DETECTING HATE SPEECH AND CYBER HARASSMENT FROM MULTIPLE SOCIAL MEDIA PLATFORMS USING MACHINE LEARNING

A Dissertation

Submitted

In Partial Fulfillment of the Requirements for

The Degree of

MASTER OF TECHNOLOGY

In

Advanced Computing and Data Science

Submitted by:

Mohiyaddeen

Enroll. No. 1900104070

Roll No. 1901209003

Under the Supervision of:

Dr. Sifatullah Siddiqi

(Assistant Professor)

Department of Computer Science & Engineering

Faculty of Engineering

INTEGRAL UNIVERSITY, LUCKNOW, INDIA

August 2021

CERTIFICATE

This is to certify that **Mr. Mohiyaddeen** (Enroll. No.1900104070) has carried out the research work presented in the dissertation titled **"Detecting Hate Speech and Cyber Harassment from Multiple Social Media Platform using Machine Learning"** submitted for partial fulfillment for the award of the **Master of Technology in Advanced Computing and Data Science** from **Integral University, Lucknow** under my supervision.

It is also certified that:

(i) This thesis embodies the original work of the candidate and has not been earlier submitted elsewhere for the award of any degree/diploma/certificate.

(ii) The candidate has worked under my supervision for the prescribed period.

(iii) The thesis fulfills the requirements of the norms and standards prescribed by the University Grants Commission and Integral University, Lucknow, India.

(iv) No published work (figure, data, table etc) has been reproduced in the thesis without express permission of the copyright owner(s).

Therefore, I deem this work fit and recommend for submission for the award of the aforesaid degree.

_____

Dr. Sifatullah Siddiqi
Dissertation Guide
(Assistant Professor)
Department of CSE
Integral University, Lucknow

Date: 09/ 08/ 2021

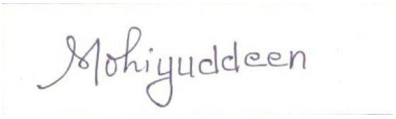
Place: Lucknow

DECLARATION

I hereby declare that the thesis titled **"Detecting Hate Speech and Cyber Harassment from Multiple Social Media Platform using Machine learning"** submitted to Computer Science and Engineering Department, Integral University, Lucknow in partial fulfillment of the requirements for the award of the Master of Technology degree, is an authentic record of the research work carried out by me under the supervision of Dr. Sifatullah Siddiqi, Department of Computer Science & Engineering, Integral University, Lucknow. No part of this thesis has been presented elsewhere for any other degree or diploma earlier.

I declare that I have faithfully acknowledged and referred to the works of other researchers wherever their published works have been cited in the thesis. I further certify that I have not willfully taken other's work, para, text, data, results, tables, figures etc. reported in the journals, books, magazines, reports, dissertations, theses, etc., or available at websites without their permission, and have not included those in this M.Tech. thesis citing as my own work.

Date: 09/ 08/ 2021



Signature

Name. Mohiyaddeen

Enroll. No. 1900104070

Roll. No. 1901209003

COPYRIGHT TRANSFER CERTIFICATE

Title of the Dissertation: **Detecting Hate Speech and Cyber Harassment from Multiple Social Media Platform using Machine learning**

Candidate Name: **Mohiyaddeen**

The undersigned hereby assigns to Integral University all rights under copyright that may exist in and for the above dissertation, authored by the undersigned and submitted to the University for the Award of the M.Tech degree.

The Candidate may reproduce or authorize others to reproduce material extracted verbatim from the dissertation or derivative of the dissertation for personal and/or publication purpose(s) provided that the source and the University's copyright notices are indicated.

MOHIYADDEEN

ACKNOWLEDGEMENT

I am highly grateful to the Head of Department of Computer Science and Engineering for giving me proper guidance and advice and facility for the successful completion of my dissertation.

It gives me a great pleasure to express my deep sense of gratitude and indebtedness to my guide **Dr. Sifatullah Siddiqi, Assistant Professor, Department of Computer Science and Engineering,** for his valuable support and encouraging mentality throughout the project. I am highly obliged to him for providing me this opportunity to carry out the ideas and work during my project period and helping me to gain the successful completion of my Project.

I am also highly obliged to the Head of department, **Dr. Mohammadi Akheela Khanum (Associate Professor, Department Of Computer Science and Engineering)** and PG Program Coordinator **Dr. Faiyaz Ahmad, Assistant Professor, Department of Computer Science and Engineering,** for providing me all the facilities in all activities and for his support and valuable encouragement throughout my project.

My special thanks are going to all of the faculties for encouraging me constantly to work hard in this project. I pay my respect and love to my parents and all other family members and friends for their help and encouragement throughout this course of project work.

Date: 09 / 08 / 2021

Place: Lucknow

RECOMMENDATION

On the basis of the declaration submitted by **"Mohiyaddeen"**, a student of M.Tech CSE (ACDS), successful completion of Pre presentation on 09/08/2021 and the certificate issued by the supervisor **Dr. Sifatullah Siddiqi**, Assistant Professor Computer Science and Engineering Department, Integral University, the work entitled "**Detecting Hate Speech and Cyber Harassment from Multiple Social Media Platform using Machine learning**" , submitted to department of CSE, in partial fulfillment of the requirement for award of the **Master of Technology in Advanced Computing and Data Science**, is recommended for examination.

Program Coordinator	Signature HOD Signature
Dr. Faiyaz Ahmad	Dr. M Akheela Khanum
Dept. of Computer Science & Engineering	Dept. of Computer Science & Engineering
Date: 09/08/21	Date: 09/08/21

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

LIST OF EQUATIONS

LIST OF ABBREVIATIONS AND SYMBOLS

Table no.	Description
ML	Machine Learning
AI	Artificial Intelligence
NLP	Natural Language Processing
SVM	Support Vector Machine
SSE	Sum of Squared Errors
DBSCAN	Density-Based Spatial Clustering
DBCLASD	Distribution-Based Clustering of Large Spatial Databases
HT	Hate Speech
FT	Free Speech
VADER	Valence Aware Dictionary for Sentiment Reasoning
TFIDF	Term Frequency -Inverse Document Frequency
GHSOM	Growing Hierarchical Self-Organizing Map
CNN	Convolutional Neural Network
LSTM	Long Short-Term Memory Network
BiGRU	Bi-directional Gated Recurrent Unit
LIWC	Linguistic Inquiry and Word Count
SWN	Sent WordNet
WSD	Word-Sense Disambiguation
ANEW	Affective Norms for English Words
GI	General Inquirer
BoW	Bag-of-Words
IDF	Inverse Document Frequency
DF	Document Frequency
LSA	Latent Semantic Analysis
BERT	Bidirectional Encoder Representations from Transformers
LDA	Linear Discriminant Analysis

LM	Local Meta Classifier
GM	Global Meta Classifier
MLP	Multilayer Perceptron

ABSTRACT

Online social media platforms provide freedom to their users to post or comment anything they want. The lack of regulation empowers social media users to post any hateful comments without any control that leads to riot, terrorism, and cyber harassment.

There has been a lot of research done on hate speech detection using various machine learning approach such as using logistic regression, random forest, or support vector machine. All these previous research focuses on finding better accuracy based on single machine learning algorithm.

In this dissertation, work has been done to improve the accuracy of already existing researches on hate speech detection using hybrid machine learning approach. It aimed at developing a more accurate hybrid machine learning algorithm that contributed to overcome some of the major disadvantages of the prior system. The objective was not only to focus on the accuracy but improve the precision, recall and f-score. Another advantage of this research is that it also gives better accuracy with minimum number of datasets.

The combination of six different machine learning algorithm including support vector machine, k-nearest neighbor, decision tree, random forest, naïve bayes and logistic regression were used to create a hybrid machine learning model. While the TFIDF approach was used to extract the important features from the textual data. We achieved 94.78% accuracy with 14000 rows of dataset. A comparison graph plot was illustrated to demonstrate the performance of the hybrid machine learning model.

CHAPTER - 1

INTRODUCTION

1.1 SOCIAL MEDIA

In recent years, social networking has become a popular way for people to connect with one another on a regular basis.. Facebook, Instagram, and Twitter are examples of social media platforms that allow you to interact with family and friends, as well as individuals who have similar interests to your own.

If anyhow you're involved in social networking, it implies you're making use of social media sites, also known as social networks, to interact with other people and share information. Twitter, Facebook, Reddit, YouTube, Telegram, TikTok, Instagram, LinkedIn, Pinterest, and Snapchat are just a few of the most popular social networking platforms available today.

Even though different social media sites are suitable for different users, Facebook serves as an excellent example of a broad social network. When you sign up for Facebook, you may come across some other individuals who are already members of the site, and you can add them as casual or close friends. As you get more familiar with the site, you may be able to add friends who share your type of interests or find individuals you already know and invite them to join your group. Other individuals may come across your profile on Facebook and attempt to establish a connection with you.

Your connections and hobbies increase the more you use social networking sites like Facebook. It's comparable to networking in real life, such as at a business conference or a social gathering. The more you connect with other people and find that you have similar friends and hobbies, the larger your circle of friends grows.

Each social networking site and the app has its own set of features and points of view, although the majority share certain characteristics. The following terminology will be

encountered whether you're new to Facebook, Twitter, or another social media platform. e.g., Public Profile, Followers And Friends, Shares, Comments, And Likes, Groups, Tagging And Hashtags.

It has been a concern of humans for centuries to interact with friends and family over long distances. People have traditionally relied on communication to strengthen their relationships as social animals. When face-to-face communication is impossible or inconvenient, people have devised a variety of inventive methods.

1.2 HISTORY OF SOCIAL MEDIA

The oldest ways of communication across long distances relied on written writing handed by hand from one person to another. To put it another way, letters. Since the year 550 BC, there has been a continuous evolution of postal service, with the most basic delivery method eventually becoming more ubiquitous and streamlined over the course of several centuries.

The invention of the telegraph occurred in 1792. As a result, messages could be sent across great distances far more quickly than a horse and rider could carry them. Despite the fact that we can send short messages through the telegraph, they were a revolutionary method of sending messages.

The pneumatic post, which was invented in 1865, provided another method for letters to be transported swiftly between their respective receivers. Pneumatic technology-enabled telegraph firms to transmit messages via sealed pipes constructed under the sewers.

In the last decade of the nineteenth century, two significant discoveries were made. The telephone was invented in 1890, while the radio was invented in 1891. Although

the newer versions are more powerful, these technologies are still widely employed. People could talk instantly across large distances using telephone and radio transmissions,

In the twentieth century, technology began to evolve rapidly. In the early days of supercomputers, developers started developing methods for connecting such computers, which eventually resulted in the development of the internet.

In the 1960s, the first versions of the internet were created, such as CompuServe. During this time period, primitive versions of the email were also developed. Users could communicate with one another via a virtual newsletter when UseNet was established in 1979 due to advances in networking technology in the 1970s.

Between the 1970s and the 1980s, the number of personal computers increased as social media steadily became more complex. In the 1980s, Internet Relay Chat (IRC), was developed and widely utilized into the 1990s.

Six Degrees was established in 1997 as the first recognized website of social networking. This allows users to create a profile and establish friends. In 1999, the first blogs were popular and created an impression of social media that remains popular today.

After a few years, Friendster was founded in 2002 to give fair competition to Six Degrees. It allows users to register and join friends, like with Six Degrees. Using this platform, people could also share Audio, Pictures, Videos, Text with other friends. Commenting on other profiles was also enabled in Friendster.

In the early 2000s, websites like LinkedIn and MySpace gained popularity, while websites like Flickr and Photobucket facilitated online photo sharing. YouTube was established in 2005, giving people a whole new way of connecting and sharing

around the world.

Mark Zuckerberg launched Facebook in 2004, and the company has grown rapidly since then. Following its debut, Facebook achieved rapid growth, eventually overtaking MySpace as the most visited website on the internet in 2008.

In 2006, Jack Dorsey founded Twitter. The SMS protocol standard used to have a restriction of 140 characters enforced by mobile operators. The character limit for tweets was raised to 280 in 2017.

In 2010, Mike Krieger and Kevin Systrom co-founded Instagram, which debuted on October 6, 2010. It distinguished itself from the competition by being a smartphone-only application that specialized only in picture and video sharing.

In 2011, Reggie Brown, Bobby Murphy, and Evan Spiegel founded Snapchat. This application's differentiating feature was the ability for users to transmit pictures to each other that would vanish after a time period.

The availability of various social networking sites has increased dramatically in recent years. This created an atmosphere in which users may contact as many people as possible without compromising the privacy of conversation between individuals.

Figure 1.1 Infographic of Social media consumption

Mr. Chadd Callahan and Lori Lewis released an infographic to show that how many people are using social media sites every single minute, as shown in figure 1. According to this research, 973k Facebook users check in every second, over 1 million individuals swipe on Tinder, and over 174,000k users scroll over Instagram. Furthermore, 38 million texts are sent and received on WhatsApp by various individuals. [1]

1.3 HATE SPEECH

According to United Nations, "Any kind of communication in speech, writing or behavior, that attacks or uses pejorative or discriminatory language with reference to a person or a group on the basis of who they are, in other words, based on their religion, ethnicity, nationality, race, color, descent, gender or other identity factors" [2]

A study called "Umati study" divided hate speech into three categories based on their violence. [3]

The three categories are as follows:

1.3.1 OFFENSIVE SPEECH

The majority of the posts in this category are meant to be offensive to a certain group. The speaker typically has limited influence over the public. The text's substance is slightly offensive. It does not normally call on the listener to take destructive action against the targeted group. Statements in this category are unlikely to incite violence.

1.3.2 MODERATELY DANGEROUS SPEECH

Statements in this category are somewhat controversial in nature and are often made by speakers who have minimal to moderate influence on their respective audiences. They may be very provocative to some people while just slightly inflammatory to others,

depending on the remarks' topic. Despite the fact that some of them have the potential to be very provocative, they are included in this section because we take into account the moderate influence the speaker has on the public, which is a factor in the minimal to moderate response the remark got from the listeners.

1.3.3 EXTREMELY DANGEROUS SPEECH

This category includes statements posted by speakers who have moderate to high influence on the audience and dangerously misleading comments and have the greatest potential to instigate violence.

1.3.4 RECOGNITION OF DANGEROUS SPEECH

1.3.4.1. DIRECTED AT A GROUP OF PEOPLE RATHER THAN A SINGLE INDIVIDUAL.

Hate speech encourages listeners to engage in violent actions against certain people, including religious, political, tribal, gender, and ethnic divisions.

a) AGAINST PEOPLE OF DIFFERENT TRIBAL GROUPS

For instance: Some people who are against any tribal group

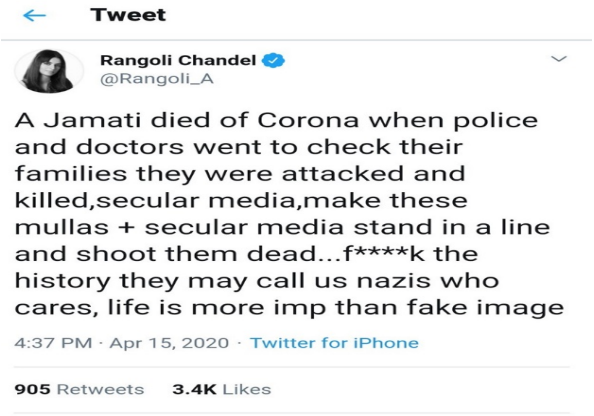


Figure 1.2. Example of tweet against a tribal group

b) AGAINST WOMEN

Unnessacery gender-based comments can not be allowed on social media.

For Instance:

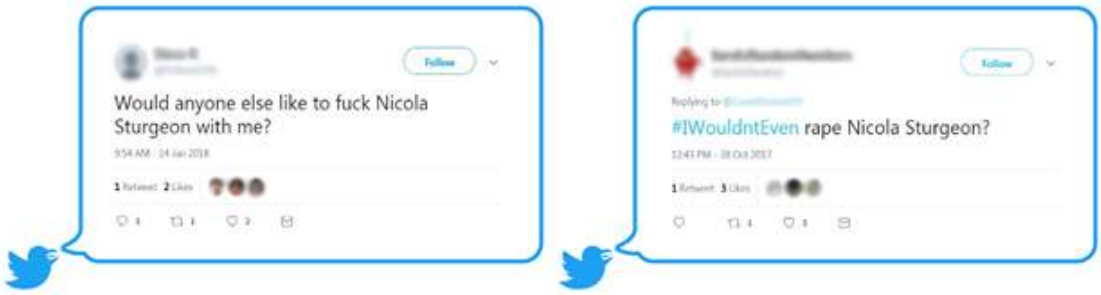


Figure 1.3. Example of tweet against women

c) AGAINST PEOPLE OF DIFFERENT RELIGION

Some people comment on opposite religions and politics could be another reason

For Instance:

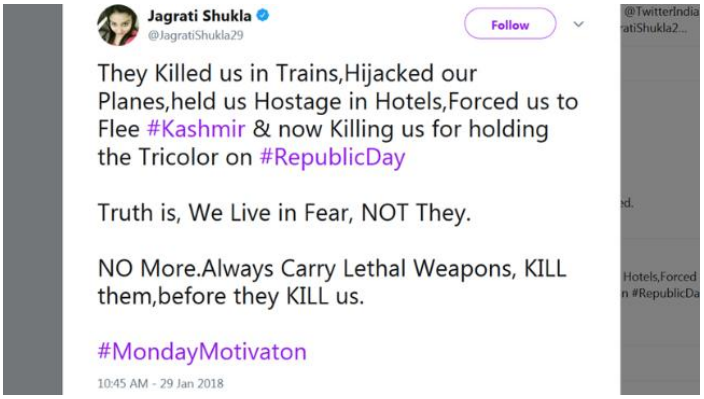


Figure 1.4 Example of tweet against religion

d) AGAINST PEOPLE OF DIffERENT SEXUAL ORIENTATION

For Instance

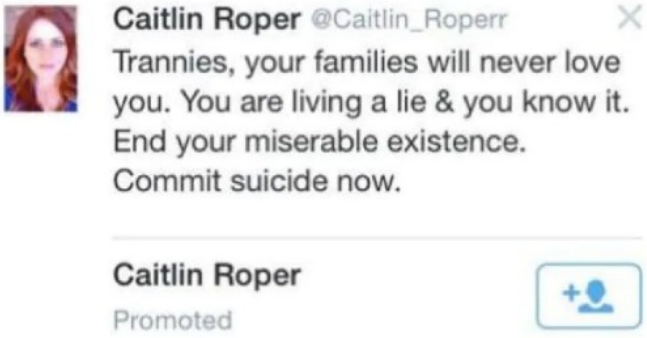


Figure 1.5 Example of tweet against sexuality

e) AGAINST PEOPLE OF DIffERENT RACE

For Example:

Figure 1.6. Example of tweet against race

1.3.4.2 INCLUDE A SIGN CONTAINING TOXIC LANGUAGE

There are three common signs which are in the category of Hate speech.

a. Compare a person as an animal.

b. People set a mindset that the Majority community is in danger from another minority community.

c. Proposes that a group of individuals ruin the integrity of the group of speakers.

