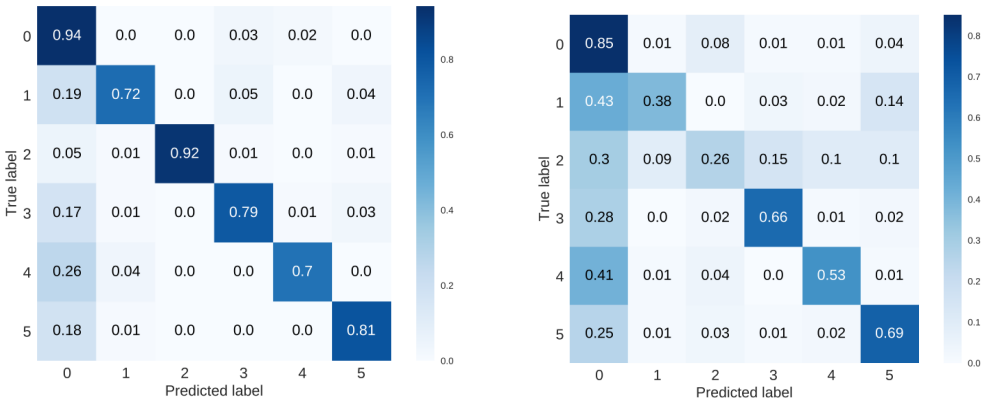
Fig.-5.1: Confusion matrices for Identification model-Logistic Regression with BOW embedding: (a) train data, (b) test data



(a) (b)

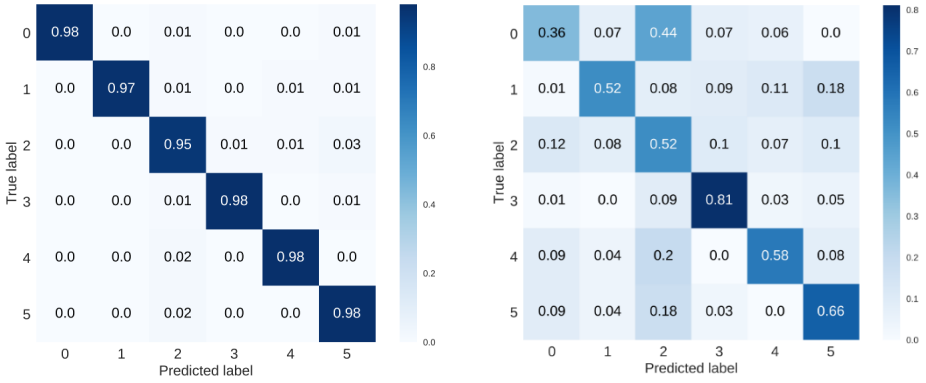
Fig.-5.2: Confusion matrices for Identification model- RF with TFIDF embedding: (a) Train data, (b) test data

The performance of Bully classification models built with machine learning algorithms for the training and testing instances was measured in terms of precision, recall, and f1 score, and the results are shown in table 5.2. It is evident from table 5.2 that the models with TFIDF embedding works better than BoW embedding and in both embedding methods Random Forest performs better than others. The confusion matrices of both models are provided in Fig. 5.3 and Fig. 5.4, where class 0=attacking, 1=misogyny, 2=neutral, 3=offensive, 4=positive and 5= racism as per proposed taxonomy.



(a) (b)

Fig.-5.3: Confusion matrices for Classification model- RF with BOW embedding: (a) train data, (b) test data



(a) (b)

Fig.-5.4: Confusion matrices for Classification model- RF with BOW embedding: (a) train data, (b) test data

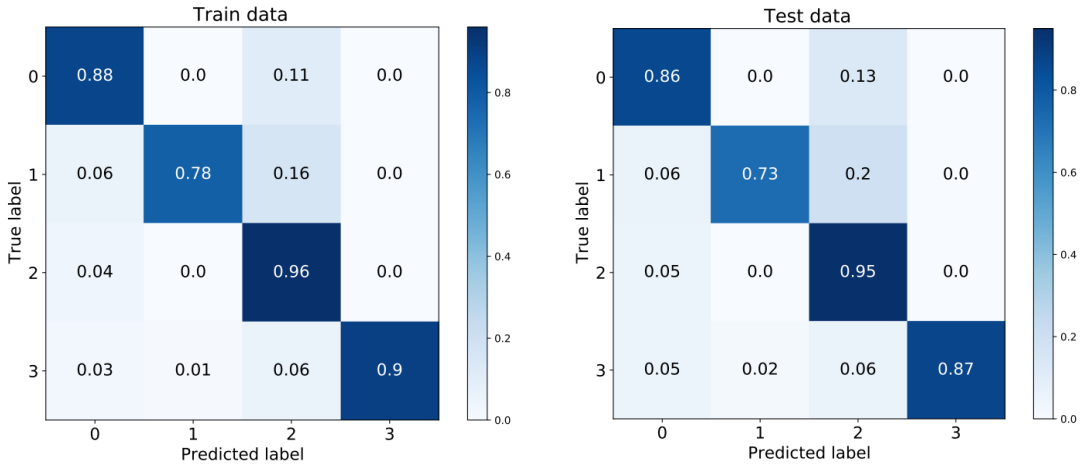
Table-5.2: Performance of Bully Classification Model

Model		**Train Set**			**Test Set**		
Embedding	**Algorithm**	**Precision**	**Recall**	**F1**	**Precision**	**Recall**	**F1**
BOW	Log Regression	0.823	0.759	0.771	0.617	0.515	0.515
	Naive Bayes	0.766	0.705	0.714	0.571	0.491	0.490
	SVM	0.860	0.792	0.805	0.648	0.506	0.511
	RF	0.872	0.806	0.818	0.637	0.527	0.524
TFIDF	Log Regression	0.834	0.838	0.835	0.626	0.559	0.576
	Naive Bayes	0.835	0.827	0.815	0.553	0.543	0.544
	SVM	0.888	0.885	0.886	0.596	0.489	0.505
	RF	0.980	0.979	0.979	0.635	0.575	0.584
GloVe	BiLSTM RNN	0.927	0.879	0.901	0.908	0.852	0.877

However, BiLSTM model with GloVe embedding outperformed machine learning models with precision, recall and F1 score of 90.8, 85.2 and 87.7 respectively. The confusion matrices of both models are provided in 5.5, where class 0=attacking, 1=misogyny, 2=neutral and 3= racism.