1.3.4.3 INSTIGATING THE VIOLENCE

Hate speech often urges the public to support or conduct acts of violence against the targeted group. In Hateful words, the six frequent requests for action are:

1. Riot, **2.** Treat as Inferior, **3.** Loot, **4.** Forcibly Evict, **5.** Beat and **6**. Murder

Hate speech consists of more than simply a few scathing statements. They may be any kind of statement designed to disgrace or degrade a class or group of people or to instigate hate. It may happen online, offline, or on both ways. You may use pictures, emoticons, words, symbols, videos, and notes to communicate. Memes, for instance, maybe images or pictures that seem funny or even good.

Hateful memes may be dismissed by people who engage in hate speech by stating things like “we’re just joking” or “it’s a simple joke,” but hateful memes are significantly involved in hate speeches when taken out of context.

1.3.5 WHY DO PEOPLE USE HATE SPEECH?

There are a number of reasons why people engage in hate speech. Hate speech may sometimes be a valid expression of a person’s ideological opinions or contempt for a certain community of individuals. If the speakers come from a culture where stereotypes are prevalent, hate speech may also be a result of a lack of education, understanding, or thoughtfulness on the subject matter at hand. Members of the targeted

group may not be known to speakers and may not be aware of the stereotyped or harmful language usage. They may also be unaware of the facts. For example, consider the case of someone who makes insulting comments about religion without firsthand information of the teachings of that religion or even the history of prejudice towards that group. It's possible that they have a genuine reason to hate certain individuals within that group and are under the impression that the majority of people within that group have the same bad traits.

Another possibility is that the speaker is deliberately offensive in order to elicit a reaction from others, which is known as "trolling." This is something that is often seen on the internet, where trollers take pleasure in participating in this kind of behavior as a form of recreational activity. Sometimes hate speech is motivated by ignorance, such as when someone uses an offensive term without realizing that they are using an offensive term, which may be unpleasant and harmful.

1.4 CYBER HARASSMENT

Using the Internet and Telecom to harass, control, manipulate or degrade a child, adult, or organization without a direct threat of physical damage is referred to as Cyber Harassment. Cyber harassment involves the use of the internet and is emotional, verbal, social, and sexual abuse by a person, a group, or an organization, unlike physical harassment with face-to-face interaction. The main aim of a cyber harasser is to impose authority and control over the person who has been attacked.

When teenagers are engaged, the word "Cyberbullying" is used to describe Cyber Harassment; however, when direct or indirect physical danger is involved, the phrase "Cyberstalking" is used to describe Cyber Harassment. Another comparable word is

the Internet Troll that is frequently used to describe cyber harassment but is significantly different in attacker type.

There may be numerous kinds of cyber harassment. When the victim becomes aware that their online actions are being monitored, stalking turns into harassment. On the other hand, cyberbullies want the victim to be aware that they are being bullied and will engage in open conversation with them over the internet. In many cases, the following methods are used:

a. Inappropriate comments and posts on Social Media

b. Unusual Emails

c. Harmful Texts

d. Graphical animation and pictures directed at the victim

e. Instant messaging

There are many different kinds of media and information available on the internet. Blogs, forums, and portals that promote hate speech may be found in a number of locations. Despite the fact that hate speech is usually banned, it is fairly unusual to come across it on popular social media sites such as Twitter and Facebook. Although these companies are still trying to make a strong artificial intelligence to remove hate speech, they are not always successful.

We can describe hate speech as abusive language, hatefulness, threats, racism, cyberbullying, aggression, insults, provocation, personal attacks, or sexism. These are some major threats to online social media platforms.

[4] According to a study commissioned by cybersecurity solutions provider Norton by Symantec, eight out of ten individuals in India have encountered some kind of online abuse at some point in their lives. On the internet, 41 percent of women reported being the target of sexual harassment. Norton by Symantec conducted a survey on 1,035 people in India, according to the survey. Most of the people said that online harassment is very prevalent among them. Sixty-three percent of people accepted that they have been facing cyber harassment in their usual life. Almost 50 percent of young people experienced malicious and threatening comments on social networking sites, and 50 percent of them had suffered online trolling. Among 1035 people, 49 percent had got abused by a specific group. This research discovered that, In India, out of the four nations in the Asian countries that were studied (India, Australia, New Zealand, and Japan), had the greatest incidence of online harassment, according to the study figure 1.7.

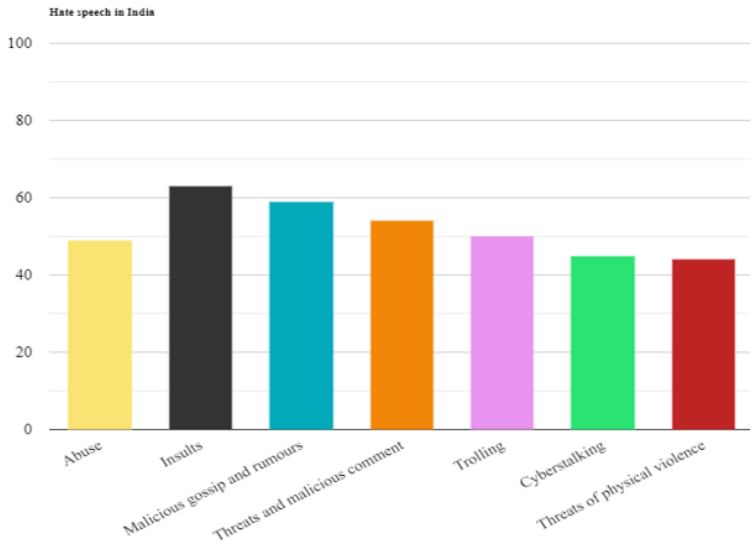


Figure 1.7 Types of Hatespeech Indian citizens are facing in Percentage

Social media platforms are the most pre-eminent fields for such toxic activity. Some social media platforms provide flagging techniques to prevent hateful content, but only 18% of all adults have flagged or reported hate speech or harassing conversation, whereas only 13% of adults have flagged or reported someone for abusive comments. Manual techniques like reporting or flagging comments are not effective and have a risk of favoritism under judgments by human reviewers. As we all know, an automated system can be faster than a human reviewer. A machine learning model to automatically detect online hate speech from social media platforms will be very useful. Online social media platforms, including Twitter, YouTube, Wikipedia, Reddit, etc., have been observed as the most hateful content providers.

CHAPTER - 2

MACHINE LEARNING

ALGORITHMS

2.1 MACHINE LEARNING

Machine learning is a part of Artificial Intelligence. It is basically a set of algorithms that takes a dataset as an input and learns patterns from that dataset. The process of learning data makes the algorithm a model that can solve complex problems. Throughout the learning process, We don't need to write code explicitly.

Each machine learning algorithm uses different calculations to solve the problem, such as sigmoid function, probability function, etc. These algorithms are capable of handling countless features to make the model better to solve a specific problem.

2.2 TYPE OF MACHINE LEARNING

Machine learning can be classified into three basic categories

1. Supervised Machine Learning
2. Unsupervised Machine Learning
3. Reinforcement Machine Learning

2.2.1 SUPERVISED MACHINE LEARNING

Supervised Learning problem contains target label attached with each set of features. Supervised learning algorithms try to find out the relationship between the target variable and the corresponding set of features. A supervised Learning Algorithm can easily solve the following two types of problems.

2.2.1.1 CLASSIFICATION ALGORITHMS

In the classification problem, the Supervised Learning algorithm creates a mapping function that maps a set of features with the target class, and the outcome of the algorithm must be discrete in nature. The outcome can also be referred to as a

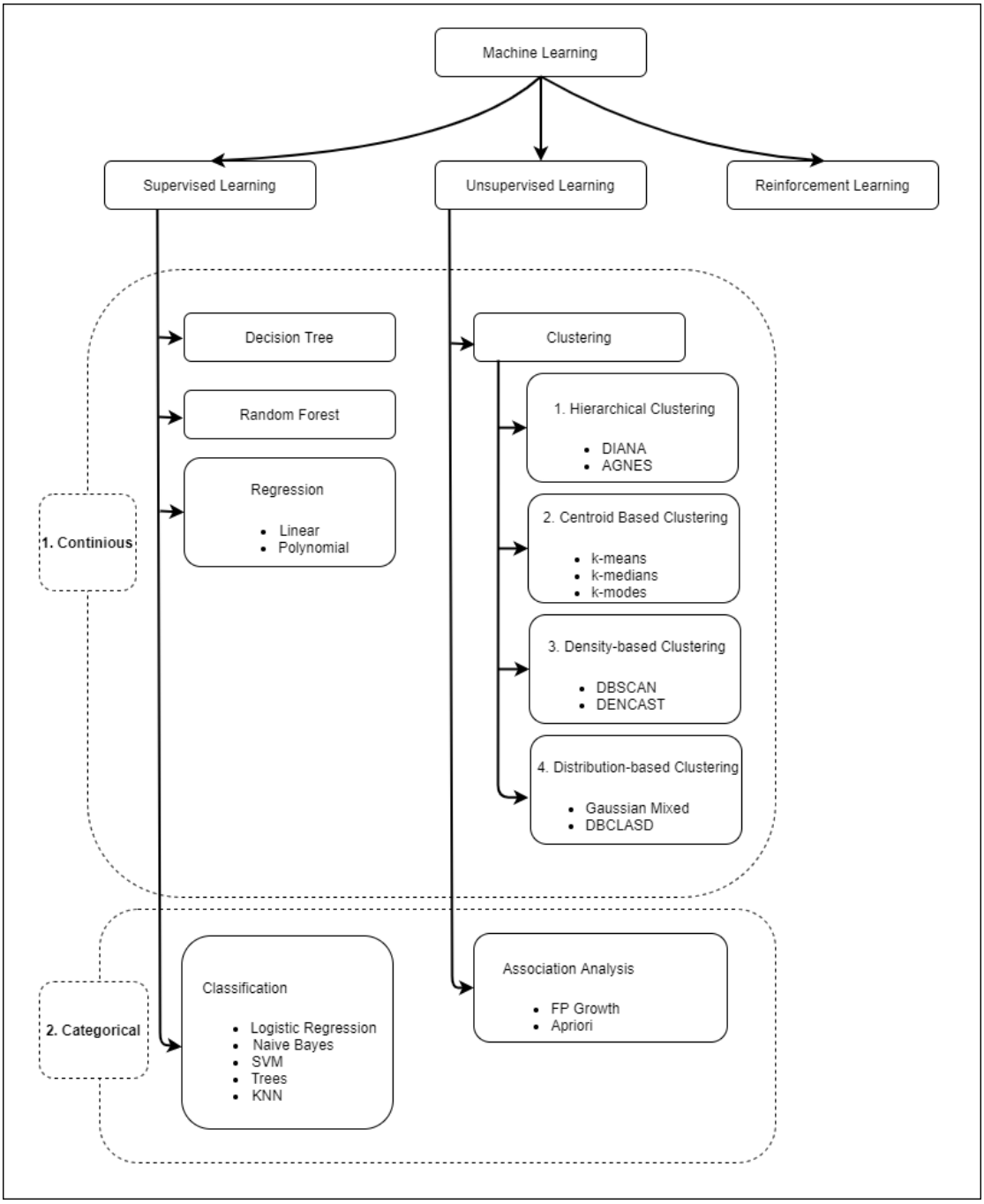


Figure 2.1. Type of Machine Learning

Categorical or class label. Each tuple in the dataset must have a target class/label associated with it that helps to train the model. After training with the labeled dataset, the model is capable of predicting the future without a labeled dataset.

Example. Disease prediction is based on certain characteristics of the Patient's historical report.

a) NAIVE BAYES CLASSIFIER

Naïve Bayes is a classifier that is based on the Bayes Theorem. It has had great success in solving many problems, but it performs particularly well in the context of natural language processing (NLP) issues.

$$P(A \mid B) = \frac{P(A)\ P(B \mid A)}{P(B)} \qquad (1)$$

Where P(A|B) is a probability of A occurring given evidence B has already occurred, P(B) shows Probability of B occurring, P(B|A) shows that Probability of B occurring given evidence A has already happened, and P(A) shows the probability of A occurring.

b) SUPPORT VECTOR MACHINE

SVM learning methods are able to separate or classify any given data point into multiple groups. A support vector machine analyzes the data points and outputs the hyperplane that better divides the groups.

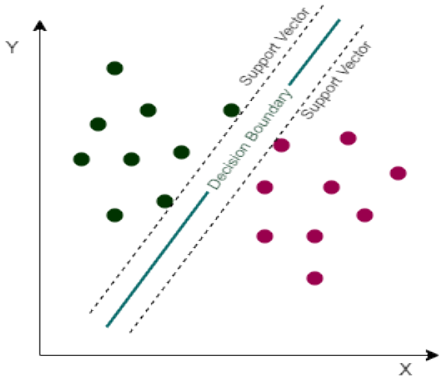


Figure 2.2 Support Vector Machine plot

The resultant line is the decision boundary, and everything that falls on one side of it will be classed as green, and everything that falls on the other will be classed as pink.

c) K-NEAREST NEIGHBOR

K-Nearest Neighbor is a classification and prediction algorithm that is used to identify and predict the class of a given data point based on the distance between the data points. K-Nearest Neighbor theory suggests that points that are close to one another must be identical, and so they will be bound together to form a cluster.

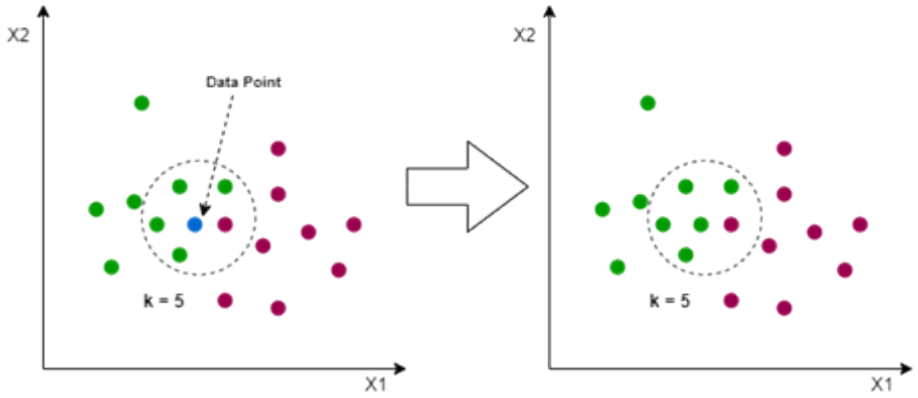


Figure 2.3 K-Nearest Neighbor plot

If there are four green points and one pink point surrounding a data point, this data point is likely to be a green one by majority vote. The “k” in k -NN is a parameter that corresponds to the number of neighbors used in the majority voting mechanism. The value of k in the example above is 5. It is necessary to find the right k value to tune the prediction model correctly, and this is called parameter tuning.

d) LOGISTIC REGRESSION

Logistic regression is nearly equivalent to Linear Regression in that they are implemented in a similar manner. Linear regression approach is used to solve the linear problems and Logistic regression is use to solve classification problems. In logistic

regression, we have an “S” shaped logistic function that predicts only two values (0,1) rather than fitting a regression line. The curve from the logistic function specifies the probability of something, such as whether the tweets are hate speech or not hate speech.

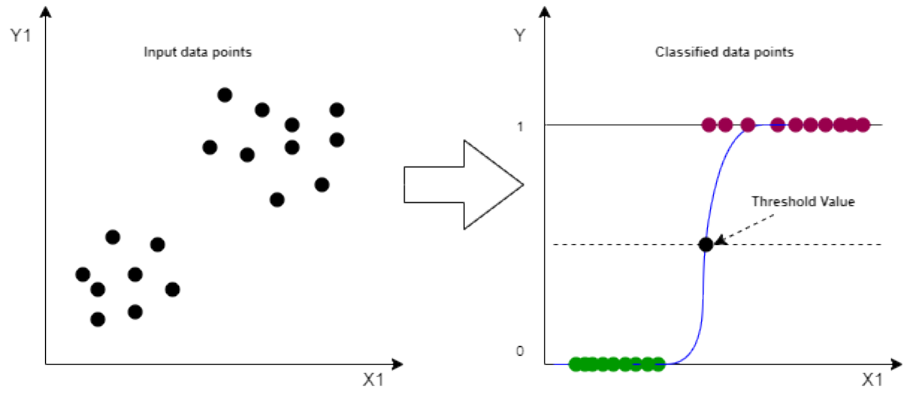


Figure 2.4 Logistic Regression plot

First, we apply the linear equation, and then we apply the Sigmoid function for the result, so we obtain a value which is between 0 and 1. To make the logistic regression work for new data points, we calculate the Maximum log-likelihood and Gradient descent at the end.

2.2.1.2 REGRESSION ALGORITHMS

In the Regression problem, the Supervised learning algorithm creates a map between the set of features with other target classes, but the outcome of the algorithm is continuous. The outcome of the regression problem is a real value that can be integer or float.

Example Price prediction of the house based on different characteristics of the house.

2.2.2 UNSUPERVISED MACHINE LEARNING

Unsupervised Learning techniques use a dataset without a label or target class and find out the hidden pattern from the given dataset. Basically, The main objective of Unsupervised Machine Learning is to group the data according to their similar features.

Unsupervised Learning algorithms can be classified into two different types of algorithms.

2.2.2.1 CLUSTERING ALGORITHMS

Clustering is an unsupervised machine learning algorithm that separates dataset points into smaller subgroups based on their similar features. Datapoint lies into the same cluster consist similar features, whereas data points that exist in two different clusters differ based on features.

a. Hierarchical Clustering (Connectivity-based Clustering)

Hierarchical Clustering is a clustering approach in which either data divides or combines into groups in a hierarchical manner. Based on the direction of the procedure, Hierarchical Clustering can be divided into two types 1. Divisive Approach 2. Agglomerative Approach

❖ **Divisive Approach**

Divisive Hierarchical Clustering is based on the top-down approach in which the whole dataset consider as one cluster initially. It divides the cluster into two clusters based on the sum of squared errors (SSE) of each data point in the cluster. SSE decides which cluster to be separated. To divide the cluster, we use various approaches such as single-

linkage algorithm (MIN), complete linkage algorithm (MAX), Group Average, Ward's Method, Distance between centroids.

- **Agglomerative Approach**

Agglomerative Hierarchical Clustering is based on a bottom-up approach. Each data point in the dataset is considered a cluster. Based on the similarity of two clusters, it merges the two similar clusters into one cluster.

We can visualize the hierarchical Clustering using Dendrogram. The Dendrogram represents the sequence of the divide or merges in a tree-like structure.

b. Model-based methods (Density-based Clustering)

The density-based clustering approach separates the high-density data points area with low-density data points areas.

For example. DBSCAN and DENCAST

c. Distribution based Clustering

Distribution-based Clustering creates a cluster of data based on probability distribution such as Binomial distribution or Gaussian Distribution in the dataset.

For e.g., Gaussian Mixed and DBCLASD algorithm

d. Centroid-based models (Partitioning methods)

Centroid-based clustering algorithms take the number of k clusters which is called centroid and try to find out the distance between the data point and identify the cluster based on the given centroid.

E.g., K-mean, node, and k-median algorithms

2.2.2.2 ASSOCIATION ALGORITHMS

Association Analysis represents the specific relationship between data points in the huge datasets. It represents the relationship as the association rules

For eg. { Bread } → { Milk }

The above rule shows that Bread has a strong relationship with Milk because a customer who buys Bread also buys Milk with it.

Some applications of Association analysis are market basket analysis, bioinformatics, medical diagnosis, web mining, and scientific data analysis.

2.2.3 REINFORCEMENT LEARNING

In the reinforcement learning model, the model tries to learn by interacting with its environment. For instance, if a robot is learning to deliver a courier, it uses various strategies or paths to deliver the courier to the correct address. If the robot delivers the courier to the correct address, then a specific point/reward is included in that strategy taken by a robot, and if the robot doesn't deliver the courier at the correct address, then a specific point is deducted as a punishment. Over time the model learners and improve the strategy by using reward and punishment points.

CHAPTER - 3

LITERATURE REVIEW

3. LITERATURE SUMMARY

Hate speech identification can be accomplished via a variety of methods, including machine learning, deep learning, and the rule-based approach. In this section, we reviewed various approaches that have already been implemented on hate speech identification by various researchers

3.1 SUPERVISED LEARNING APPROACH

Fatahillah et al. (2017) utilized the Naive Bayes Classifier Algorithm to identify hate speech on Instagram using the k-nearest neighbor classifier [5], while Fatahillah et al. (2017) used the Naive Bayes Classifier Algorithm to detect hate speech on Twitter [4]. They obtained the data set from Twitter using the Twitter API and manually annotated the data set once it was collected. Following the pre-processing and feature engineering phases, they used the Naive Bayes Classifier algorithm and discovered that it had a 93 percent accuracy rate in classification.

M. Ali Fauzi and colleagues (2018) developed a method for identifying hate speech by using a collection of supervised learning algorithms [6]. Among the classification methods used were K-Nearest Neighbours, Random Forest, Naive Bayes, Support Vector Machine, and Maximum Entropy, all of which were combined to form an ensemble. They gathered the data set using the Twitter API and manually annotated the information in the data collection. They used tokenization, filtering, stemming, and term weighting techniques throughout the pre-processing phase. They used the bag of words characteristics with TFIDF methods to get their results. Among the five stand-alone classifiers tested, the Naive Bayes algorithm fared the best, with an accuracy of 78.3 percent in the process.

P. Sari and colleagues published a paper in 2019 proposing a method for detecting hate speech on Twitter based on Logistic Regression. [7] They gathered the data from Twitter and used Case Folding, Tokenizing, Filtering, and Stemming techniques in the pre-processing step to further refine the information. For vectorization, the TF-IDF method is employed once the pre-processing step has been completed. The Logistic regression method has been used, and they have discovered that it has an accuracy of 84 percent.

In 2020, Oluwafemi Oriola et al. suggested a method for detecting abusive comments on Twitter 2020 [8]. A dataset was gathered and annotated by the author using the Twitter API, and the data set was divided into two sections: free speech ("FS") and hate speech ("HT"). Pre-processing involves removing special characters, emojis, punctuation, symbols, hashtags, and stopwords from data in order to make it more readable and usable. For feature engineering purposes, they used the TF-IDF method to convert the text into vectors of feature information (feature vectors). They discovered that the accuracy of an optimized support vector machine with n-gram was 89.4 percent after using the algorithm.

In 2020, Annisa Briliani et al. developed a method for detecting hate speech on Instagram using the nearest neighbor classifier [9]. They obtained the data set through Instagram's API and manually marked it. They classified the dataset into two categories, zero and one. They cleansed the data during the pre-processing phase and used the TF-IDF method during the feature engineering step. They then used the k-nearest neighbor method and discovered an accuracy of 98.13 percent.

3.2 UNSUPERVISED LEARNING APPROACH

Rui Zhao et al. (2015) suggested utilizing a Semantic-Enhanced Marginalized Denoising Auto-Encoder to identify cyberbullying [10]. They utilized two data sets from different sources. Twitter is the first source, while Myspace is the second. Twitter data was gathered using the Twitter stream API, while Myspace data was gathered using the site crawling method. They obtained 84.9 percent accuracy with smSDA on the Twitter dataset and 89.7 percent accuracy with smSDA on the MySpace dataset.

Axel Rodrguez et al. (2019) developed a method for detecting hate speech material on Facebook using sentiment analysis [11]. They extracted the post and comments from Facebook using the Graph API. VADER and JAMMIN were employed to eliminate the irrelevant texts. During the pre-processing step, they removed any superfluous stopwords or symbols. TFIDF was used to transform pre-processed documents into vectors. The resultant matrix is used as an input matrix by the k-means clustering method. Sentiment and emotion analysis were used to gather the most unfavorable articles and comments.

Sylvia Jaki et al. (2019) developed a method for detecting hate speech material on Twitter using unsupervised learning [12]. Using the Twitter API, they gathered over 50,00 data sets. They utilized NLP methods to classify the words into clusters. They used spherical k-means clustering and skip-grams to create three groups of the top 250 most biassed words. As a consequence, they have an F1 score of 84.21 percent.

Michele Di Capua et al. (2019) suggested utilizing unsupervised learning to identify cyberbullying. [13] They gathered approximately 54,000 data sets from YouTube and carefully annotated each data set. The SOM-Toolbox-2 platform was used to build the

GHSOM network method. They used a K-fold approach with K = 10 to train and test GHSOM. As a consequence, they have a 64 percent accuracy rate.

3.3 LINGUISTIC RULE-BASED APPROACH

J. Hutto et al. described a technique to classify sentiment using VADER in 2014 [14], which is a rule-based approach to sentiment classification. They began by compiling a list of linguistic characteristics that are extremely responsive to the sentiment of social media postings. They then coupled that set of lexical characteristics with five generic rules encapsulating syntactical and grammatical principles for expressing emotion strength. Finally, they discovered that VADER performed 96 percent accurately on Twitter sentiments while utilizing the rule-based approach.

Dennis Gitari et al. presented a rule-based approach for identifying sentiment analysis in social media text in 2015 [15]. They divided the hate speech issue into three categories in their work: religion, nationality, and race. The primary goal of this article is to create a sentiment analysis-based classification model. Not only can the created model identify subjective statements, but it also classifies and ranks the polarity of emotion expressions. They then associate the semantic and subjective characteristics with hate speech. Finally, utilizing the lexicon-based method, they obtained 71.55 percent accuracy.

3.4 DEEP LEARNING APPROACHES

Hugo Rosa et al. (2018) presented a deep learning-based method for detecting cyberbullying [16]. The training and testing data sets for this study were obtained from Kaggle. Initially, they launched CNN, which has some resemblance to the problem of cyberbullying. It begins with a single-layer CNN and progresses to a fully connected

layer with 0.5 dropouts and softmax performance. Then, to attain optimum accuracy, they integrated CNN-DNN-LSTM. They used TFIDF to represent vectors. Using Google embeddings, they obtained an accuracy of 64.9 percent.

Tin Van Huynh et al. (2019) developed a method for detecting hate speech based on the Bi-GRU-CNN-LSTM Model [17]. They gathered data from Twitter and classified it into three categories in their study (OFFENSIVE, HATE, and CLEAN). After cleaning the data, they used three neural network models to detect hate speech: BiGRU-LSTM-CNN, BiGRU-CNN, and TextCNN. As a consequence, they received a 70.57 percent F1 score.

To identify hate speech on Twitter, Gambäck et al. (2019) used a deep learning system [18]. They gathered data from Twitter and categorized it into four categories (sexism, racism, combination (sexism and racism), and non-hate-speech) in this study. They used four CNN models that were trained using character n-grams, word2vec, and random vectors (word2vec and character n-gram). To enhance the model's accuracy, the author used a 10-fold approach. The word2vec-based CNN model outperformed the other three models, with a 78.3 percent F-score.

3.5 HYBRID BASED APPROACH

To identify hate speech, Viviana Patti et al. (2019) developed a Hybrid-based method. [19] They used two models in this study. They used a linear support vector classifier (LSVC) in their first model and a long short-term memory (LSTM) neural model with word embedding in their second model. They combined 17 categories, including HurtLex, with two types, LSVC and LSTM. Using 68.7 percent of the F1-score, joint

learning with a multilingual word embedding model, including HurtLex, performed best.

Safa Alsafari et al. (2020) developed a methodology for detecting hate speech in Arabic social media [20]. They gathered the data set in this study using the Twitter search API, and the data set is divided into four categories (Religious, Nationality, Gender, and Ethnicity). During the preparation step, they sanitized the data set by eliminating extraneous terms such as URLs, punctuation, symbols, tags, and stop words. They used CNN and Bert to perform a three-class categorization and received 75.51 percent of the F1 score. Both regular validation and on-demand validation may produce significant and sometimes needless network traffic, and the latter eliminates most of the latency savings provided by caching. In such cases, a resource-driven invalidation is a feasible option, in which the server calls a callback on the cache to notify it whenever an update occurs. Although this approach requires the server to keep track of its caches, there will be applications that are prepared to accept these memory costs over the communication overhead of polling-based invalidation.

3.6 COMPARATIVE ANALYSIS

A comparative analysis of hate speech detection is shown in table 3.1 – 3.5

Table 3.1

Linguistic Rule-Based Approach							
Author	**Year**	**Platform**	**Features and Algorithm**	**Precision (%)**	**Recall (%)**	**Accuracy (%)**	**F1-score (%)**
[13]	2014	Micro blogging sites	SentiWordNet, VDER,	-	-	96.0	-
[14]	2015	Twitter, Amazon	LIWC, GI, ANEW, SCN,WSD	81.0	75.0	75.0	-

Table 3.2

Supervised Learning Approach							
Author	**Year**	**Platform**	**Features and Algorithm**	**Precision (%)**	**Recall (%)**	**Accuracy (%)**	**F1-score (%)**
[4]	2017	Twitter	TF-IDF, Naive Bayes	-	-	93.0	-
[5]	2018	Twitter	TF-IDF, Essembled method	-	-	83.4	79.8
[6]	2019	Twitter	TF-IDF, Multinomial Logistic Regression	80.02	82.0	87.68	-
[7]	2020	Twitter	n-gram, Optimized Gradient Boosting	-	-	80.3	-
[8]	2020	Instagram	TF-IDF , K-Nearest Neighbor	94.0	93.0	97.19	93.0

Table 3.3

Unsupervised Learning Approach							
Author	**Year**	**Platform**	**Features and Algorithm**	**Precision (%)**	**Recall (%)**	**Accuracy (%)**	**F1-score (%)**
[9]	2015	MySpace , Twitter	Bag-of-words (BoW), Latent Semantic Analysis (LSA), smSDA	-	-	87.70	77.60
[10]	2019	Facebook	VADER and JAMMIN, TF-IDF, k-means	-	-	74.42	-
[11]	2019	Twitter	n-gram and k-means	84.21	83.97	-	84.21
[12]	2019	Twitter, YouTube , Formspring	GHSOM network algorithm, SOM-Toolbox-2	60.0	94.0	69.0	74.0

Table 3.4

Deep Learning Approach							
Author	**Year**	**Platform**	**Features and Algorithm**	**Precision (%)**	**Recall (%)**	**Accuracy (%)**	**F1-score (%)**
[15]	2018	Kaggle dataset, Formspring, Google,Twitter	CNN-LSTM, Twitter Embedding	84.5	84.2	-	84.2
[16]	2019	Twitter	Bi-GRU-CNN, Bi-GRU-LSTM-CNN, TextCNN,	-	-	-	70.57
[17]	2019	Twitter	CNN, word2vec, character n-grams,	86.61	70.42	-	77.38

Table 3.5

Hybrid Approach							
Author	**Year**	**Platform**	**Features and Algorithm**	**Precision (%)**	**Recall (%)**	**Accuracy (%)**	**F1-score (%)**
[18]	2019	Benchmark corpora	Word embedding, LSVC, LSTM and HurtLex	60.4	79.8	-	68.7
[19]	2020	Twitter	CNN and mBert	76.95	81.52	--	78.99

3.7 CONCLUSION

Various researches have already been done in the field of hate speech detection, and We can categories these researches based on what approach they have been used. There are various approaches to detect hate speech, such as the supervised learning approach, unsupervised learning approach, deep learning approach, and rule-based linguistic approach. Research papers [4-8] are based on supervised learning approaches. Researchers use supervised machine learning algorithms such as Naïve Bayes, Logistic regression, support vector machine, etc., to classify hate speech. Previous work [9-12] is based on an unsupervised learning approach. In these works, researchers employed various unsupervised learning algorithms such as k-mean, Dbscan, LDA, etc. Other researches [15-17] are based on deep learning approaches such as CNN etc. There is a lack of a pure hybrid machine learning approach that is needed to be implemented. We have also found that the accuracy of the model is more dependent on the number of the dataset we use. If we use a fewer number of datasets, the machine learning model's accuracy decreases drastically.

CHAPTER - 4

PROPOSED WORK

4.1 PROBLEM STATEMENT

Every social media platform provides a facility to flag/report any tweet to prevent hateful content from their platforms. But very few people use this technique to report someone's tweet as hate speech or abusive. To overcome this problem, various researchers started working on automatic detection systems using machine learning or deep learning. All recent researches are mostly based on single machine learning algorithm. The accuracy gained by them is very low, between 70% to 90%.

In this work, Our main idea is to create a hybrid machine learning model using six different machine learning algorithms. The performance of the hybrid machine learning model with the TFIDF technique gives the accuracy better than other conventional machine learning models.

4.2 MOTIVATION

Most of the researchers are focusing on straight machine learning problems and solutions. Every single recent research is based on a conventional machine learning model. Our objective was to mix all conventional machine learning models and make a hybrid model which performs better than other conventional machine learning models. The main challenge of using a hybrid model is maintaining the balance between precision, recall, accuracy, f-score and gain accuracy of more than 90 percent in the case of few dataset availabilities.

4.3 PROPOSED GOAL

The objective of this work is to develop an automatic machine learning-based approach for detecting hate speech and cyber harassment. The main aim is to achieve more than 90 percent accuracy with the limited number of the dataset. To achieve the aim, we

classified tweets and posts into two classes (hate speech or Normal speech). In this work, we are focusing on a Hybrid Machine Learning Model for better performance. There are the following procedures which I have taken to accomplish our aim.

4.4 DATA COLLECTION

To accomplish the objective of this work, we collected the dataset from two sources Crowd Flower dataset and Hatespeechdata dataset. Initially, we pre-processed the data on the 55591 rows of the dataset and performed training on 14000 rows of the dataset. The collected dataset contains two columns: label and tweet. Label feature has two classes, 0 and 1, where 0 refers to normal speech and 1 represents hate speech text. Figure 4 is showing the dataset used by this work.

	label	tweet
0	0	@user why is your mother friending your ex on...
1	0	enjoy #lifeofaking starring cuba gooding jr ! ...
2	1	Find Waldo at Escape bitches ! 🔴&#9898...
3	1	@Liveitupjersey yeah, I seen you tweet about b...
4	1	sea shepherd suppoers are racist! #antirac...
...	...	...
55586	0	off to italy tomorrow for some client pr work!...
55587	1	Okay but your number ain't been saved for how ...
55588	1	RT @tmaeleen: this bitch just blew my high &am...
55589	0	@user today at @user for @user ðŸ□□□ðŸ□□□ðŸ□□□ðŸ□...
55590	1	RT @Hi_____Bye: "@UglyAssAyeKay: Ugh I fu...