Table-5.3: AUC scores of Classification Models

Classes	Attacking	Misogyny	Neutral	Racism	Offensive	Positive
RF-BOW	82	80	76	89	88	87
RF-TFIDF	87	84	74	94	90	85
BiLSTM RNN-GloVe	98	99	97	100	-	-



(a) (b)

Fig.-5.5: Confusion matrices for Classification model- BiLSTM RNN with Glove embedding: (a) train data, (b) test data

5.2 Evaluating performance through ROC and AUC Score

Receiver operating characteristics (ROC) curves are valuable graphical tools for evaluating classifiers. ROC curve plots true positive rate (TPR) and false-positive rate (FPR).

The area under the ROC curve (AUC) measures the entire area underneath the ROC curve. AUC leverages helpful properties such as increased sensitivity in the analysis of variance (ANOVA) tests, independence from decision threshold, invariance to a priori class probabilities, and an indication of how well classes are in regarding the decision index.

For Identification models the ROC curve having the highest F1 score in BoW embedding is presented in Fig. 5.6, having an AUC score of 0.86 for both classes. The RF model

achieved the highest F1 score in TFIDF embedding having an AUC score of 0.88 for both classes. The ROC curve is shown in Fig. 5.7.

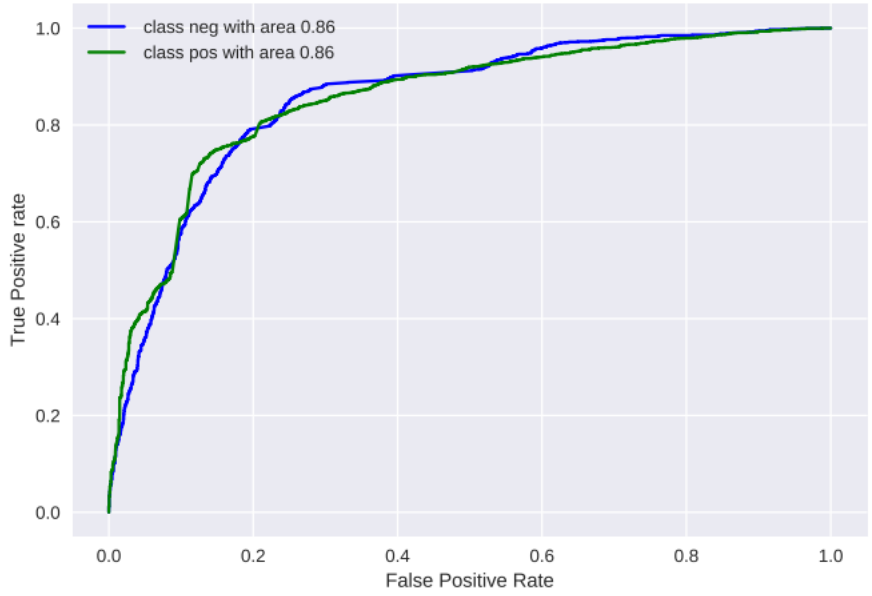


Fig.-5.6: ROC curve for Logistic Regression model with BOW embedding

The AUC scores of classification models achieving the highest F1 scores in each embedding technique is provided in table 5.3. With the BoW technique, the RF model achieved the highest AUC score with class racism at 89% and the lowest with the neutral class at 76%. The ROC curve is provided in Fig. 5.8.

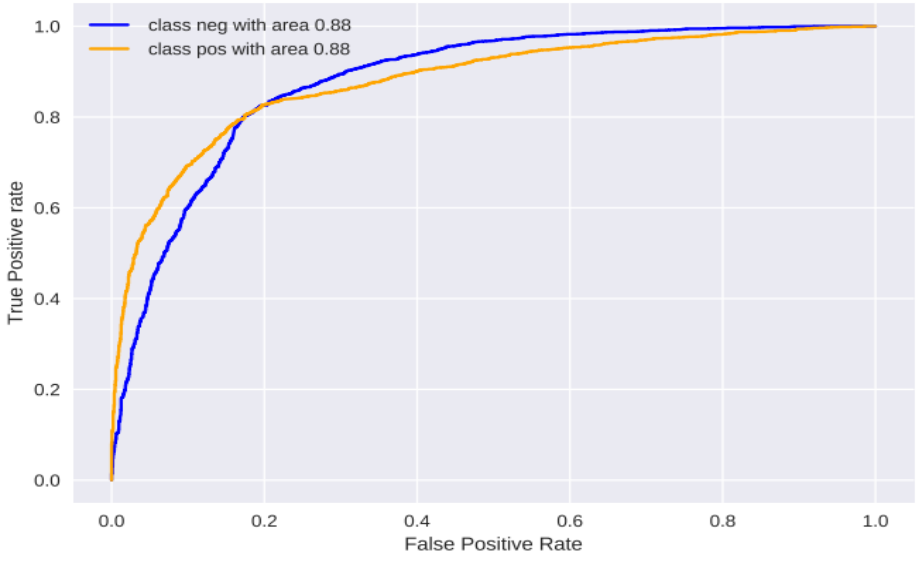


Fig.-5.7: ROC curve for Random Forest model with TFIDF embedding

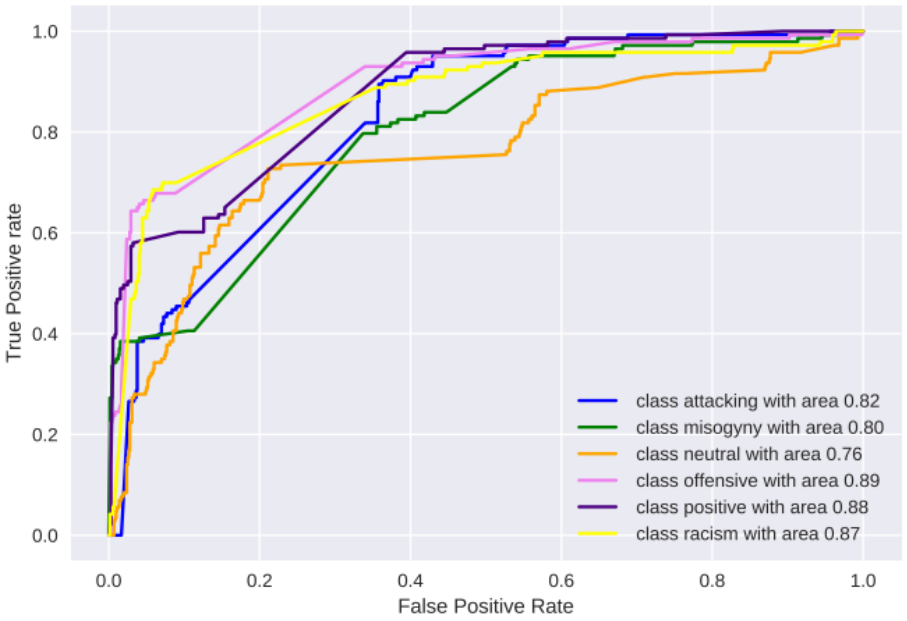


Fig.-5.8: ROC curve for Random Forest model with BOW embedding

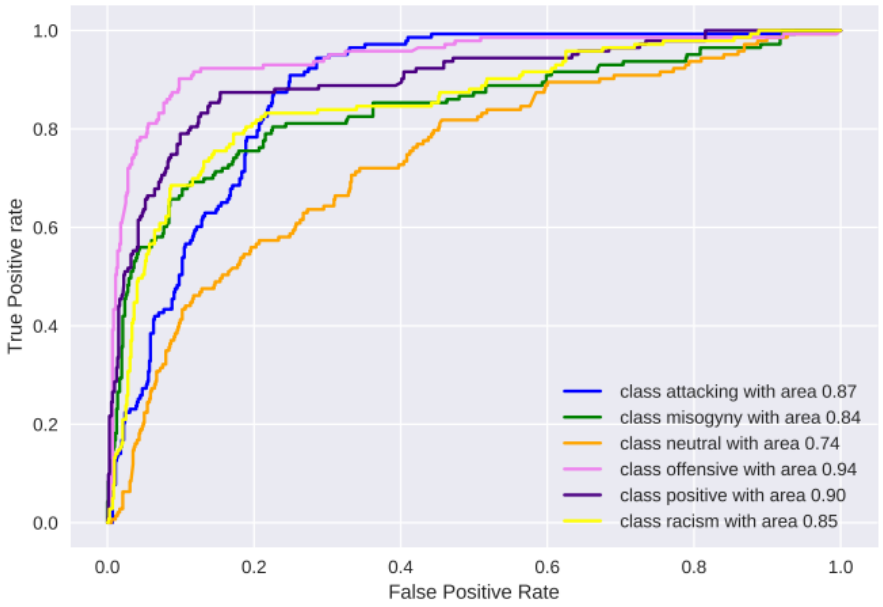


Fig.-5.9: ROC curve for Random Forest model with TFIDF embedding

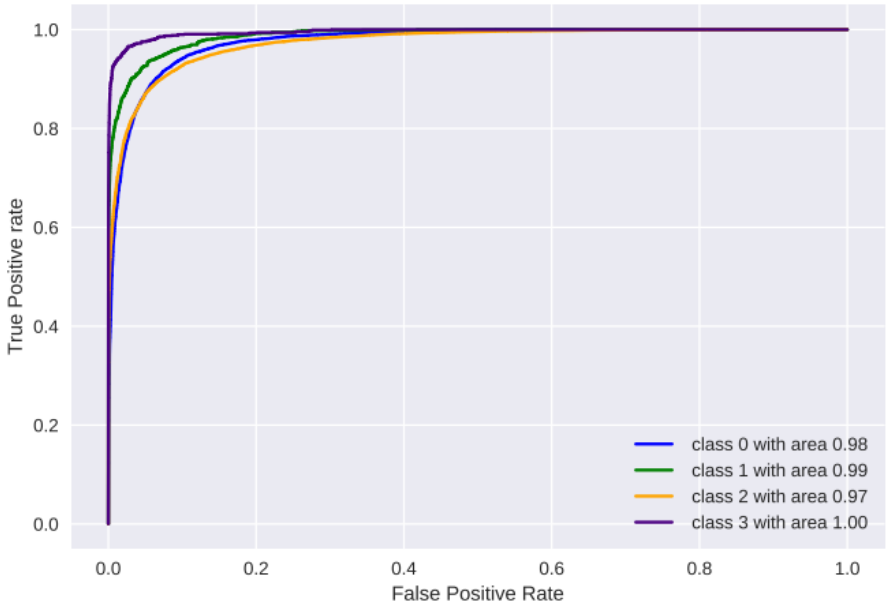


Fig.-5.10: ROC curve for BiLSTM RNN model

The RF with TFIDF showed improvement in performance with highest AUC score of 94% in a racism class. In comparison with the BoW technique, AUC scores increased with every class except the “neutral”. The ROC curve is provided in Fig. 5.9.

The BiLSTM model with GloVe outperformed previously mentioned models with AUC scores of 98%, 99%, 97% and 100% with “attacking”, “misogyny”, “neutral” and “racism” classes ‘respectively. The ROC curve is in Fig. 5.10.

CHAPTER 6
CONCEPTUAL FRAMEWORK

To apply cyberbullying detection, a conceptual framework is proposed and shown in Fig. 6.1. There are three modules in this system which are consisted of the user interaction module, the Decision-making module, and the analysis module. This proposal is essentially proposed to the authority of a social media website. This can be used to abate cyberbullying on social media.

6.1 User Interaction Module

Primarily, the user interaction module is the website. It can be a mobile application or web page. The content that will be taken from here is the posts and comments. This system can be used to monitor messages as well. Users will have an individual account from which a user can post or comment on others' posts. The contents are taken into the analytics module.

6.2 Analytics Module

This module contains the best performing machine learning model which is deployed in the cloud server. The model pertains so it will be able to detect bully and nonbully contents. When bully content is detected it can classify the type of bully which result will be stored in a database for further monitoring.

6.3 Decision-Making Module

This module takes action on the result given from the analytics module. This module contains database storage which keeps the history of a bullet account for further monitoring. If the content is nonbully, this module takes no action. But if the content is a bully, after taking the type, the module will decide to give a badge in the bully profile to minimize the cyberbullying from the social media website.