55591 rows × 2 columns

Figure 4.1 Dataset

4.5 DATA IMBALANCE

Imbalance data in the dataset gives good accuracy in numbers, but it can be misleading if data is unequal in the dataset. It is necessary to balance the dataset before using it in

the model. In this work, we have two classes, “Hate speech” and “Normalspeech.”. During the pre-processing phase, we balanced the dataset with 24995 rows of Hatespeech data and 24929 rows of the Normal speech dataset.

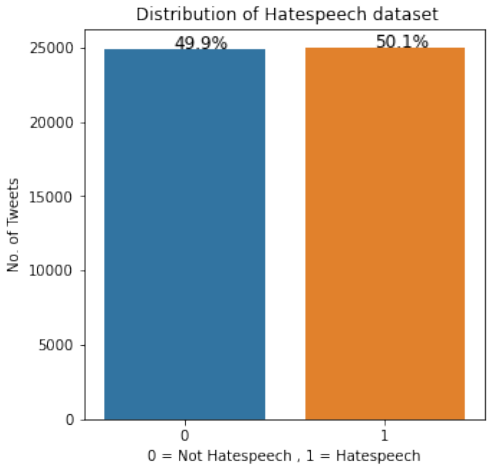


Figure 4.2 Bar plot representation of Balanced dataset

4.6 DATA PRE-PROCESSING

The collected dataset may include unnecessary and redundant data that may reduce the accuracy of the final model. Redundant data and unnecessary characters can be responsible for reducing the model’s performance. Therefore, Preprocessing is a very important procedure to be employed before feeding it to the machine learning model.

The raw data collected contains inconsistent, redundant, incomplete, and duplicate data that are incapable of providing greater accuracy. Thus it degrades the system’s performance. This is a vital step that must be performed before the dataset is sent to the machine learning model/algorithm. The different pre-processing steps are shown in Figure 4.3.

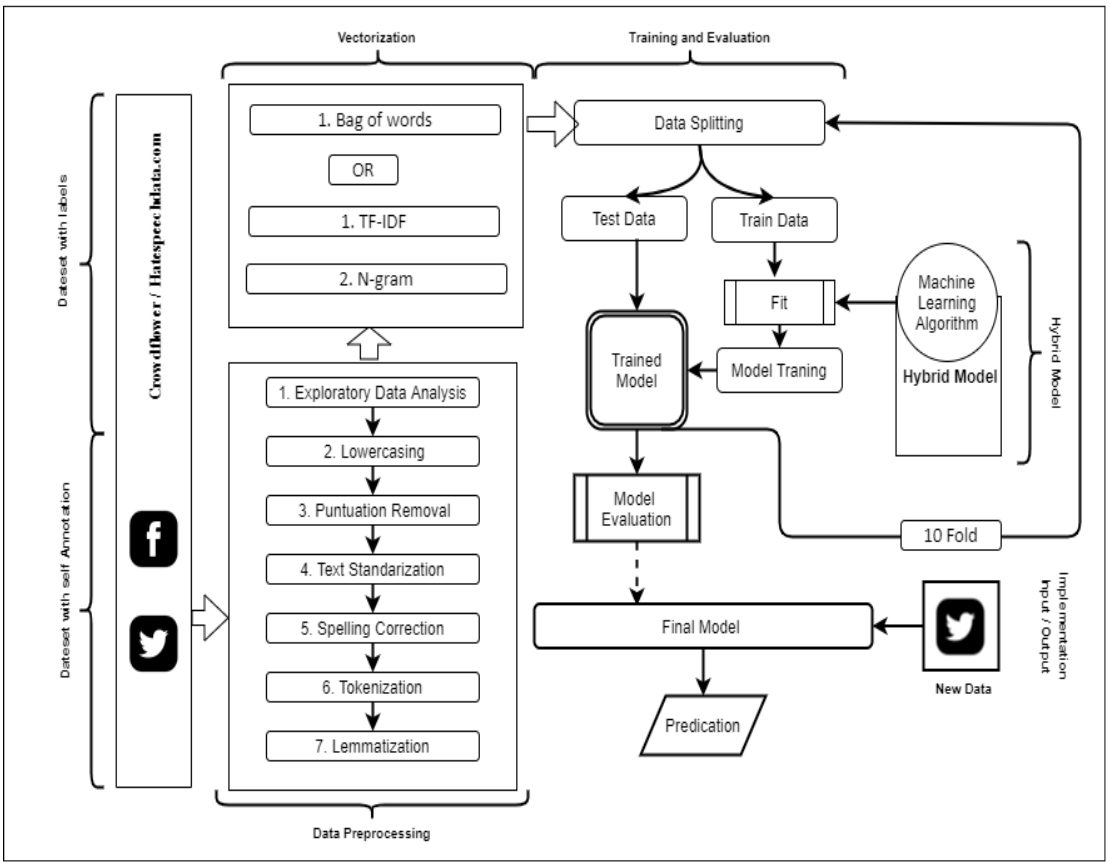


Figure 4.3 Framework to detect hate speech

The following steps are taken to prepare the dataset for processing:

4.6.1 LOWERING TEXTS

In this Phase, We transform the dataset. This phase is required in order to vectorize the dataset, and it must be completed at the beginning of the pre-processing phase.

4.6.2 LABEL ENCODING

Our dataset contains two classes, “hate speech” and “NormalSpeech.”. The Machine Learning algorithm does not work with textual data, so we must convert the textual data into numerical format. Label encoding is the method to transform the categorical

text into numerical form. In the case of our dataset, we transform the “Hate speech” class into 1, and Normalspeech class turned into 0.

4.6.3 REMOVED ACCENTED CHARACTERS

Some meaningless information such as emojis and accented characters are useless for our research. So we have removed all unnecessary accented and emojis characters.

4.6.4 REMOVED PUNCTUATION

Generally, all social media users use some special characters such as @(Usertag) and #(Hashtag). These special character does not contain any useful information. Thus, we removed all useless special characters and numbers.

4.6.5 STOP WORDS

Stop words are words that do not include any special information in the sentence, and if we remove stop words from the sentence, it does not modify the meaning of the sentence. In this work, we have employed Spacy Library to eliminate stopwords from the dataset because Spacy library provides 326 stopwords. We have also found some extra stopwords in my dataset which is not available in the Spacy library, so we created a list of custom stopwords

4.6.6 LEMMATIZATION

We use verbs in our sentences to express the tense. Verbs can be in any form, such as the present, past, or past participle. Transforming the verb’s various forms into base form is called lemmatization. We have used lemmatization to transform the various verb form into base forms.

4.6.7 STEPS INVOLVED IN DATA PRE-PROCESSING

Step 1: Import Dataset

```
import pandas as pd
df = pd.read_csv(r"..\Raw_Dataset\combined_dataset.csv")
df
```

	Unnamed: 0	label	tweet
0	0	1	bad bitch
1	1	0	@user here's what's happening twitter... i'm ...
2	2	1	nothing worse than a raw cock....oh wait....#MKR
3	3	0	@user #fathersday gift for proud dads.avail...
4	4	1	The bitch is freeee!!!! ;D
...	...	...	...
55586	55586	0	he put a ring on it ð□□□ð□□□ð□□¯ #proposal #s...
55587	55587	0	i am thankful for good health. #thankful #posi...
55588	55588	1	"Walked in that bitch like, NIGGA WE MADE IT!"
55589	55589	0	well done wood green sd on incentive win!!! #3...
55590	55590	0	@user #review #2016 out today featuring someon...

55591 rows × 3 columns

Figure 4.4 Import Dataset

Step 2: Delete Unnecessary columns

```
df.drop(labels = ['Unnamed: 0'], axis = "columns", inplace=True)
df.head(3)
```

	label	tweet
0	1	bad bitch
1	0	@user here's what's happening twitter... i'm ...
2	1	nothing worse than a raw cock....oh wait....#MKR

Figure 4.5 Delete Unnecessary columns

Step 3: Removing Duplicate Tuples

```
df.drop(labels = ['Unnamed: 0'], axis = "columns", inplace=True)
df.head(3)
```

	label	tweet
0	1	bad bitch
1	0	@user here's what's happening twitter... i'm ...
2	1	nothing worse than a raw cock....oh wait....#MKR

Figure 4.6 Removing Duplicate Tuples

Step 4: Balancing the dataset

```
from sklearn.utils import shuffle
df = shuffle(df)
print(df.groupby('label').count())
label0 =  df.loc[df.label==0][:25000]
label1 =  df.loc[df.label==1][:25000]
frames = [label0,label1]
df = pd.concat(frames)
df
from sklearn.utils import shuffle
df = shuffle(df)
df
```

```
       tweet
label
0      27517
1      25598
```

	label	tweet
6931	1	#realtalk if you a hoe an try to get serious w...
46955	0	#racket #broken #yonex #nanoray #white #blue...
14050	1	@YeridZee bitches better watch out because we'...
8971	0	so tired of dealing with customers, i have lit...
43032	1	The little bitch told me sloppy top was a hopp...
...	...	...
49766	0	kids out here complaining about the schools th...
12368	1	@kaitlynshae33 You say this based on what?
2978	1	tricking off ya paper ain't wat u do... gotta ...
54699	1	RT @jessielee1028: @Hawkdaddy300 @GavinMHawks2...
5148	0	@user ahhhh might have guessed #euro2016

Figure 4.7 Balancing the dataset

Step 5: Separating Independent and Dependent Features

```
# 5. Separating Independent and Dependent features (Optional)
```

```
X = df["tweet"]
y = df["label"]
```

Figure 4.8 Separating Independent and Dependent Features

Step 6: Tokenization

```
from textblob import TextBlob

def tokenize(string):
    return(TextBlob(string).words)

df['tokenized_tweet'] = df['tweet'].apply(tokenize)
df.head(5)
```

	label	tweet	tokenized_tweet
0	1	#realtalk if you a hoe an try to get serious w...	[realtalk, if, you, a, hoe, an, try, to, get, ...
1	0	#racket #broken #yonex #nanoray #white #blue...	[racket, broken, yonex, nanoray, white, blue, ...
2	1	@YeridZee bitches better watch out because we'...	[YeridZee, bitches, better, watch, out, becaus...
3	0	so tired of dealing with customers, i have lit...	[so, tired, of, dealing, with, customers, i, h...
4	1	The little bitch told me sloppy top was a hopp...	[The, little, bitch, told, me, sloppy, top, wa...

Figure 4.9 Tokenization

Step 7: Removing (URL, EMOJI, NUMBER, SMILEY, RESERVED, MENTION) and Lowercasing

```
#Returns cleaned list ( remove URL, EMOJI, NUMBER, SMILEY, RESERVED, MENTION)
#!pip install tweet-preprocessor
import preprocessor as p
p.set_options(p.OPT.URL, p.OPT.EMOJI, p.OPT.NUMBER, p.OPT.SMILEY,p.OPT.RESERVED, p.OPT.MENTION)

def remove_U_E_N_S_R_M_lower(df_tweet):
    temp_list2 =[]
    for x in df_tweet:
        temp_list1 =[]
        for y in x:
            temp_list1.append(p.clean(y).lower())       #converting all into lower case first
        temp_list2.append(temp_list1)
    return(temp_list2)

df['preproc_step_1'] = remove_U_E_N_S_R_M_lower(df['tokenized_tweet'])
df.head()
```

	label	tweet	tokenized_tweet	preproc_step_1
0	1	#realtalk if you a hoe an try to get serious w...	[realtalk, if, you, a, hoe, an, try, to, get, ...	[realtalk, if, you, a, hoe, an, try, to, get, ...
1	0	#racket #broken #yonex #nanoray #white #blue...	[racket, broken, yonex, nanoray, white, blue, ...	[racket, broken, yonex, nanoray, white, blue, ...
2	1	@YeridZee bitches better watch out because we'...	[YeridZee, bitches, better, watch, out, becaus...	[yeridzee, bitches, better, watch, out, becaus...
3	0	so tired of dealing with customers, i have lit...	[so, tired, of, dealing, with, customers, i, h...	[so, tired, of, dealing, with, customers, i, h...
4	1	The little bitch told me sloppy top was a hopp...	[The, little, bitch, told, me, sloppy, top, wa...	[the, little, bitch, told, me, sloppy, top, wa...

Figure 4.10 Removing unnecessary characters and Lowercasing

Step 8: Removing Accented Characters

```
import unicodedata
# Remove accented characters like é, ñ etc.

def remove_accented_chars(df_tweet):
    temp_list2 = []
    for row in df_tweet:
        temp_list1 = []
        for word in row:
            if(word != ""):
                temp_word = unicodedata.normalize('NFKD',word).encode('ascii','ignore').decode('utf-8','ignore'
                temp_list1.append(temp_word)
            else:
                continue
        temp_list2.append(temp_list1)
    return(temp_list2)


df['preproc_step_2'] = remove_accented_chars(df['preproc_step_1'])
df.head()
```

	label	tweet	tokenized_tweet	preproc_step_1	preproc_step_2
0	1	#realtalk if you a hoe an try to get serious w...	[realtalk, if, you, a, hoe, an, try, to, get, ...	[realtalk, if, you, a, hoe, an, try, to, get, ...	[realtalk, if, you, a, hoe, an, try, to, get, ...
1	0	#racket #broken #yonex #nanoray #white #blue...	[racket, broken, yonex, nanoray, white, blue, ...	[racket, broken, yonex, nanoray, white, blue, ...	[racket, broken, yonex, nanoray, white, blue, ...
2	1	@YeridZee bitches better watch out because we'...	[YeridZee, bitches, better, watch, out, becaus...	[yeridzee, bitches, better, watch, out, becaus...	[yeridzee, bitches, better, watch, out, becaus...
3	0	so tired of dealing with customers, i have lit...	[so, tired, of, dealing, with, customers, i, h...	[so, tired, of, dealing, with, customers, i, h...	[so, tired, of, dealing, with, customers, i, h...
4	1	The little bitch told me sloppy top was a hopp...	[The, little, bitch, told, me, sloppy, top, wa...	[the, little, bitch, told, me, sloppy, top, wa...	[the, little, bitch, told, me, sloppy, top, wa...

Figure 4.11 Removing Accented Characters

Step 9: Removing Default Stop words and Custom Stop words

```
import spacy
en = spacy.load('en_core_web_sm')
stop_words = en.Defaults.stop_words


def remove_stop_words(df_tweet):
    temp_list1 = []
    for row in df_tweet:
        temp_list1.append([word for word in row if word.lower() not in stop_words])
    return(temp_list1)


df['preproc_step_3'] = remove_stop_words(df['preproc_step_2'])
df.head()
```

	label	tweet	tokenized_tweet	preproc_step_1	preproc_step_2	preproc_step_3
0	1	#realtalk if you a hoe an try to get serious w...	[realtalk, if, you, a, hoe, an, try, to, get, ...	[realtalk, if, you, a, hoe, an, try, to, get, ...	[realtalk, if, you, a, hoe, an, try, to, get, ...	[realtalk, hoe, try, expect, piece, meat]
1	0	#racket #broken #yonex #nanoray #white #blue...	[racket, broken, yonex, nanoray, white, blue, ...	[racket, broken, yonex, nanoray, white, blue, ...	[racket, broken, yonex, nanoray, white, blue, ...	[racket, broken, yonex, nanoray, white, blue, ...
2	1	@YeridZee bitches better watch out because we'...	[YeridZee, bitches, better, watch, out, becaus...	[yeridzee, bitches, better, watch, out, becaus...	[yeridzee, bitches, better, watch, out, becaus...	[yeridzee, bitches, better, watch, running, ta...
3	0	so tired of dealing with customers, i have lit...	[so, tired, of, dealing, with, customers, i, h...	[so, tired, of, dealing, with, customers, i, h...	[so, tired, of, dealing, with, customers, i, h...	[tired, dealing, customers, literally, reached...
4	1	The little bitch told me sloppy top was a hopp...	[The, little, bitch, told, me, sloppy, top, wa...	[the, little, bitch, told, me, sloppy, top, wa...	[the, little, bitch, told, me, sloppy, top, wa...	[little, bitch, told, sloppy, hoppy, gave, slo...

Figure 4.12 Removing Default Stop words and Custom Stop words

Step 10: Removing Punctuation

```
import re
import preprocessor as p

def remove_punctuation(df_tweet):
    PUNCTUATION = re.compile("(\")|(\>)|(\})|(\])|(\{)|(\;)|(\<)|(\%)|(\|)|(\()|(\))|(\$)|(\!)|(\.)|(\[)|(\?)|
    PUNCTUATION_NEEDS_SPACE = re.compile("([0-9])|(:)|(_)|(-)|(<br)|(/)|(\s\s+)|(?<!\w)\w(?!\w)")
    temp_list2 = []
    for row in df_tweet:
        temp_list1 = []
        for word in row:
            if(word != ""):
                temp_word = PUNCTUATION.sub("", word)
                temp_word = PUNCTUATION_NEEDS_SPACE.sub(" ",temp_word)
                temp_list1.append(temp_word)
            else:
                continue
        temp_list2.append(temp_list1)
    return(temp_list2)

df['preproc_step_5'] = remove_punctuation(df['preproc_step_4'])
df.head()
```

	label	tweet	tokenized_tweet	preproc_step_1	preproc_step_2	preproc_step_3	preproc_step_4	preproc_step_5
0	1	#realtalk if you a hoe an try to get serious w...	[realtalk, if, you, a, hoe, an, try, to, get, ...	[realtalk, if, you, a, hoe, an, try, to, get, ...	[realtalk, if, you, a, hoe, an, try, to, get, ...	[realtalk, hoe, try, expect, piece, meat]	[realtalk, hoe, try, expect, piece, meat]	[realtalk, hoe, try, expect, piece, meat]
1	0	#racket #broken #yonex #nanoray #white #blue...	[racket, broken, yonex, nanoray, white, blue, ...	[racket, broken, yonex, nanoray, white, blue, ...	[racket, broken, yonex, nanoray, white, blue, ...	[racket, broken, yonex, nanoray, white, blue, ...	[racket, broken, yonex, nanoray, white, blue, ...	[racket, broken, yonex, nanoray, white, blue, ...
2	1	@YeridZee bitches better watch out because we'...	[YeridZee, bitches, better, watch, out, becaus...	[yeridzee, bitches, better, watch, out, becaus...	[yeridzee, bitches, better, watch, out, becaus...	[yeridzee, bitches, better, watch, running, ta...	[yeridzee, bitches, better, watch, running, ta...	[yeridzee, bitches, better, watch, running, ta...