This proposal is essentially proposed to the authority of a social media website to abate cyberbullying on social media. After content is identified as a bully, the type of bully will be determined by the model. The model will give probabilities of the kinds of bullying mentioned in the taxonomy development section and the kind with the highest probability will be selected.

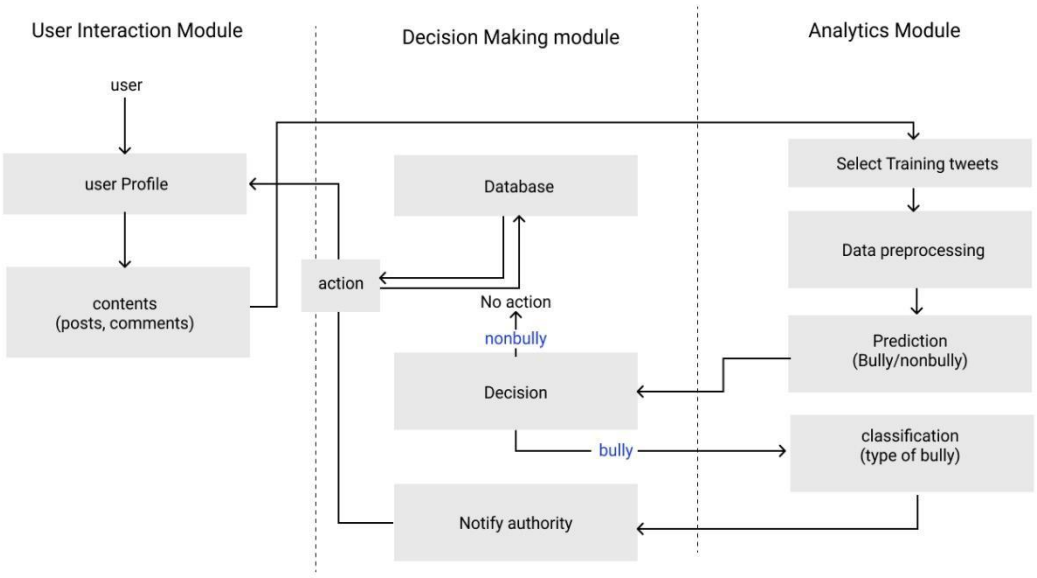


Fig.-6.1: Framework

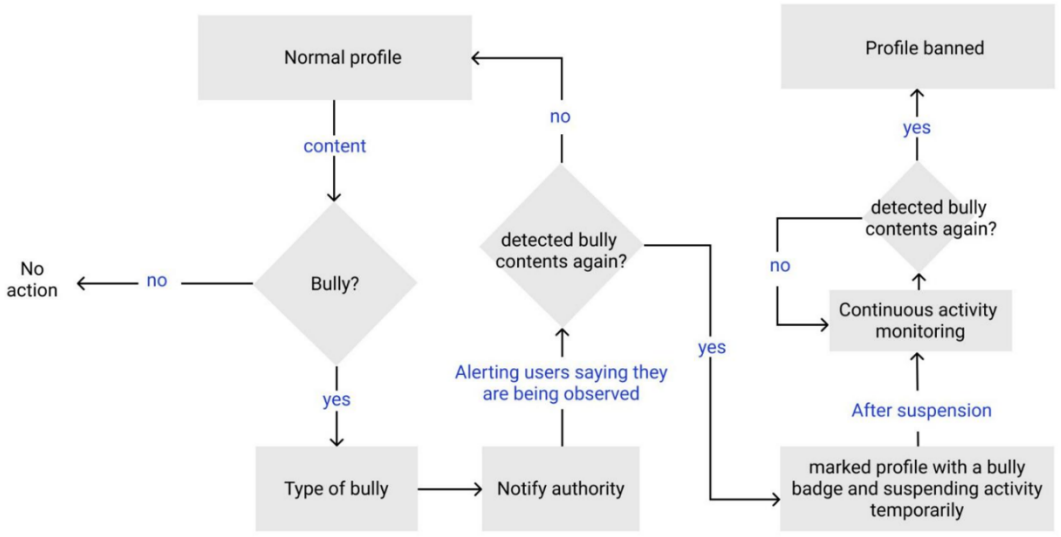


Fig.-6.2: Protocol

Classification of the bully can be used according to the preference of a website, as different websites are built on different topics where the severity of the kinds of bully differs. When content is marked as a bully, the history will be collected against the user profile activity history (Database). If user posts or comments on a particular number of bullies within 24 hours, that user will be given an alert from the website letting them know that they are being monitored and authorities will be notified with the victim

profiles in a database. if the user bullies again in the next 168 hours user will be unable to post or comment for 168 hours and the user's profile will contain a badge showing that the user is
a bully. The time allotment can be determined by authorities. If the user does not comment within a particular time, the suspension will be removed but the user will be marked as a bully in the user's history (Database). The bully badge will play a psychological role so that the user will not be willing to post bully content because that way they will gain the bully badge which is not a positive thing. After the action is over, if the user still comments on a bully, the profile will be banned permanently. The protocol is shown in Fig. 6.2.

CHAPTER 7
CYBERBULLY ANALYSIS

For exploring, how people's sentiment and responses to events change, public tweets concerning global politics and social-economic issues worldwide and in Bangladesh have been collected from the timeline after the pandemic of COVID-19 (2020-2021) and classified into four categories (attacking, misogyny, neutral and racism) using our recurrent neural network-based model. The objective of this analysis of public sentiment is to observe how cyberbullying occurs on Twitter revolving around global concerns and trending topics worldwide and the frequently stated words used in cyberbullying on social media.

A histogram of distinct sentiments vs the number of frequencies in the predicted tweets is presented in Fig. 7.1. The results revealed that from all collected tweets 61.5% of tweets were classified as neutral tweets, 37% of the tweets are classified as attacking tweets, whereas 1% and 0.5% of tweets were classified as misogynistic and racial tweets, respectively. We have used Python pandas and matplotlib packages for data analysis and visualization.