Figure 4.13 Removing Punctuation

Step 11: Lemmatization

```
import spacy
nlp = spacy.load('en_core_web_sm')

def lemmatization(df_tweet):
    temp_list2 = []
    for row in df_tweet:
        temp_list1 = []
        listToStr = ' '.join([element for element in row])
        for word in nlp(listToStr):
            lemma = word.lemma_
            temp_list1.append(lemma)
        temp_list2.append(temp_list1)
    return(temp_list2)


df['preproc_step_6'] = lemmatization(df['preproc_step_5'])
df.head()
```

	label	tweet	tokenized_tweet	preproc_step_1	preproc_step_2	preproc_step_3	preproc_step_4	preproc_step_5	preproc_step_6
0	1	#realtalk if you a hoe an try to get serious w...	[realtalk, if, you, a, hoe, an, try, to, get, ...	[realtalk, if, you, a, hoe, an, try, to, get, ...	[realtalk, if, you, a, hoe, an, try, to, get, ...	[realtalk, hoe, try, expect, piece, meat]	[realtalk, hoe, try, expect, piece, meat]	[realtalk, hoe, try, expect, piece, meat]	[realtalk, hoe, try, expect, piece, meat]
1	0	#racket #broken #yonex #nanoray #white #blue...	[racket, broken, yonex, nanoray, white, blue, ...	[racket, broken, yonex, nanoray, white, blue, ...	[racket, broken, yonex, nanoray, white, blue, ...	[racket, broken, yonex, nanoray, white, blue, ...	[racket, broken, yonex, nanoray, white, blue, ...	[racket, broken, yonex, nanoray, white, blue, ...	[racket, break, yonex, nanoray, white, blue, b...
2	1	@YeridZee bitches better watch out because we'...	[YeridZee, bitches, better, watch, out, becaus...	[yeridzee, bitches, better, watch, out, becaus...	[yeridzee, bitches, better, watch, out, becaus...	[yeridzee, bitches, better, watch, running, ta...	[yeridzee, bitches, better, watch, running, ta...	[yeridzee, bitches, better, watch, running, ta...	[yeridzee, bitch, well, watch, run, table, ton...

Figure 4.14 Lemmatization

Step 12: Removing single and double remaining characters

```
def remove_singleChar(df_tweet):
    SINGLE_OR_DOUBLE_CHARACTERS = re.compile(r'(?<![\w\-])\w(?![\w\-])|(?<![\w\-])\w\w(?![\w\-])')
    temp_list2 = []
    for row in df_tweet:
        temp_list1 = []
        for word in row:
            if(word != ""):
                temp_word = SINGLE_OR_DOUBLE_CHARACTERS.sub("",word)
                temp_list1.append(temp_word)
            else:
                continue
        temp_list2.append(temp_list1)
    return(temp_list2)

df['preproc_step_7'] = remove_singleChar(df['preproc_step_6'])
df.head()
```

	label	tweet	tokenized_tweet	preproc_step_1	preproc_step_2	preproc_step_3	preproc_step_4	preproc_step_5	preproc_step_6	preproc_s
0	1	#realtalk if you a hoe an try to get serious w...	[realtalk, if, you, a, hoe, an, try, to, get, ...	[realtalk, if, you, a, hoe, an, try, to, get, ...	[realtalk, if, you, a, hoe, an, try, to, get, ...	[realtalk, hoe, try, expect, piece, meat]	[realtalk, hoe, try, expect, piece, meat]	[realtalk, hoe, try, expect, piece, meat]	[realtalk, hoe, try, expect, piece, meat]	[realtalk try, e piece,
1	0	#racket #broken #yonex #nanoray #white #blue...	[racket, broken, yonex, nanoray, white, blue, ...	[racket, broken, yonex, nanoray, white, blue, ...	[racket, broken, yonex, nanoray, white, blue, ...	[racket, broken, yonex, nanoray, white, blue, ...	[racket, broken, yonex, nanoray, white, blue, ...	[racket, broken, yonex, nanoray, white, blue, ...	[racket, break, yonex, nanoray, white, blue, b...	[racket, yonex, na white, bl

Figure 4.15 Removing single and double remaining characters

Step 13 Untokenized final pre-processed dataset

```
from nltk.tokenize.treebank import TreebankWordDetokenizer
def untokenize(df_tweet):
    return(df_tweet.apply(TreebankWordDetokenizer().detokenize))
df['preproc_step_8'] = untokenize(df['preproc_step_7'])
```

```
df.head(10)
```

	label	tweet	tokenized_tweet	preproc_step_1	preproc_step_2	preproc_step_3	preproc_step_4	prepr
0	1	RT @EmilyFaye2: My signature white-bitch move ...	[RT, EmilyFaye2, My, signature, white-bitch, m...	[, emilyfaye2, my, signature, white-bitch, mov...	[emilyfaye2, my, signature, white-bitch, move,...	[emilyfaye2, signature, white-bitch, profusely...	[emilyfaye2, signature, white-bitch, profusely...	[signat bitch,
1	1	i love bad bitches thats my fucking problem	[i, love, bad, bitches, thats, my, fucking, pr...	[i, love, bad, bitches, thats, my, fucking, pr...	[i, love, bad, bitches, thats, my, fucking, pr...	[love, bad, bitches, thats, fucking, problem]	[love, bad, bitches, thats, fucking, problem]	[love, ba that
2	1	who's pussy better then a crazy bitch?	[who, 's, pussy, better, then, a, crazy, bitch]	[who, 's, pussy, better, then, a, crazy, bitch]	[who, 's, pussy, better, then, a, crazy, bitch]	[pussy, better, crazy, bitch]	[pussy, better, crazy, bitch]	[pussy, be
3	0	@user its official- we're making a tv series!...	[user, its, official, we, 're, making, a, tv, ...	[user, its, official, we, 're, making, a, tv, ...	[user, its, official, we, 're, making, a, tv, ...	[user, official, making, tv, series, theyard, ...	[official, making, tv, series, theyard, teamra...	[official, r serie

Figure 4.16 Separating Independent and Dependent Features

i. WORD FREQUENCY

Word frequency is the way to visualize the top occurring words in the dataset. Figure 3 shows the top 20 most frequent words in our Hatespeech dataset.

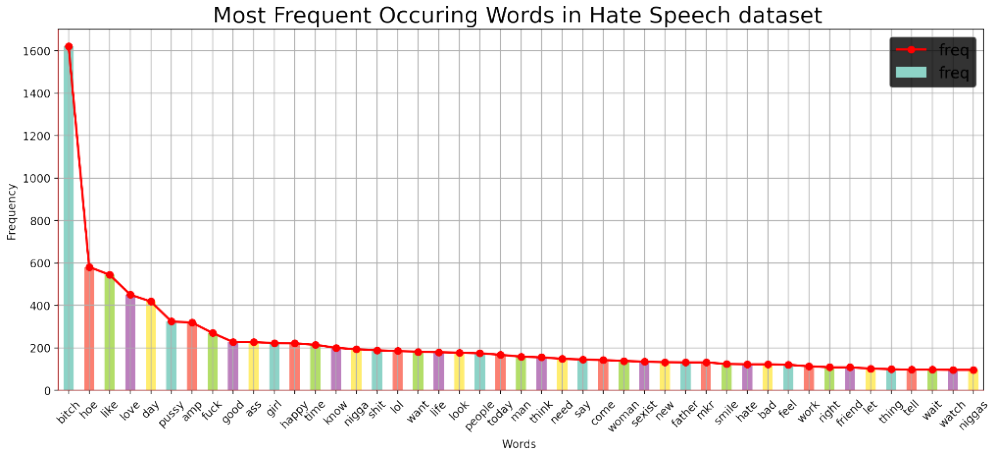


Figure 4.17 Word frequency

ii. WORD CLOUD

After pre-processing phase, We represented a word cloud view on the result dataset. Figure 8 is showing the word cloud view of our dataset, and the most frequent word is in a bigger size, and less frequent words are referred to as small-sized word

Figure 4.18 Word Cloud

4.7 FEATURE EXTRACTION

In the Machine Learning classification model, It is necessary to extract the important features from the texts in order to train the model. The process of transforming the text into features is called feature extraction. We use various feature extraction techniques to extract the features from the texts. In this work, we employed TFIDF and Bag of Words to extract the important features from the textual data. The result of this process will then be given to a machine learning model to classify the text into two categories.

4.7.1 BAG OF WORDS

Bag of words is a technique to transform the unstructured text into vectors by calculating the frequency of the word in that text. Vectorization is another term for this procedure. At first, we transform each word into a feature. Then we calculate the frequency of each word in the document. Bag of words only focuses on how many times a word occurs in the document, but it does not give any information about the word's location.

In this work, we are using the bag of words technique to compare the result with the TFIDF technique, which is another method for vectorization.

4.7.2 TERM FREQUENCY-INVERSE DOCUMENT FREQUENCY (TFIDF)

TFIDF is used to evaluate the weight of the term based on how important that term is in the corpus. TFIDF weight increases based on the frequency of the word but decreases by the number of words in the corpus.

TFIDF contains two terminologies. The first terminology is Term frequency. It is the ratio between the number of times a term occurred in the document and the total number of words in that document. The second terminology is Inverse Document frequency. It can be calculated by taking the logarithm with the ratio of the total number of documents and the number of documents contains term t

a. Term Frequency(TF)

Term frequency calculates the frequency of the word in a single document.

$$TF(w) = \frac{(frequency\ of\ the\ word\ w\ in\ a\ document)}{(Total\ number\ of\ words\ in\ that\ document)} \qquad (1)$$

b. Inverse Document Frequency(IDF)

It evaluates how important a word is in the corpus. Such as “she” or “they” appears very frequently in the corpus, but these words do not include that much information. So using IDF, we decrease the weightage of most frequent words and increase the weightage of the words that are rare in the document.

$$IDF(w) = \log \frac{(Number\ of\ times\ word\ w\ occurs\ in\ a\ document)}{(Total\ number\ of\ words\ in\ that\ document)} \quad (2)$$

In this work, we are using Bag of words and TFIDF both to evaluate the final result.

4.8 METHODOLOGY

In this work, we have implemented a Multi-layers Hybrid Machine Learning model for better accuracy. To develop the model, we have created two layers of machine learning algorithm, and in the first layer, we are using Decision Tree, Naive Bayes, Support Vector Machine, Logistic Regression, Random Forest, and for the second layer, we are using logistic regression algorithm. To accomplish our objective, we have followed the following procedure.

4.8.1 IMPLEMENTATION OF MULTI-LAYER HYBRID ML MODEL

In this work, we are using a hybrid machine learning model to improve the performance of the model. In order to put this concept into action, we have divided the hybrid algorithm into two layers: Baselayer and Meta Layer. Using six different machine learning algorithms, we have built the base layer of our model, including Decision Tree, Support Vector Machine, Random Forest, K-Neighbor Classifier, Naive Bayes, Logistic Regression.

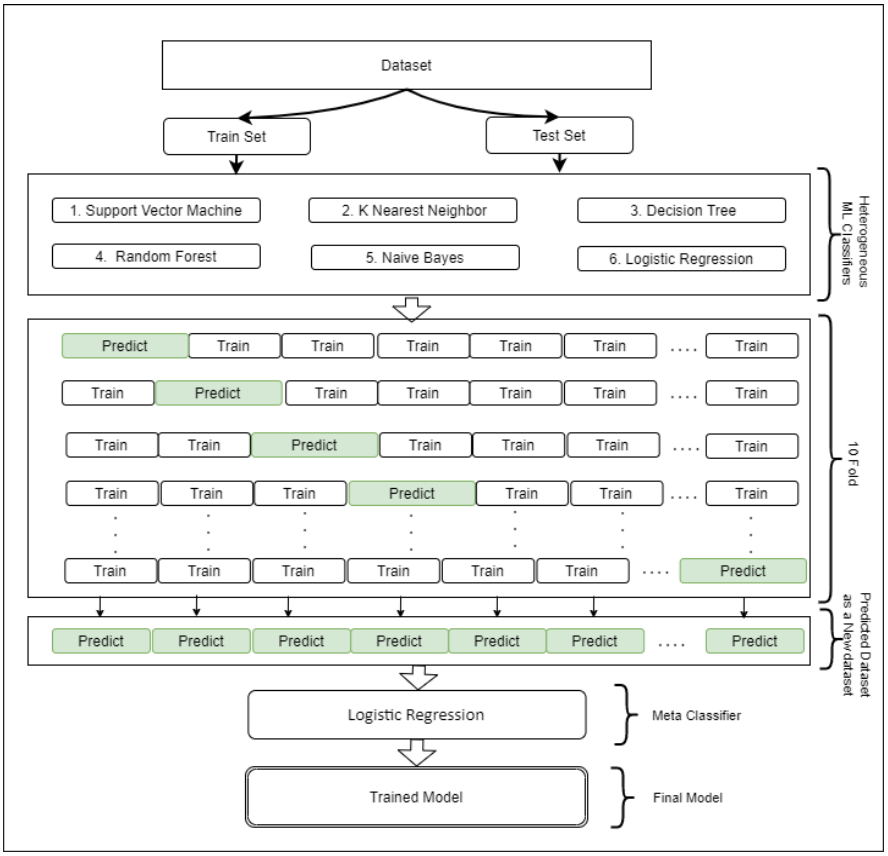


Figure 4.19 Hybrid Machine Learning Implementation Flow chart

We are using the following formula to implement our Hybrid machine learning model

$$\sum_{j=1}^{n}\left(\left(\sum_{i=1}^{m} ML_i\right)+LM_j\right)+GM \qquad \begin{cases} n \in \mathrm{N} \\ n \geq 2 \end{cases} \qquad (3)$$

Where, ML_i = Combination Machine Learning Models

LM_j = Local Meta Classifiers

GM = Global Meta Classifier

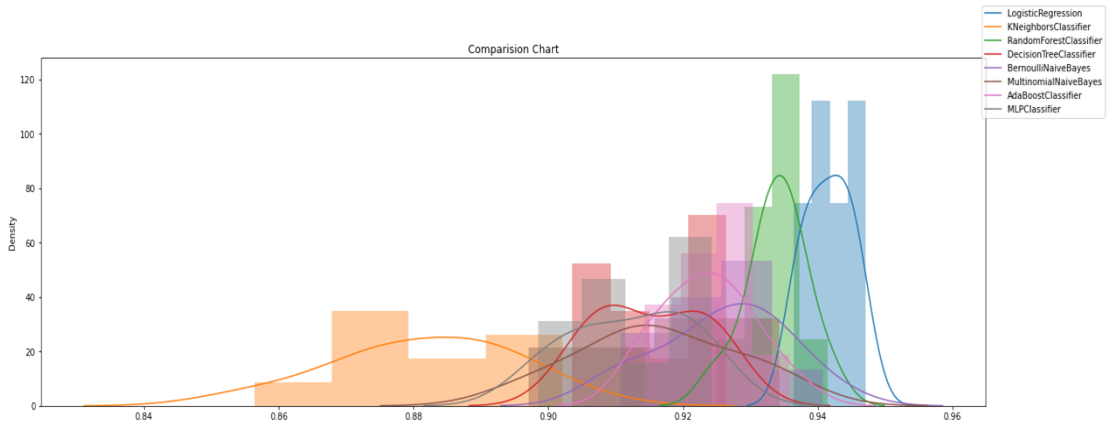
To implement the hybrid machine learning algorithm. We utilized a 10-fold cross-validation technique to split the dataset into train and test and then feed that output into

six different machine learning models. Each machine learning model will generate a set of the predicted dataset. Those datasets have been combined, and build a new dataset for another layer of our hybrid algorithm.

In the second layer, we have utilized a logistic regression algorithm in the second layer. The output of the first layer will be given to the Logistic regression algorithm in the second layer.

4.8.2 PERFORMANCE OF CONVENTIONAL ML MODEL

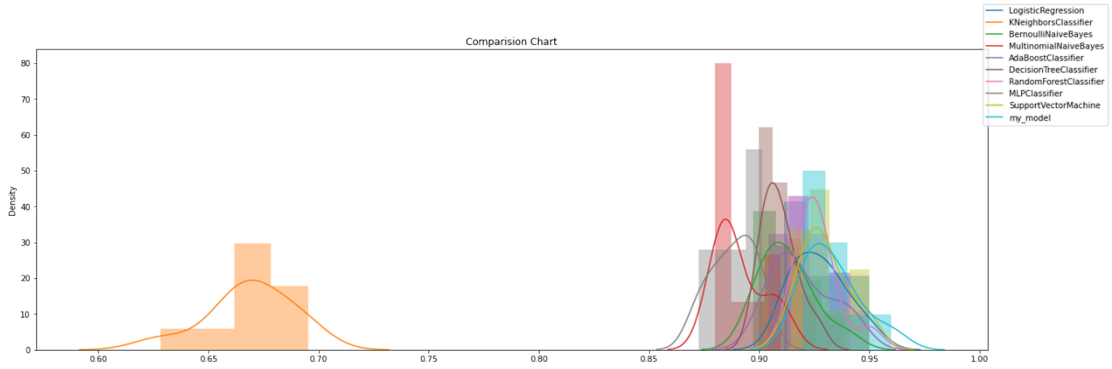
To evaluate the performance of the conventional machine learning model. We have trained nine different machine learning models, including Logistic Regression, KNeighbor classifier, Bernoulli Naïve Bayes, Multinomial Naïve Bayes, Ada Boost Classifier, Random Forest, MLP classifier, and Support Vector Machine on our pre-processed hate speech dataset with TFIDF vectorization techniques. A visualization of the result is shown in Figure 7, which compares the two-layer ML model to other ML models.



Graph 4.1 Performance plot of Conventional ML model

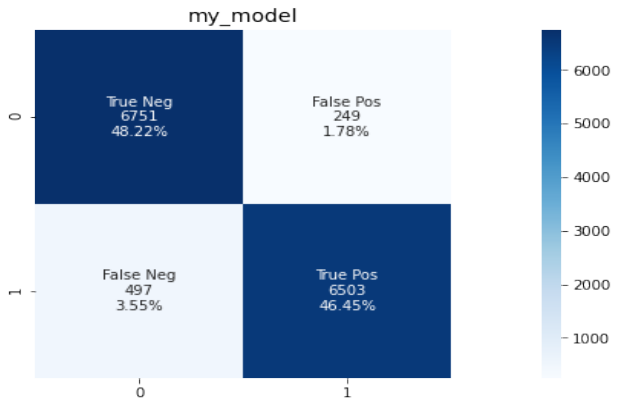
4.8.3 PERFORMANCE OF TWO-LAYER HYBRID ML MODEL

To evaluate the performance of the hybrid machine learning model, we used our pre-processed hate speech dataset with the TF-IDF vectorization technique to train the model. With regard to the other machine learning models, we observed that the Hybrid ML Model Approach worked better than the other nine Machine learning models.



Graph 4.2 Comparision Chart of Hybrid Model and Conventional Models

To determine the accuracy of our hybrid machine learning model, we depicted a confusion matrix on the final result, and we observed that the accuracy of the model is 94.67 %, and the F-score is 94.57%. The type 1 and type 2 error is very minimal 1.78% and 3.55% respectively. The confusion matrix of the hybrid model, as shown in the figure is showing our model is more accurate than other conventional machine learning models.



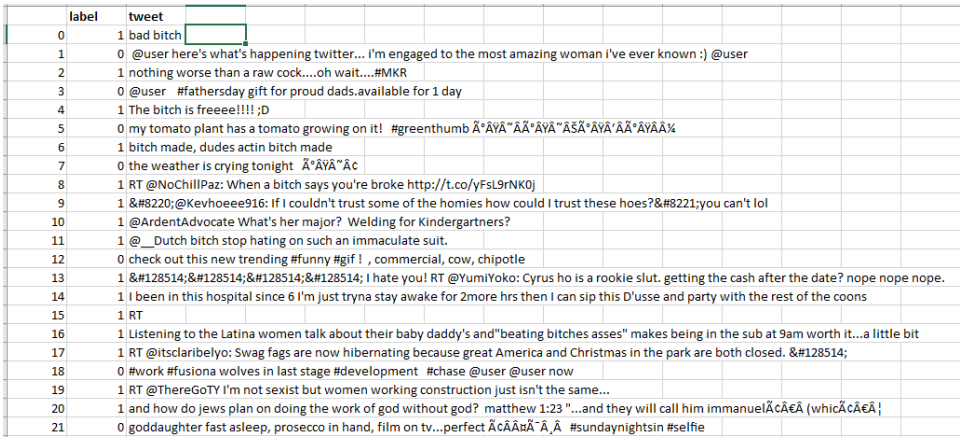
Graph 4.3 Confusion Matrix of Hybrid ML Model

CHAPTER - 5

RESULT ANALYSIS AND DISCUSSION

5.1 DATASET RESULT

We have collected the dataset from two sources Crowdflower dataset and the Hatespeechdata dataset. The gathered dataset included several redundant characteristics, which I eliminated, and the final dataset included just two attributes, as seen in the figure. Depending on the two attributes, we merged two datasets together. The total number of rows in the dataset we obtained after merging it is 55591 rows. There are two attributes in our dataset set: tweet and label. The tweet property has text values, while the label has numerical values 0 and 1. The numbers 0 and 1 indicate hate speech and normal speech, respectively.



	label	tweet
0	1	bad bitch
1	0	@user here's what's happening twitter... i'm engaged to the most amazing woman i've ever known :) @user
2	1	nothing worse than a raw cock....oh wait....#MKR
3	0	@user #fathersday gift for proud dads.available for 1 day
4	1	The bitch is freeee!!!! ;D
5	0	my tomato plant has a tomato growing on it! #greenthumb Ã°Ã¥Ã˜ÃÃ°Ã¥Ã˜ÃŠÃ°Ã¥Ã'ÃÃ°Ã¥ÃÃ¼
6	1	bitch made, dudes actin bitch made
7	0	the weather is crying tonight Ã°Ã¥Ã˜Âc
8	1	RT @NoChillPaz: When a bitch says you're broke http://t.co/yFsL9rNK0j
9	1	“@Kevhoeee916: If I couldn't trust some of the homies how could I trust these hoes?”you can't lol
10	1	@ArdentAdvocate What's her major? Welding for Kindergartners?
11	1	@__Dutch bitch stop hating on such an immaculate suit.
12	0	check out this new trending #funny #gif ! , commercial, cow, chipotle
13	1	😂😂😂😂 I hate you! RT @YumiYoko: Cyrus ho is a rookie slut. getting the cash after the date? nope nope nope.
14	1	I been in this hospital since 6 I'm just tryna stay awake for 2more hrs then I can sip this D'usse and party with the rest of the coons
15	1	RT
16	1	Listening to the Latina women talk about their baby daddy's and"beating bitches asses" makes being in the sub at 9am worth it...a little bit
17	1	RT @itsclaribelyo: Swag fags are now hibernating because great America and Christmas in the park are both closed. 😂
18	0	#work #fusiona wolves in last stage #development #chase @user @user now
19	1	RT @ThereGoTY I'm not sexist but women working construction just isn't the same...
20	1	and how do jews plan on doing the work of god without god? matthew 1:23 "...and they will call him immanuelÃcÂ€Â (whicÃcÂ€Â¦
21	0	goddaughter fast asleep, prosecco in hand, film on tv...perfect ÃcÂÂ¤Ã¯Â‚Â #sundaynightsin #selfie

Figure 5.1 Dataset in Excel

5.2 PRE-PROCESSING RESULT

After collecting the dataset, we have two attributes in our dataset: tweet and label. The label is a dependent value for our model, and the tweet is the independent attribute for our model. The independent attribute contained some unnecessary and redundant text such as hashtag, emoji, punctuation, null value, and other unnecessary words like stopwords which don't include any useful information. To remove unnecessary

information from the tweet attribute, we are using various steps such as lowercasing, punctuation removal, accented characters removal, stopwords removal, tokenization, and lemmatization. The result of the final dataset is shown in figure 5.2

label	tweet	tokenized_tweet	preproc_step_1	preproc_step_2	preproc_step_3	preproc_step_4	preproc_step_5	preproc_step_6	preproc_step_7
1	RT @OMGitsBabyJoe: Don't talk behind my back I...	[RT, OMGitsBabyJoe, Do, n't, talk, behind, my,...	[, omgitsbabyjoe, do, n't, talk, behind, my, b...	[omgitsbabyjoe, do, n't, talk, behind, my, bac...	[omgitsbabyjoe, talk, hate, pussy, shit]	[omgitsbabyjoe, talk, hate, pussy, shit]	[omgitsbabyjoe, talk, hate, pussy, shit]	[omgitsbabyjoe, talk, hate, pussy, shit]	[omgitsbabyjoe, talk, hate, pussy, shit]
0	first day of schooll #quinnhaley.	[first, day, of, school, quinnhaley]	[first, day, of, school, quinnhaley]	[first, day, of, school, quinnhaley]	[day, school, quinnhaley]	[day, school, quinnhaley]	[day, school, quinnhaley]	[day, school, quinnhaley]	[day, school, quinnhaley]
1	@user todays donkey #zionazis are proud of mu...	[user, todays, donkey, zionazis, are, proud, o...	[user, todays, donkey, zionazis, are, proud, o...	[user, todays, donkey, zionazis, are, proud, o...	[user, todays, donkey, zionazis, proud, murder...	[todays, donkey, zionazis, proud, murdering, r...	[todays, donkey, zionazis, proud, murdering, r...	[today, donkey, zionazis, proud, murder, real,...	[today, donkey, zionazis, proud, murder, real,...
0	today is the first sunny day of the year	[today, is, the, first, sunny, day, of, the, y...	[today, is, the, first, sunny, day, of, the, y...	[today, is, the, first, sunny, day, of, the, y...	[today, sunny, day, year]	[today, sunny, day, year]	[today, sunny, day, year]	[today, sunny, day, year]	[today, sunny, day, year]
0	damn i wish this orlando shooter shoot me&...	[damn, i, wish, this, orlando, shooter, shoot,...	[damn, i, wish, this, orlando, shooter, shoot,...	[damn, i, wish, this, orlando, shooter, shoot,...	[damn, wish, orlando, shooter, shoot, amp, kil...	[damn, wish, orlando, shooter, shoot, amp, kil...	[damn, wish, orlando, shooter, shoot, amp, kil...	[damn, wish, orlando, shooter, shoot, amp, kil...	[damn, wish, orlando, shooter, shoot, amp, kil...

Figure 5.2 Output Result after pre-processing

5.3 FEATURE EXTRACTION RESULTS

5.3.1 BAG OF WORDS RESULT

In this technique, we extracted the important features from the tweet attribute and calculated the frequency of the particular word in the document. Based on the frequency, we decide which attribute is more important in the document. Here in the figure is showing the data frame of the bag of word output. Where single words show the features and each row represents the document of the corpus. The sparse matrix created by computing the frequency of the word appears in the document, as shown in figure 5.3.

```
------------------------------------------------
Feature Length:  10000
------------------------------------------------
---------------Bag of Word Matrix-----------------
   aaa  ability  able  able sleep  abo  abortion  abraham  abraham lincoln  \
0    0        0     0           0    0         0        0                0
1    0        0     0           0    0         0        0                0
2    0        0     0           0    0         0        0                0
3    0        0     0           0    0         0        0                0
4    0        0     0           0    0         0        0                0

   abrupt  absolute  ...  zero  zero bitch  zitlalyl  zitlalyl vic  zoe  zone  \
0       0         0  ...     0           0         0             0    0     0
1       0         0  ...     0           0         0             0    0     0
2       0         0  ...     0           0         0             0    0     0
3       0         0  ...     0           0         0             0    0     0
4       0         0  ...     0           0         0             0    0     0

   zoo  zuma  zurich  zzachbarness
0    0     0       0             0
1    0     0       0             0
2    0     0       0             0
3    0     0       0             0
4    0     0       0             0

[5 rows x 10000 columns]
------------------------------------------------
------------------------------------------------
Data / BOW fit-transform Matrix Shape:  (14000, 10000)
------------------------------------------------
```

Figure 5.3 Sparse Matrix after Countvectorizer Transformation

5.3.2 TFIDF RESULT

TFIDF calculates the weight of the term/word based on how much importance that word is in the whole corpus. The result of the TFIDF on the hate speech dataset is shown in figure 5.4

```
-----------------------------------------------
Feature Length:  3000
-----------------------------------------------
---------------TFIDF Matrix-----------------
   ability  able  abo  absolutely  abt  abuse  accept  accessory  accord  \
0      0.0   0.0  0.0         0.0  0.0    0.0     0.0        0.0     0.0
1      0.0   0.0  0.0         0.0  0.0    0.0     0.0        0.0     0.0
2      0.0   0.0  0.0         0.0  0.0    0.0     0.0        0.0     0.0
3      0.0   0.0  0.0         0.0  0.0    0.0     0.0        0.0     0.0
4      0.0   0.0  0.0         0.0  0.0    0.0     0.0        0.0     0.0

   account  ...  youth  youtube  youtuber  yrs  yum  yummy  yung  zen  zero  \
0      0.0  ...    0.0      0.0       0.0  0.0  0.0    0.0   0.0  0.0   0.0
1      0.0  ...    0.0      0.0       0.0  0.0  0.0    0.0   0.0  0.0   0.0
2      0.0  ...    0.0      0.0       0.0  0.0  0.0    0.0   0.0  0.0   0.0
3      0.0  ...    0.0      0.0       0.0  0.0  0.0    0.0   0.0  0.0   0.0
4      0.0  ...    0.0      0.0       0.0  0.0  0.0    0.0   0.0  0.0   0.0

   zone
0   0.0
1   0.0
2   0.0
3   0.0
4   0.0

[5 rows x 3000 columns]
-----------------------------------------------
Data / TFIDF fit-transform Matrix Shape:  (14000, 3000)
-----------------------------------------------
```

Figure 5.4 Sparse Matrix after TFIDF Transformation

5.4 MODEL RESULT

After the implementation of the hybrid model, we compared our proposed hybrid model with the existing conventional machine learning model. The figure is the comparison plot which is showing various models' performance. Each model is represented in different colors, and the blue color is representing our hybrid model. The deviation of the blue line in the figure is leading compare to other models, which is showing the performance of our model is better than other conventional models.

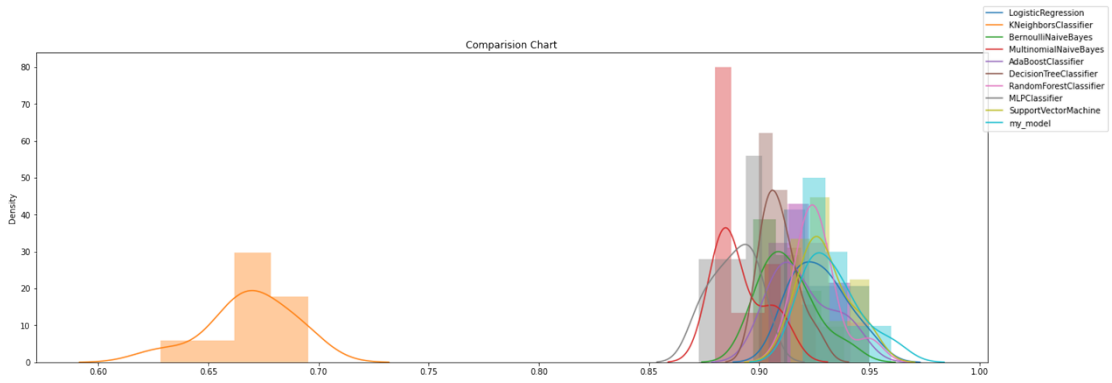


Figure 5.5 Comparison plot

We have also depicted the result in a line plot to make the visualization more clear. The figure is a line plot of the various machine learning model with a hybrid model. The blue line is showing the performance of our hybrid machine learning model implemented on the hate speech dataset.

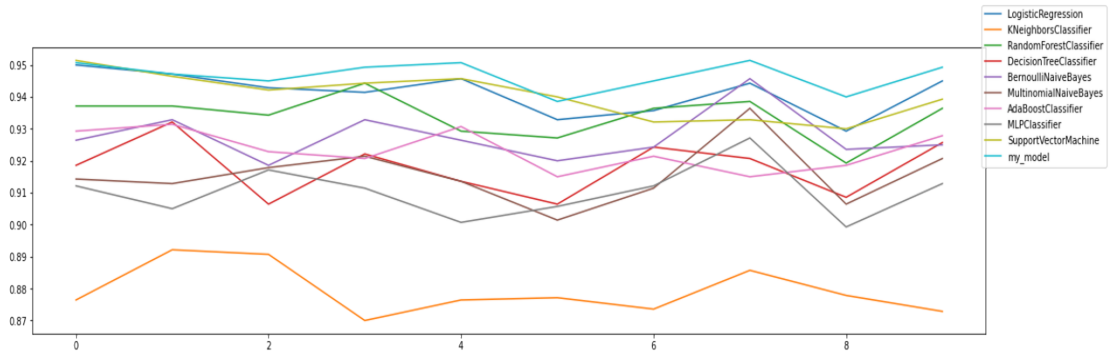


Figure 5.6 Comparison line plot

The box plot in the figure represents the average of all 10-fold cross-validation approaches for each model. The last box in the box plot is very dense and at the top.

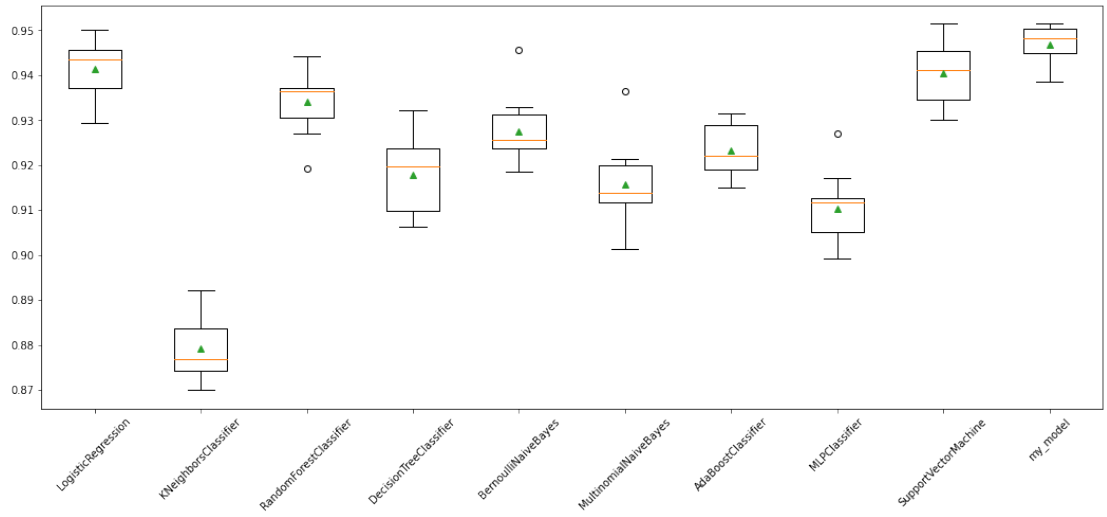


Figure 5.7 Comparison Boxplot

5.5 COMPARATIVE ANALYSIS

I have implemented our pre-processed dataset into all possible machine learning algorithms and compared the accuracy, precision, recall, and f-score with our hybrid model. We have found that the accuracy of our model is 94.78% which is far better than other conventional machine learning algorithms. Table 5.1 compares the accuracy of different machine learning models, as well as the accuracy of our hybrid machine learning models.