7.1 Frequently Stated Words in Cyberbully

The most frequent 10 words from all the tweets classified as a bully (attacking, racism and misogyny classes altogether) and their number of occurrences in the dataset are plotted in Fig. 7.2.

This analysis shows a significant side of the research on cyberbullying from social media. The most used words used in the case of cyberbullying are common offensive words like: 'fuck', 'dick', 'shit', 'shame' and chauvinistic or racist words like: 'nigga', 'terrorist'. The other non-negative words which have been observed in the list of ten are: 'stand', 'war', 'women' which conveys the message that the most hate comments on social media are around the topics relating to women or warfare.

The word clouds are generated all for four classes: attacking, misogyny, neutral, and racism are presented in Fig. 7.3. The size of each word indicates its frequency or importance in the word cloud. The words observed in the word clouds for each class indicates the word used for each type of bully characteristics. The word clouds also highlight the specific frequently used words in each category of the cyberbully detection process. A critical point to observe is that for expressing attacking sentiment people use offensive language mostly whereas language relating to the female gender along with offensive words are used to express misogynistic behavior. Words like 'zionist', 'terrorist', 'jews' etc. expressing racial segregation or chauvinistic nature are used in the tweets classified as racism.

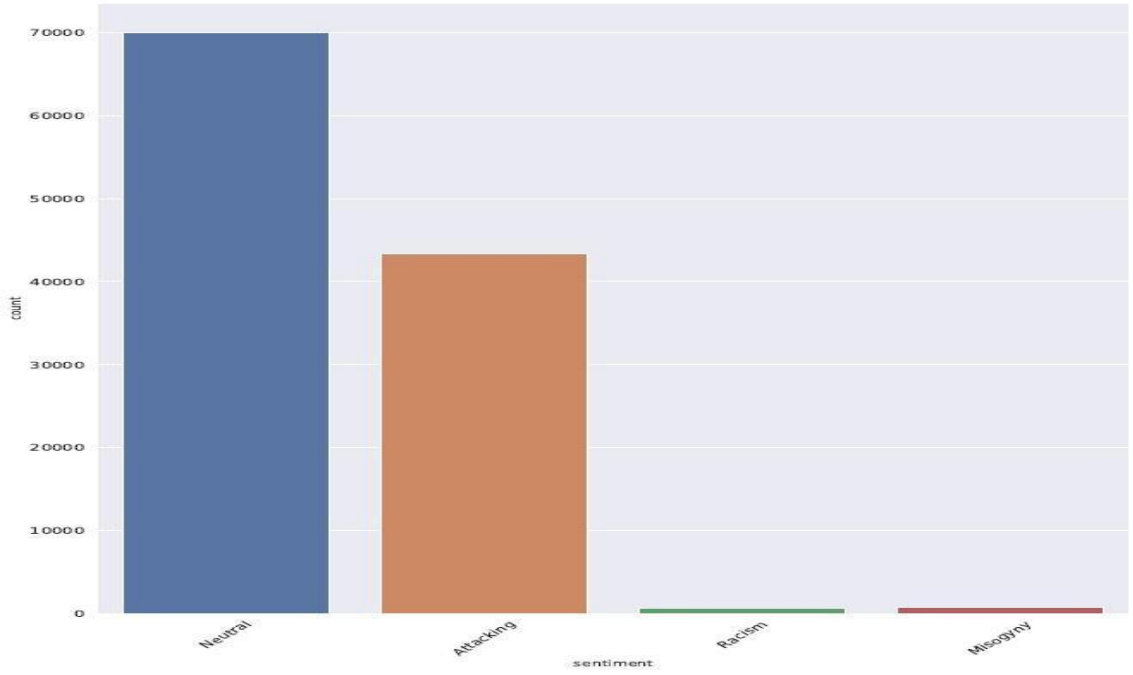


Fig.-7.1: Histogram of distinct sentiments vs the number of frequencies in the predicted tweets

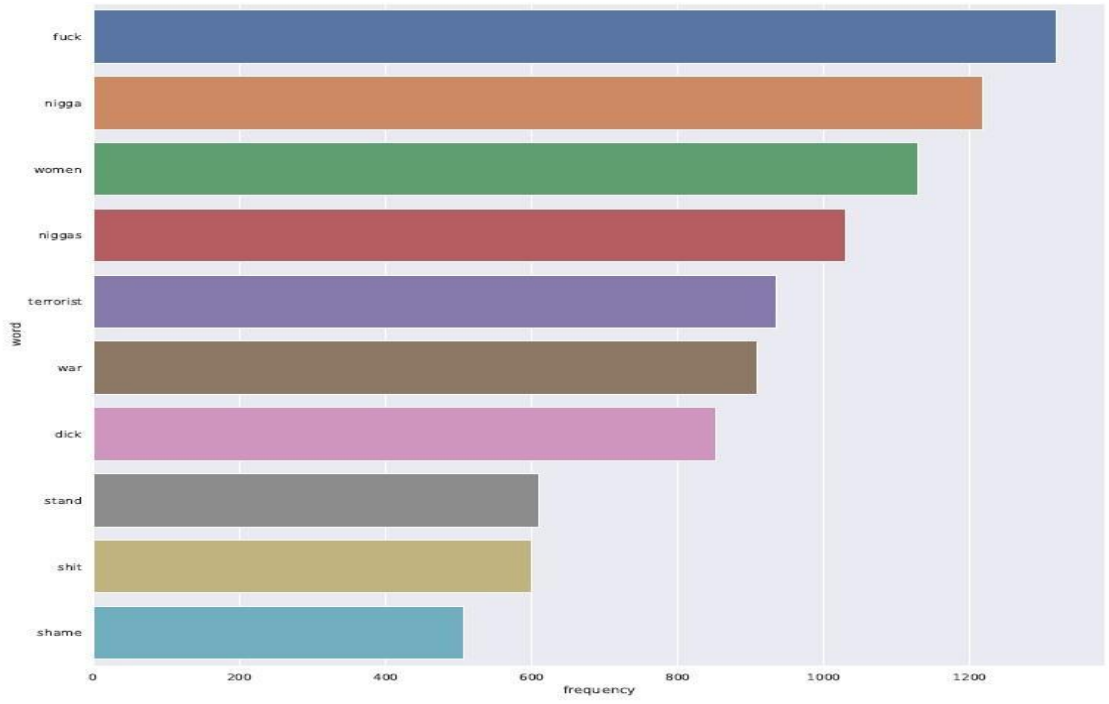
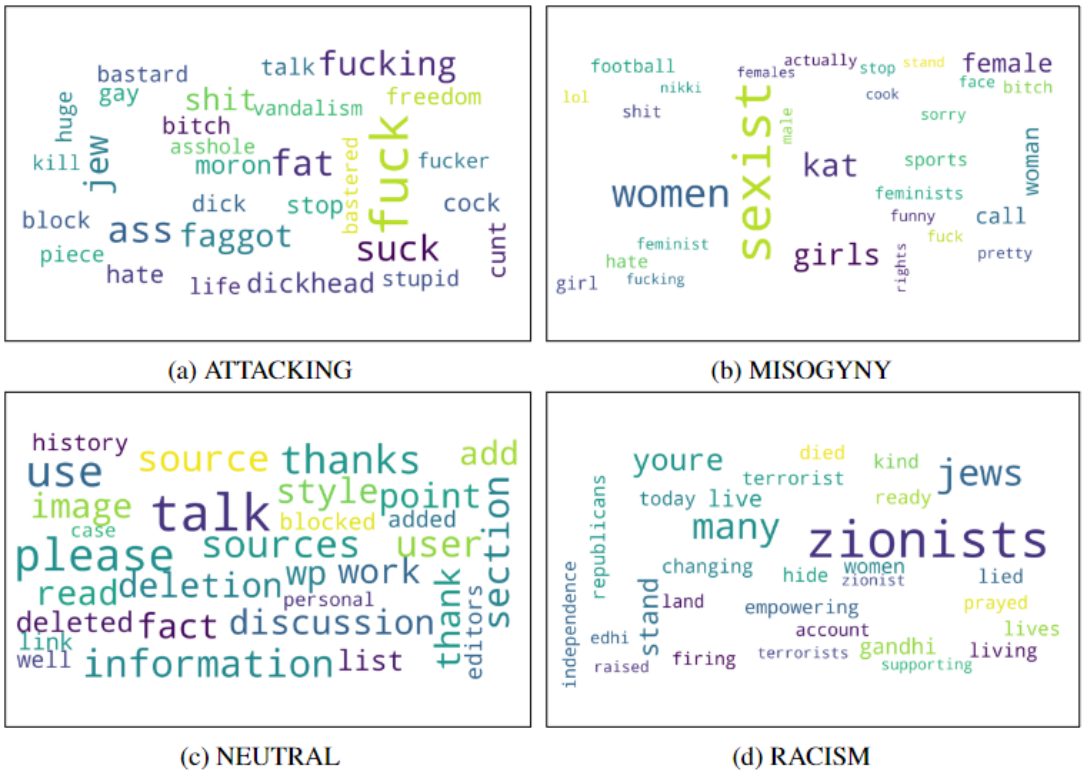


Fig. 7.2: Most frequent 10 words from all the tweets classified as bully



(a) ATTACKING (b) MISOGYNY

(c) NEUTRAL (d) RACISM

Fig.-7.3: Word Clouds for Predicted Classes

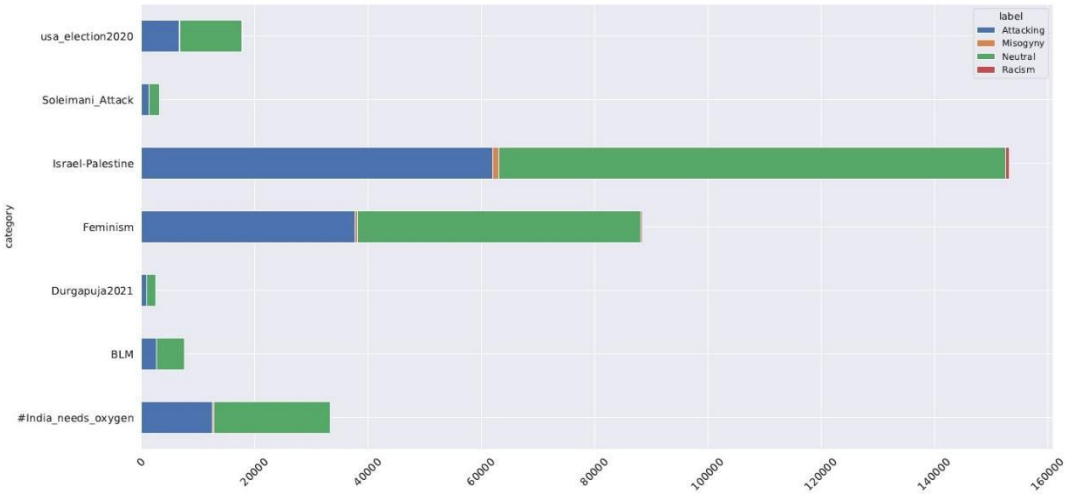


Fig. 7.4: Classes of cyberbully against categories of collected tweets

7.2 Cyberbullying against Global Concerns

The data collection for this analysis part have been done to extract tweets regarding worldwide trending topics on Twitter and global concerns or events from timeline 2020-2021 to understand the public sentiment upon these issues. For this, each tweet is grouped in some specific categories based on the presence of the words regarding the events or concerns. The tweets and the respective categories are plotted in Fig. 7.4 where each of the bars denotes the total number of attacking, misogyny, neutral and racism tweets indistinguishable colors. We can see all of the categories defined in the bar chart have a significant amount of attacking tweets, which reveals people's aggressive attitude toward attacking the government or spreading hatred to the communities regarding the issues. The tweets relating to the Israel-Palestine conflict, USA election 2020, feminism, Durga puja 2021 (communal riots relating to this event) have more negative sentiments than others. The tweets relating to feminism and the Israel-Palestine conflict have misogynistic tweets whereas the latter one is the only category having racist tweets. Fig. 7.5 provides the frequency of the cyberbully-related tweets upon the USA election 2020 with the year 2020. The information is displayed in Fig 7.5. can be interpreted in the way that people's negative sentiment relating to an event increases as the event gets closer to its scheduled date.