Table 5.1. Comparison between conventional and hybrid model results

Sr. No.	Model Name	Precision (%)	Recall (%)	Accuracy (%)	F1_score (%)
1	Logistic Regression	97.17	90.99	94.16	93.97
2	Random Forest Classifier	94.02	92.67	93.39	93.34
3	Bernoulli Naïve Bayes	91.64	93.74	92.59	92.68
4	Multinomial Naïve Bayes	89.22	94.70	91.62	91.87
5	AdaBoost Classifier	97.71	86.74	92.36	91.90
6	Support Vector Machine	97.80	90.37	94.17	93.94
7	Decision Tree Classifier	90.36	92.96	91.52	91.64
8	MLP Classifier	90.86	91.69	91.23	91.27
9	K-Neighbors Classifier	95.56	80.01	88.15	87.10

10	My Hybrid Model	96.31	93.00	94.78	94.59

5.6 DISCUSSION

According to S. Ahammed et al. [21], they implemented hate speech detection using Naïve Bayes and Support Vector Machine using the TFIDF vectorization approach. With 1339 rows of the dataset, they achieved 70% accuracy with the Support vector machine and 72% accuracy with Naïve Bayes. Whereas according to N. Rai et al. [22], employed a Random forest machine learning algorithm using the bag of words approach and achieved 83% of accuracy with 24782 rows of the dataset. In the same way, there are various other researches where single machine learning algorithms were utilized, but the accuracy of the model declined because of the fewer number datasets Table 5.2.

Table 5.2. Previous researches on hate speech detection

Published Year	Methodology	Dataset Rows	Result	Reference
[IEEE] 2020	Web scraping, count vectorizer and TF-IDF, Naive Bayes and SVM	1339	SVM (Accuracy =70%) Naive Bayes (Accuracy=72%)	[21]
[IEEE] 2020	Random forest with Bag of words	24782	Accuracy= 83%	[22]
[IEEE] 2020	Word2vec and Convolutional Neural Network	13029	Accuracy = 80.15%	[23]

Throughout this research, it has been proven that a Hybrid machine learning model can perform better with a limited number of datasets and gain more than 90 percent accuracy.

CHAPTER - 6

CONCLUSION

AND

FUTURE WORK

6.1 CONCLUSION

In this work, we have proposed a hybrid machine learning model using six different machine learning algorithms in two layers. Initially, we collected the dataset from two sources and combined it into one dataset. Our final dataset contains only two features: label and tweet. The tweet feature had some unnecessary special characters and redundant data. We applied pre-processing steps to clean the data and balanced the label attribute to improve the precision and recall. We employed the TFIDF and Bag words approach to transform text data into a sparse matrix since machine learning does not work with text data. Using TFIDF with the hybrid model, we gained 94.68% accuracy, while in the bag of words with a 2-gram approach, our model achieved 94.62% of accuracy. Other researches on hate speech detection are based on the conventional machine learning model, which gives accuracy around 70% to 90%. The proposed work is based on a hybrid machine learning algorithm and achieved more than 90% accuracy.

6.2 LIMITATIONS

The proposed work is based on a hybrid model. In this work, we are using 14000 rows of the dataset, which requires a minimum of 4 GB RAM to process the dataset. As we are utilizing six different machine learning models to make a hybrid model hence, it requires plenty of memory to process the data simultaneously.

6.3 FUTURE WORK

Our main objective was to detect hate speech on text datasets, but this work can be extended to images, videos, and audio. We can make a hybrid model with the combination of neural network and machine learning to efficiently analyze the image, video, and audio data.

REFERENCES

[1] "This is What Happens in an Internet Minute [Infographic] | Social Media Today." https://www.socialmediatoday.com/news/this-is-what-happens-in-an-internet-minute-infographic/524426/ (accessed Jun. 21, 2021).

[2] United Nations, "United Nations Strategy and Plan of Action on Hate Speech," *United Nations Rep.*, no. May, pp. 1–5, 2019.

[3] N. Sambuli, F. Morara, and C. Mahihu, "Umati: Monitoring Online Dangerous Speech," no. March, p. 55, 2013, [Online]. Available: http://www.ihub.co.ke/blog/wp-content/uploads/2014/06/2013-report-1.pdf.

[4] Y. Bhargava, "8 out of 10 Indians have faced online harassment - The Hindu," *The Hindu*, Oct. 04, 2017. https://www.thehindu.com/news/national/8-out-of-10-indians-have-faced-online-harassment/article19798215.ece (accessed Feb. 18, 2021).

[5] N. R. Fatahillah, P. Suryati, and C. Haryawan, "Implementation of Naive Bayes classifier algorithm on social media (Twitter) to the teaching of Indonesian hate speech," *Proc. - 2017 Int. Conf. Sustain. Inf. Eng. Technol. SIET 2017*, vol. 2018-Janua, pp. 128–131, 2018, doi: 10.1109/SIET.2017.8304122.

[6] M. A. Fauzi and A. Yuniarti, "Ensemble Method for Indonesian Twitter Hate Speech Detection," no. July, pp. 294–299, 2018, doi: 10.11591/ijeecs.v11.i1.pp294-299.

[7] P. Sari and B. Ginting, "Hate Speech Detection on Twitter Using Multinomial Logistic Regression Classification Method," pp. 105–111, 2019.

[8] O. Oriola and E. Kotze, “Evaluating Machine Learning Techniques for Detecting Offensive and Hate Speech in South African Tweets,” *IEEE Access*, vol. 8, pp. 21496–21509, 2020, doi: 10.1109/ACCESS.2020.2968173.

[9] A. Briliani, B. Irawan, and C. Setianingsih, “Hate speech detection in indonesian language on instagram comment section using K-nearest neighbor classification method,” *Proc. - 2019 IEEE Int. Conf. Internet Things Intell. Syst. IoTaIS 2019*, pp. 98–104, 2019, doi: 10.1109/IoTaIS47347.2019.8980398.

[10] R. Zhao and K. Mao, “Cyberbullying Detection based on Semantic-Enhanced Marginalized Denoising,” vol. 3045, no. c, pp. 1–12, 2016, doi: 10.1109/TAFFC.2016.2531682.

[11] A. Rodriguez, C. Argueta, and Y. L. Chen, “Automatic Detection of Hate Speech on Facebook Using Sentiment and Emotion Analysis,” *1st Int. Conf. Artif. Intell. Inf. Commun. ICAIIC 2019*, pp. 169–174, 2019, doi: 10.1109/ICAIIC.2019.8669073.

[12] S. Jaki and T. De Smedt, “Right-wing German Hate Speech on Twitter : Analysis and Automatic Detection,” no. May, pp. 1–31, 2018.

[13] M. Di Capua, E. Di Nardo, and A. Petrosino, “Unsupervised Cyber Bullying Detection in Social Networks,” pp. 432–437, 2016.

[14] C. J. Hutto and E. Gilbert, “VADER : A Parsimonious Rule-based Model for Sentiment Analysis of Social Media Text,” pp. 216–225.

[15] N. D. Gitari, Z. Zuping, H. Damien, and J. Long, “A Lexicon-based Approach for Hate Speech Detection,” vol. 10, no. 4, pp. 215–230, 2015.

[16] H. Rosa, D. Matos, L. Coheur, and P. Carvalho, “A ‘ Deeper ’ Look at Detecting Cyberbullying in Social Networks,” *2018 Int. Jt. Conf. Neural Networks*, pp. 1–8, 2018, doi: 10.1109/IJCNN.2018.8489211.

[17] T. Van Huynh, D. Nguyen, K. Van Nguyen, N. L. Nguyen, and A. G. Nguyen, “Hate Speech Detection on Vietnamese Social Media Text using the Bi-GRU-LSTM-CNN Model,” 2019.

[18] B. Gambäck and U. K. Sikdar, “Using Convolutional Neural Networks to Classify Hate-Speech,” no. 7491, pp. 85–90, 2017.

[19] E. W. Pamungkas, V. Patti, and D. Informatica, “Cross-domain and Cross-lingual Abusive Language Detection : a Hybrid Approach with Deep Learning and a Multilingual Lexicon,” pp. 363–370, 2019.

[20] S. Alsafari, S. Sadaoui, and M. Mouhoub, “Hate and offensive speech detection on Arabic social media,” *Online Soc. Networks Media*, vol. 19, no. September, p. 100096, 2020, doi: 10.1016/j.osnem.2020.100096.

[21] S. Ahammed, M. Rahman, M. H. Niloy, and S. M. M. H. Chowdhury, “Implementation of Machine Learning to Detect Hate Speech in Bangla Language,” *Proc. 2019 8th Int. Conf. Syst. Model. Adv. Res. Trends, SMART 2019*, pp. 317–320, 2020, doi: 10.1109/SMART46866.2019.9117214.

[22] N. Rai, P. Meena, and C. Agrawal, “Improving the hate speech analysis through dimensionality reduction approach,” in *2020 6th International Conference on Advanced Computing and Communication Systems (ICACCS)*, 2020, pp. 321–325, doi: 10.1109/ICACCS48705.2020.9074240.

[23] A. Chaudhari, A. Parseja, and A. Patyal, “CNN based Hate-o-Meter: A Hate Speech Detecting Tool,” in *2020 Third International Conference on Smart Systems and Inventive Technology (ICSSIT)*, 2020, pp. 940–944, doi: 10.1109/ICSSIT48917.2020.9214247.

PLAGARISHM CHECK REPORT

Plagiarism Checker X Originality Report

Plagiarism Quantity: 4% Duplicate

Date	Wednesday, August 04, 2021
Words	343 Plagiarized Words / Total 8629 Words
Sources	More than 69 Sources Identified.
Remarks	Low Plagiarism Detected - Your Document needs Optional Improvement.

SOCIAL MEDIAIn recent years, social networking has become a popular way for people to connect with one another on a regular basis.. Facebook, Instagram, and Twitter are examples of social media platforms that allow you to interact with family and friends, as well as individuals who have similar interests to your own.If anyhow you're involved in social networking, it implies you're making use of social media sites, also known as social networks, to interact with other people and share information. Twitter, Facebook, Reddit, YouTube, Telegram, TikTok, Instagram, LinkedIn, Pinterest, and Snapchat are just a few of the most popular social networking platforms available today.Even though different social media sites are suitable for different users, Facebook serves as an excellent example of a broad social network.

When you sign up for Facebook, you may come across some other individuals who are already members of the site, and you can add them as casual or close friends. As you get more familiar with the site, you may be able to add friends who share your type of interests or find individuals you already know and invite them to join your group. Other individuals may come across your profile on Facebook and attempt to establish a connection with you.Your connections and hobbies increase the more you use social networking sites like Facebook. It's comparable to networking in real life, such as at a business conference or a social gathering. The more you connect with other people and find that you have similar friends and hobbies, the larger your circle of friends grows.Each social networking site and the app has its own set of features and points of view, although the majority share certain characteristics. The following terminology will be encountered whether you're new to Facebook, Twitter, or another social media platform.E.g.,

Public Profile, Followers And Friends, Shares, Comments, And Likes, Groups, Tagging And Hashtags.It has been a concern of humans for centuries to interact with friends and family over long distances. People have traditionally relied on communication to strengthen their relationships as social animals. When face-to-face communication is impossible or inconvenient, people have devised a variety of inventive methods. HISTORY OF SOCIAL MEDIAThe oldest ways of communication across long distances relied on written writing handed by hand from one person to another. To put it another way, letters. Since the year 550 BC, there has been a continuous evolution of postal service, with the most basic delivery method eventually becoming more ubiquitous and streamlined over the course of several centuries.The invention of the telegraph occurred in 1792.

As a result, messages could be sent across great distances far more quickly than a horse and rider could carry them. Despite the fact that we can send short messages through the telegraph, they were a revolutionary method of sending messages.The pneumatic post, which was invented in 1865, provided

Sources found:

Click on the highlighted sentenc

Internet Pages

<1% https://startupmindse
<1% https://www.fannit.cc
<1% https://adespresso.c
<1% https://www.researcl
<1% https://www.academ
<1% https://rampages.us/
<1% https://dokumen.pub
<1% https://www.researcl
<1% https://www.researcl
<1% https://www.canada.
<1% https://www.publicsa
<1% https://www.cambrid
<1% https://hcis-journal.s
<1% https://www.researcl
<1% https://vas3k.com/bl
<1% https://www.researcl
<1% https://towardsdatas
<1% https://www.researcl
<1% https://www.mygreat
<1% https://pediaa.com/d
<1% https://www.sas.upe
<1% https://online.stat.ps
<1% https://www.geeksfo
<1% https://machinelearn
<1% https://www.quora.c
<1% https://www.science
<1% http://www.ijcstjourn
<1% https://www.uni-man
<1% https://link.springer.c
<1% https://www.datanov

Turnitin Plagiarism Report

M.tech Dissertation

turnitin

ORIGINALITY REPORT

10%	5%	7%	3%
SIMILARITY INDEX	INTERNET SOURCES	PUBLICATIONS	STUDENT PAPERS

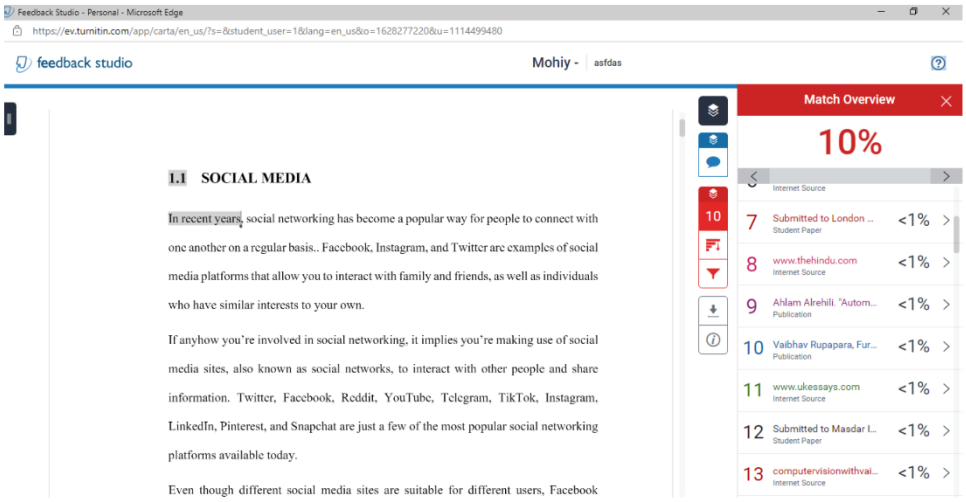
MATCH ALL SOURCES (ONLY SELECTED SOURCE PRINTED)

3%

★ "International Conference on Innovative Computing and Communications", Springer Science and Business Media LLC, 2021

Publication

Exclude quotes	Off	Exclude matches	Off
Exclude bibliography	On		



PUBLICATION FROM THIS WORK

1. **"Automatic Hate Speech Detection: A Literature Review"** has been published in the International Journal of Engineering and Management Research
2. **"Improved Hate speech Detection System using Multi-Layers Hybrid Machine Learning Model"** has been accepted in International Conference on Computer Vision and Robotics (CVR 2021) and will be published in Springer Book Series 'Algorithms for Intelligent Systems'.

PUBLICATIONS

International Journal of Engineering and Management Research
www.ijemr.net
e-ISSN: 2250-0758 | p-ISSN: 2394-6962
Volume-11, Issue-2 (April 2021)
https://doi.org/10.31033/ijemr.11.2.17

Automatic Hate Speech Detection: A Literature Review

Mohiyaddeen[1] and Dr. Sifatullah Siddiqi[2]
[1]Student, Department of Computer Science, Integral University, INDIA
[2]Professor, Department of Computer Science, Integral University, INDIA

[1]Corresponding Author: moinsheik@student.iul.ac.in

ABSTRACT

Hate speech has been an ongoing problem on the Internet for many years. Besides, social media, especially Facebook, and Twitter have given it a global stage where those hate speeches can spread far more rapidly. Every social media platform needs to implement an effective hate speech detection system to remove offensive content in real-time. There are various approaches to identify hate speech, such as Rule-Based, Machine Learning based, deep learning based and Hybrid approach. Since this is a review paper, we explained the valuable works of various authors who have invested their valuable time in studying to identifying hate speech using various approaches.

Keywords— Classification Algorithm, Machine Learning, Hate Speech, Deep Learning, Supervised Learning

I. INTRODUCTION

Social networking sites are the most efficient way to meet new people. However, as social networking sites have grown in popularity, people have discovered an illegal and immoral way to use them. The most commonly encountered and most dangerous misuses of online social media are the expression of hate and harassment. Hate speech may be characterized as violence, hate, intimidation, racism, threats, harassment, insults, provocation, or sexism. These are some of the biggest threats to a social media site online. Several studies have already been worked into the identification of hateful messages in social media platforms[1], along with the dissemination of hateful messages on the dark web[2]. Certain studies have implemented the domain of detection of hate speech but are primarily focused on supervised learning approaches[3]–[5].instruction set. The electronic file of your paper will be formatted further at IJEMR. Define all symbols used in the abstract.

1.1 Hate Speech on Social Media

Hate speech is a form of writing that disparages and is likely to cause damage or danger to the victim on social media. It is a partial, aggressive and malicious speech that targets an individual or a group of people because of their conscious or unconscious intrinsic characteristics[6]. It is a type of speech that shows a strong intent to cause harm, provoke violence, or encourage hate. The social media environment and collaborative worldwide web offer a conducive environment for hate messages against an alleged enemy group to be created, shared, and exchanged.

In 2013, N. Sambuli et al. worked on a project called "Umati: Monitoring Online Dangerous Speech." The project was based on monitoring Hatebase and dangerous speech[7]. According to them, dangerous expressions can be observed in the following ways:

a) It is targeted to a group of people and not a single person. Dangerous speech is an offensive speech that encourages the audience to participate in acts of violence against a particular group of people, therefore In the internet domain, the most prevalent forms of hate speech are related to religion, race, sexual orientation, nationality, class, and gender.

b) Hate Speech may contain one of the pillars of dangerous speech, for instance, statements that classify people as vermin, which claims that a group of people is like rodents or insects.

c) Dangerous speech often incites the listener to support or commit acts of violence against the specific group. The six most common calls to action in dangerous speech are: kill, riot, beat, loot, forcefully evict, and discrimination.