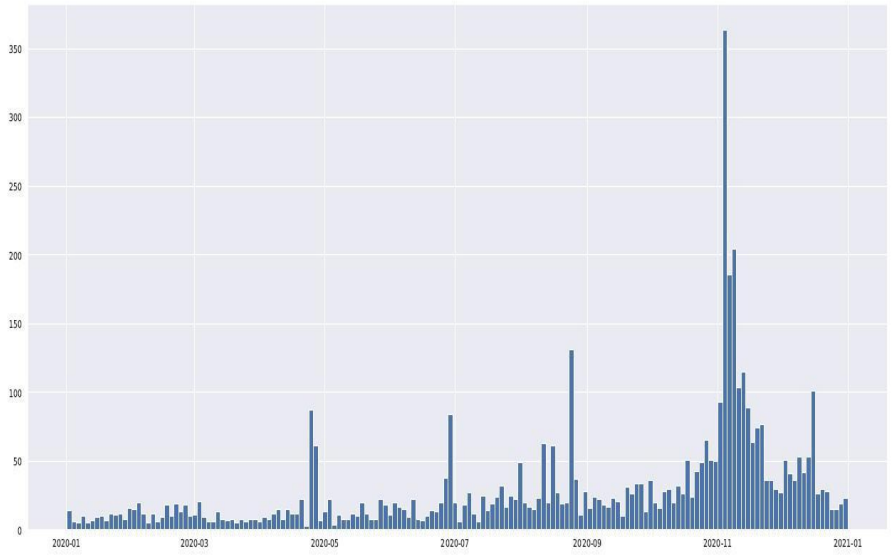


Fig. 7.5: Frequency of the cyberbully related tweets upon the USA election 2020 with the year 2020

CHAPTER 8
DISCUSSION AND CONCLUSION

This chapter describes the overall thesis by using four subsections to summarize the thesis's outcome, implications and contribution of the research, limitations, and scope for future work.

8.1 Thesis Outcome

Over the last decade, with new technological advancements, cyberbullying has become one of the most significant issues in the world. This study contributed to introducing a framework for preventing cyberbullying on SM by detecting bullying textual posts by employing sentiment analyses and bullying features through the exploration of ML algorithms. The dataset corpora extracted from Twitter are diverse in content. The taxonomy designed for bully characteristics covers most of the areas a person is bullied for such as racism, prejudice against women, disrespectful or insulting words, aggressiveness, etc. Embedding methods such as BoW and TFIDF have been used on thoroughly preprocessed tweets for applying selected classifiers namely: Liblinear based Logistic Regression, Linear SVM, Multinomial Naive Bayes, and Random Forest. It is evident from several performance evaluation measures of the models that the Random Forest model with TFIDF embedding performs better in both cases. The highest achieved accuracy (F1 score) for the bully identification model is 80.8% and for the bully classification model is 87.7%.

8.2 Thesis Implication and Contribution

The framework proposed for the prevention of cyberbullying is a prevention measure generalized for all SM platforms. In the proposed framework, the analytics module shall automatically analyze user posts/comments through deployed ML models in the cloud and the decision module shall take action accordingly if a bullying attempt is found. This framework is targeted to contribute to SM authorities so that monitoring of bullies can be automatic and certain measures can be taken against the guilty to stop cyberbullying trend for good.

8.3 Thesis Limitation

However, this paper used only two NLP techniques named BoW and TFIDF to transform text data into machine-readable vectors for feature extraction. Again, for both identification and classification of bullies, a limited amount of 1000 labeled tweets has been used exploring only the ML algorithms. The use of modern word embedding techniques along with neural networks could have improved the performance of the models. Furthermore, dataset corpora has been used is collected from Twitter only including the publicly available tweets as it provides a public API for data extraction.

Other SM platforms such as Facebook could have been investigated as it is the most used communication media worldwide with a record of severe cyberbullying.

8.4 Future Work

In the future, the constraints highlighted in this work can be addressed. Additionally, website construction can be done in order to incorporate the cyberbully prevention strategy discussed in this paper. Trolls and memes are tough to spot because they include photographs and videos. Contribution to cyberbully detection from trolls and memes may be addressed as well using image processing, ensuring that individuals have a harmless and safer SM experience.

REFERENCES

Al Marouf, A., Ajwad, R. & Ashrafi, A.F. (2019), Looking behind the mask: A framework for detecting character assassination via troll comments on social media using psycholinguistic tools, in '2019 International Conference on Electrical, Computer and Communication Engineering (ECCE)', IEEE, pp. 1–5

.

Anderson, M. (2018), 'A majority of teens have experienced some for of cyberbullying'.Anzovino, M., Fersini, E. & Rosso, P. (2018), Automatic identification and classification of misogynistic language on twitter, in 'International Conference on Applications of Natural Language to Information Systems', Springer, pp. 57–64.

Breiman, L. (2001), 'Random forests', Machine Learning 45(1), 5–32.

Cheng, L., Li, J., Silva, Y., Hall, D. & Liu, H. (2019a), Pi-bully: Personalized cyberbullying detection with peer influence, in 'The 28th International Joint Conference on Artificial Intelligence (IJCAI)'.

Cheng, L., Li, J., Silva, Y. N., Hall, D. L. & Liu, H. (2019b), Xbully: Cyberbullying detection within a multi-modal context, in 'Proceedings of the Twelfth ACM International Conference on Web Search and Data Mining', pp. 339–347.

Chowdhury, G. G. (2003), 'Natural language processing', Annual review of information science and technology 37(1), 51–89.

Christmann, A. & Steinwart, I. (2008), 'Support vector machines'.

Cook, S. (2020), Cyberbullying facts and statistics for 2020', Web-site Comparitech. com.Updated: March (6), 2020.

DiGangi, E. A. & Moore, M. K. (2012), Research methods in human skeletal biology, Academic Press

DIGITAL 2021 : GLOBAL OVERVIEW REPORT (n.d.), availableon: https://datareportal.com/reports/digital-2021-global-overview-report. Last accessed: 12 July, 2021.

Djuraskovic, O. (2021), 'Cyberbullying statistics, facts, and trends (2021) with charts'.

EFFECTS OF CYBERBULLYING (n.d.), availableon:https://americanspcc.org/impact-of-cyberbullying/. Last accessed: 22 July, 2021.

El Naqa, I. & Murphy, M. J. (2015), What is machine learning?, in 'machine learning in radiation oncology', Springer, pp. 3–11.

Fan, R.-E., Chang, K.-W., Hsieh, C.-J., Wang, X.-R. & Lin, C.-J. (2008), 'Liblinear: A library for large linear classification', the Journal of machine Learning research 9, 1871–1874.

Ge, S., Cheng, L. & Liu, H. (2021), Improving cyberbullying detection with user interaction, in 'Proceedings of the Web Conference 2021', pp. 496–506.

Hearst, M. A., Dumais, S. T., Osuna, E., Platt, J. & Scholkopf, B. (1998), 'Support vector machines', IEEE Intelligent Systems and their applications 13(4), 18–28.

Hutto, C. & Gilbert, E. (2014), Vader: A parsimonious rule-based model for sentiment analysis of social media text, in 'Proceedings of the International AAAI Conference on Web and Social Media', Vol. 8.

John, A., Glendenning, A. C., Marchant, A., Montgomery, P., Stewart, A., Wood, S., Lloyd, K. & Hawton, K. (2018), 'Self-harm, suicidal behaviours, and cyberbullying in children and young people: systematic review', Journal of medical internet research 20(4), e 9044.

Jurafsky, D. & Martin, J. H. (2018), 'Speech and language processing (draft)', preparation [cited 2020 June 1] Available from: https://web. stanford. edu/~ jurafsky/slp3 .

Kibriya, A. M., Frank, E., Pfahringer, B. & Holmes, G. (2004a), Multinomial naive bayes for text categorization revisited, in 'Australasian Joint Conference on Artificial Intelligence', Springer, pp. 488–499.

Kibriya, A. M., Frank, E., Pfahringer, B. & Holmes, G. (2004b), Multinomial naive bayes for text categorization revisited, in 'Australasian Joint Conference on Artificial Intellience', Springer, pp. 488–499.Lin, Y. (10), 'Twitter statistics every marketer should know in 2020 [infographic].'", Re-trieved 15 September 2020, from https://my. oberlo. com/blog/twitter-statistics

Luceri, L., Giordano, S. & Ferrara, E. (2020), Detecting troll behavior via inverse reinforcement learning: A case study of russian trolls in the 2016 us election, in 'Proceedings of the International AAAI Conference on Web and Social Media', Vol. 14, pp. 417–427.

Mihaylov, T. & Nakov, P. (2019), 'Hunting for troll comments in news community forums', arXiv preprint arXiv:1911.08113 .

Niklas, D. (2019), 'A complete guide to the random forest algorithm', Built In .URL: https://builtin.com/data-science/random-forest-algorithm

Oshiro, T. M., Perez, P. S. & Baranauskas, J. A. (2012), How many trees in a random forest?, in 'International workshop on machine learning and data mining in pattern recognition', Springer, pp. 154–168.

Parime, S. & Suri, V. (2014), Cyberbullying detection and prevention: Data mining and psychological perspective, in '2014 International Conference on Circuits, Power and

Computing Technologies [ICCPCT-2014]', IEEE, pp. 1541–1547.Rish, I. et al. (2001), An empirical study of the naive bayes classifier, in 'IJCAI 2001 workshop on empirical methods in artificial intelligence', Vol. 3, pp. 41–46.

Saravanaraj, A., Sheeba, J. & Devaneyan, S. P. (2016), 'Automatic detection of cyberbullying from twitter', International Journal of Computer Science and Information Technology & Security (IJCSITS) .

Sheeba, J. & Pradeep Devaneyan, S. (2016), 'Cyberbully detection using intelligent techniques'.

Smith, P. K., Mahdavi, J., Carvalho, M., Fisher, S., Russell, S. & Tippett, N. (2008), 'Cyberbullying: Its nature and impact in secondary school pupils', Journal of child psychology and psychiatry 49(4), 376–385.

Sugandhi, R., Pande, A., Agrawal, A. & Bhagat, H. (2016), 'Automatic monitoring andprevention of cyberbullying', International Journal of Computer Applications 8, 17–19.

Tokunaga, T. & Makoto, I. (1994), Text categorization based on weighted inverse document frequency, in 'Special Interest Groups and Information Process Society of Japan (SIG-IPSJ', pp. 33–39.

Tony, Y. (2019), Understanding random forest, Towards Data Science. URL: https://towardsdatascience.com/understanding-random-forest-58381e0602d2

Van Hee, C., Jacobs, G., Emmery, C., Desmet, B., Lefever, E., Verhoeven, B., De Pauw, G., Daelemans, W. & Hoste, V. (2018), 'Automatic detection of cyberbullying in social media text', PloS one 13(10), e0203794.

Vandebosch, H. & Van Cleemput, K. (2008), 'Defining cyberbullying: A qualitative research into the perceptions of youngsters', CyberPsychology & Behavior 11(4), 499–503.

Weller, H. & Woo, J. (2019), 'Identifying russian trolls on reddit with deep learning and bert word embeddings'.

Wright, R. E. (1995), 'Logistic regression.'.

Yao, M., Chelmis, C. & Zois, D. S. (2019), Cyberbullying ends here: Towards robust detection of cyberbullying in social media, in 'The World Wide Web Conference', pp. 3427–3433.

APPENDIX A

Table 9.1: Works Related to cyberbully detection

Scope	Summary
Detecting and Classifying Cyberbullying	Early detection of cyberbullying using linear kernel SVM
	Capturing troll behavior and identifying troll accounts using Inverse Reinforcement Learning
	Detecting cyberbullying rumors using Naïve Bayes and Random Forest algorithms
	Distinguishing trolls and non-trolls by training an L2-regularized Logistic Regression with Liblinear
	Detecting trolls by a three-layer neural network architecture
	Proposing a novel cyberbullying detection framework in multi-modal contexts

APPENDIX B

Table 9.2: Works Related to cyberbully prevention

Scope	Summary
Prevention or elimination of cyberbullying	Proposing a framework to monitor and control the flow of messages in social media
	Detecting trolls and a response grading system which takes action on basis of severity of bully
	Detecting and preventing of cyberbully using text mining and machine learning techniques
	Proposing a novel approach that satisfies key properties(accuracy, repetitiveness, timeliness and efficiency) for cyberbully detection