The Internet is inherently open and dynamic, but various communities have their own rules to define the limits of speech. These boundaries differ from one culture to the next and are shaped by historical events and cultural norms[6].

The manual method of detecting and eliminating hate speech posts or comments is time-consuming and computationally expensive. Because of these issues and the prevalence of hateful content on social media, there is a strong case for automated hate speech identification.

Since hate speech, abusive language, and offensive language have recently become subjects of general concern, detecting hate speech has grown to be a major topic by the community of natural language processing (NLP), as demonstrated by the creation of datasets in a variety of languages[8]–[11].

The implementation of systems for automatically detecting abusive and offensive language has followed a general pattern in NLP. Feature-based linear classifiers[8],

e-ISSN: 2250-0758 | p-ISSN: 2394-6962
Volume-11, Issue-2 (April 2021)
https://doi.org/10.31033/ijemr.11.2.17

[12], fine-tuning pre-trained language models[13], [14] , and neural network architectures [15]–[17].

There are many approaches by which hate speech detection can be carried out, such as Machine learning, Deep learning, and the Rule-based approach.

II. APPROACHES FOR HATE SPEECH DETECTION

1. Rule-Based Linguistic Approaches

In the Linguistic rule-based approach, Hate speech detection uses a linguistic engine that understands the grammar, morphology, and semantics of a specific language. Furthermore, the program adds rules that check for unique core semantic terms in the sentence in order to determine their potential meanings. For instance, if we input the keyword "bad." The linguistic engine will automatically search for the terms "terrible/awful/unsatisfactory" as well.

2. Machine Learning Approaches

Machine learning creates a mathematical model based on training data to make predictions or decisions without being explicitly programmed. The aim of Machine learning is to make a classifier or regression model through learning the training data set and then use test data set to evaluate the performance of the classifier or regression model. Machine learning can be classified into the following categories based on the nature of the training data. e.g. Supervised learning, Unsupervised learning, Semi-supervised learning.

3. Deep Learning Approaches

The deep learning approach uses neural networks to solve complex problems in an innovative way. When you feed a neural network a series of examples, such as pictures of humans, It can recognize the features that are shared by those pictures. When we use layers of neural network side by side, these layers recognize every detail of the picture to create an effective model. After sufficient training, a neural network becomes refined and capable of classifying unlabeled pictures.

4. Hybrid Approaches

Each solution has its own collection of limitations. And it seems a good solution to merge either two or more approaches into the hybrid approach where one complements another. In the Hybrid approach, we generally combined machine learning, rule-based and deep learning approaches to make an effective model.

III. RELATED WORK

A. Linguistic Rule-Based Approach

In 2014, C. J. Hutto et al. proposed an approach to classify sentiment using VADER, which is a rule-based approach [18]. At first, they created a list of lexical features that are highly sensitive to the sentiment of social media posts. After then they combined that list of lexical features with five general rules that encapsulate syntactical and grammatical rules for presenting sentiment intensity. At last, they have found that VADER performed 96% accuracy using the rule-based model on Twitter sentiments.

Dennis Gitariet al. in 2015 proposed a method to identify the Sentiment Analysis of the Social Media Text using the Rule-based method [19]. In this work, They categorized the hate speech problem into three fields religion, nationality, and race. The main objective of this paper is to develop a classification model that employs sentiment analysis. The developed model not only detects subjective sentences but also classifies and ranks the polarity of sentiment phrases. After then they relate the semantic and subjective features with hate speech. Finally, they achieved 71.55 % precision using the lexicon-based approach.

B. Supervised Learning Approach

Fatahillah et al. (2017) used Naive Bayes Classifier Algorithm to detect hate speech on Instagram using the k-nearest neighbor classifier [20]. They collected the data set using Twitter API from Twitter and annotated those data set manually. After preprocessing and feature engineering phase, they applied the Naive Bayes Classifier algorithm and found 93% of accuracy.

M. Ali Fauzi et al. (2018) proposed an approach to identify hate speech using a set of supervised learning algorithms [21]. They ensembled five different classification algorithms, including K-Nearest Neighbours, Random Forest, Naive Bayes, Support Vector Machine, and Maximum Entropy. They collected the data set using Twitter API and annotated those data set manually. In preprocessing phase, They employed tokenization, filtering, stemming, and term weighting methods. They utilized the bag of words features with TFIDF techniques. The naive Bayes algorithm performed best with 78.3 % of accuracy among all the other five stand-alone classifiers.

In 2019, P. Sari et al. proposed an approach to detect hate speech using logistic regression on Twitter. [22] They collected the data from Twitter and employed Case Folding, Tokenizing, Filtering, and Stemming methods in preprocessing phase. After Pre-processing, the TF-IDF technique is used for vectorization. After Feature engineering, the Logistic regression algorithm has been applied, and they have found 84% of accuracy.

In 2020, Oluwafemi Oriola et al. proposed an approach to detect offensive speech on tweeter [5]. The author collected the data set using Twitter API and annotated those data set into two sections, free speech 'FS' and hate speech 'HT.' In preprocessing phase, they removed special characters, emojis, punctuations, symbols, hashtags, stopwords to clean the data. In the feature

e-ISSN: 2250-0758 | p-ISSN: 2394-6962
Volume-11, Issue-2 (April 2021)
https://doi.org/10.31033/ijemr.11.2.17

engineering phase, they employed the TF-IDF technique to transform the text into feature vectors. After applying an optimized support vector machine with n-gram, they have found 89.4% of accuracy.

In 2020, Annisa Briliani et al. proposed an approach to identify hate speech on Instagram using the k-nearest neighbor classifier [23]. They collected the data set using Instagram API from Instagram and annotated those data set manually. They divided the dataset into 2 labels, namely zero and one. In preprocessing phase, they cleaned the data and employed the TF-IDF technique in the feature engineering phase. After then, they applied the k-nearest neighbor algorithm and found 98.13% of accuracy.

C. Unsupervised Learning Approach

Rui Zhao et al. (2015) proposed an approach to detect cyberbullying using Semantic-Enhanced Marginalized Denoising Auto-Encoder [24]. They used two sources of data set. The first source is Twitter, and the second source is Myspace. Twitter data was collected through Twitter stream API, and Myspace data was collected using the web crawling technique. They have achieved 84.9 % accuracy using smSDA for the Twitter dataset, and they have got 89.7% of accuracy with smSDA with the MySpace dataset.

Axel Rodríguez et al. (2019) proposed an approach to detect hate speech content using sentiment analysis on Facebook [25]. They used Graph API to extract the post and comments from Facebook. To remove the unrelated texts VADER and JAMMIN were used. In preprocessing phase, they filtered out all unnecessary stopwords or symbols. Preprocessed documents converted into the vector using TFIDF. The resulting matrix is passed to the k-means clustering algorithm as an input matrix. The most negative articles and responses were collected using sentiment and emotion analysis.

Sylvia Jaki et al. (2019) demonstrated an approach to detect hate speech content using unsupervised learning on Twitter [26]. They collected over 50,00 data set using Twitter API. They used NLP techniques to group the words into similar clusters. They computed three clusters of the top 250 most biassed terms using spherical k-means clustering and skip-grams. As a result, they have got an 84.21% F1 score.

Michele Di Capua et al. (2019) proposed an approach to detect cyberbullying using unsupervised learning [27]. They collected over 54,000 data set from YouTube and Annotated all data sets manually. The GHSOM network algorithm was implemented using the SOM-Toolbox-2 platform. They trained and tested GHSOM using a K-fold method with K = 10. As a result, they have got 64% of accuracy.

D. Deep Learning Approaches

Hugo Rosa et al. (2018) proposed an approach to detect cyberbullying using deep learning [28]. In this paper, the training and testing data set was collected from Kaggle. At first, they initiated CNN, which holds a certain similarity to the issue of cyberbullying. It starts with a single-layer CNN and continues with a completely linked layer with a dropout of 0.5 and softmax performance. Then they combined CNN-DNN-LSTM to achieve maximum accuracy. They employed TFIDF for vector representation. They achieved 64.9% precision with google embeddings.

Tin Van Huynh et al. (2019) proposed an approach to detect hate speech using Bi-GRU-CNN-LSTM Model [29]. In this paper, they collected data from Twitter and categorized their data into three labels (OFFENSIVE, HATE, and CLEAN). After cleaning the data, they implemented three neural network models such as Bi-GRU-LSTM-CNN, Bi-GRU-CNN, and TextCNN to identify hate speech. They achieved a 70.57% of F1 score as a result.

Gambäck et al. (2019) utilized a deep learning algorithm to detect hate speech on Twitter [30]. In this paper, they collected data from Twitter and divided the data set into four categories(sexism, racism, combined(sexism and racism), and non-hate-speech). They employed four CNN models that were trained with character n-gram, word2vec, random vectors combined(word2vec and character n-gram). The author utilized a 10-fold technique to improve the accuracy of the model. Among all four models, word2vec based CNN model performed well with a 78.3% of F-score.

E. Hybrid based Approach

Viviana Patti et al. (2019) proposed a Hybrid based approach to detect hate speech [31]. In this paper, they employed two models. In their first model, they implemented a linear support vector classifier (LSVC), and in the second model, they employed a long short-term memory (LSTM) neural model with word embedding. They concatenated 17 categories, such as HurtLex, with two types, namely LSVC and LSTM. Joint learning with a multilingual word embedding model, including HurtLex, performed best with 68.7% of F1-score.

Safa Alsafari et al. (2020) proposed a Hate speech detection model for Arabic social media [32]. In this paper, they collected the data set using Twitter search API, and the data set is categorized into four classes (Religious, Nationality, Gender, and Ethnicity). They cleaned the data set in preprocessing phase by removing unnecessary words such as URLs, punctuations, symbols, tags, and stopwords. They implemented a three-class classification with CNN and Bert to achieve 75.51% of the F1-score.frequent validation or on demand validation - both can generate considerable, often unnecessary, network traffic and the latter reduces much of the latency gains offered by caching. The viable alternative in such circumstances is resource-driven invalidation where the server invokes a callback on the cache to inform it whenever an update has

e-ISSN: 2250-0758 | p-ISSN: 2394-6962
Volume-11, Issue-2 (April 2021)
https://doi.org/10.31033/ijemr.11.2.17

occurred [7][8]. Although this solution involves the server maintaining knowledge of its caches there will be applications which are willing to accept these memory costs in preference to the communication costs of polling-based invalidation.

Various works have already been done in this field. We have categorized all previous works into 5 sections such as Linguistic Rule-Based, unsupervised learning, supervised learning, deep learning, and hybrid approaches. We have also pointed out algorithms and features used in respective research works (Table 1-5).

IV. COMPARATIVE ANALYSIS

Table 1: Supervised Learning Approach (Comparison Analysis)

Paper	Year	Platform	Features and Algorithm	Precision (%)	Recall (%)	Accuracy (%)	F1-score (%)
[20]	2017	Twitter	TF-IDF, Naive Bayes	-	-	93.0	-
[21]	2018	Twitter	TF-IDF, Essembled method	-	-	83.4	79.8
[20]	2019	Twitter	TF-IDF, Multinomial Logistic Regression	80.02	82.0	87.68	-
[5]	2020	Twitter	n-gram, Optimized Gradient Boosting	-	-	80.3	-
[23]	2020	Instagram	TF-IDF , K-Nearest Neighbor	94.0	93.0	97.19	93.0

Table 2: Unsupervised Learning Approach (Comparison Analysis)

Author	Year	Platform	Features and Algorithm	Precision (%)	Recall (%)	Accuracy (%)	F1-score (%)
[27]	2015	MySpace, Twitter	Bag-of-words (BoW), Latent Semantic Analysis (LSA), smSDA	-	-	87.70	77.60
[24]	2019	Facebook	VADER and JAMMIN, TF-IDF, k-means	-	-	74.42	-
[25]	2019	Twitter	n-gram and k-means	84.21	83.97	-	84.21
[26]	2019	Twitter, YouTube, Formspring	GHSOM network algorithm, SOM-Toolbox-2	60.0	94.0	69.0	74.0

e-ISSN: 2250-0758 | p-ISSN: 2394-6962
Volume-11, Issue-2 (April 2021)
https://doi.org/10.31033/ijemr.11.2.17

Table 3: Linguistic Rule-Based Approach (Comparison Analysis)

Author	Year	Platform	Features and Algorithm	Precision (%)	Recall (%)	Accuracy (%)	F1-score (%)
[19]	2014	Micro blogging sites	SentiWordNet, VADER,	-	-	96.0	-
[18]	2015	Twitter, Amazon	LIWC, GI, ANEW, SCN,WSD,	81.0	75.0	75.0	-

Table 4: Deep Learning Approach (Comparison Analysis)

Author	Year	Platform	Features and Algorithm	Precision (%)	Recall (%)	Accuracy (%)	F1-score (%)
[28]	2018	Kaggle dataset, Formspring, Google,Twitter	CNN-LSTM, Twitter Embedding	84.5	84.2	-	84.2
[29]	2019	Twitter	Bi-GRU-CNN, Bi-GRU-LSTM-CNN, TextCNN,	-	-	-	70.57
[30]	2019	Twitter	CNN, word2vec, character n-grams,	86.61	70.42	-	77.38

Table 5: Hybrid Approach (Comparison Analysis)

Author	Year	Platform	Features and Algorithm	Precision (%)	Recall (%)	Accuracy (%)	F1-score (%)
[31]	2019	Benchmark corpora	Word embedding, LSVC, LSTM and HurtLex	60.4	79.8	-	68.7
[32]	2020	Twitter	CNN and mBert	76.95	81.52	--	78.99

V. CONCLUSION

In this paper, we carried out a comprehensive review of various approaches to detect hate speech on social media platforms that have been employed in recent years, along with a brief description of comparative analysis.

The survey work is divided into five major categories: the Linguistic Rule-Based approach, Supervised Learning, Unsupervised Learning, Deep Learning, and Hybrid approaches for hate speech identification, including significant activities in those fields

Taking limited and public datasets for training hate speech detection model is one of the limitations found, and the model can be improved by using real-time big data sets. We have also found that the hate speech is not limited with texts only, but other modes of interactions, such as image and video detection, can also focus on the future.

REFERENCES

[1] E. Spertus. (1997). Smokey: automatic recognition of hostile messages. In: *Innov. Appl. Artif. Intell. - Conf. Proc.*, pp. 1058–1065.
[2] A. Abbasi & H. Chen. (2007). Affect intensity analysis of dark web forums. In: *IEEE Intell. Secur. Informatics*, pp. 282–288, 2007. DOI: 10.1109/isi.2007.379486.
[3] H. Watanabe, M. Bouazizi, & T. Ohtsuki. (2018). Hate speech on twitter: A pragmatic approach to collect hateful and offensive expressions and perform hate speech detection. *IEEE Access*, *6*, 13825–13835.

e-ISSN: 2250-0758 | p-ISSN: 2394-6962
Volume-11, Issue-2 (April 2021)
https://doi.org/10.31033/ijemr.11.2.17

[4] F. Rodriguez-Sanchez, J. Carrillo-de-Albornoz, & L. Plaza. (2020). Automatic classification of sexism in social networks: An empirical study on Twitter data. *IEEE Access*, 219563–219576.
DOI: 10.1109/ACCESS.2020.3042604.
[5] O. Oriola & E. Kotze. (2020). Evaluating machine learning techniques for detecting offensive and hate speech in South African tweets. *IEEE Access*, *8*, 21496–21509. DOI: 10.1109/ACCESS.2020.2968173.
[6] R. Cohen-Almagor. (2011). Fighting hate and bigotry on the internet. *Policy & Internet*, *3*(3), 89–114.
[7] N. Sambuli, F. Morara, & C. Mahihu. (2013). *Umati: Monitoring online dangerous speech*. Available at: http://www.ihub.co.ke/blog/wp-content/uploads/2014/06/2013-report-1.pdf.
[8] Z. Waseem & D. Hovy. (2016). *Hateful symbols or hateful people? Predictive features for hate speech detection on Twitter*.
[9] A. Founta *et al.* (2018). Large scale crowdsourcing and characterization of twitter abusive behavior. *No. Icwsm*, 491–500.
[10] M. O. Ibrohim. (2019). *Multi-label Hate Speech and Abusive Language Detection in Indonesian Twitter*.
[11] Ç. Çöltekin. (2020). *A corpus of Turkish offensive language on social media*, pp. 6174–6184.
[12] M. H. Ribeiro, P. H. Calais, Y. A. Santos, V. A. F. Almeida, & W. Meira. (2018). Characterizing and detecting hateful users on Twitter. *arXiv*, Icwsm, 676–679.
[13] P. Liu, W. Li, & L. Zou. (2019). *NULI at semeval-2019 task 6: Transfer learning for offensive language detection using bidirectional transformers*.
DOI: 10.18653/v1/s19-2011.
[14] S. D. Swamy, A. Jamatia, & B. Gambäck. (2019). Studying generalisability across abusive language detection datasets. In: *CoNLL 2019 - 23rd Conf. Comput. Nat. Lang. Learn. Proc. Conf.*, pp. 940–950.
DOI: 10.18653/v1/k19-1088.
[15] R. Kshirsagar, T. Cukuvac, K. McKeown, & S. McGregor. (2018). Predictive embeddings for hate speech detection on twitter. *arXiv*. DOI: 10.18653/v1/w18-5104.
[16] P. Mishra, H. Yannakoudakis, & E. Shutova. (2018). Neural character-based composition models for abuse detection. In: *arXiv*.
[17] J. Mitrović, B. Birkeneder, & M. Granitzer. (2015). *nlpUP at SemEval-2019 task 6: A deep neural language model for offensive language detection*, pp. 722–726.
DOI: 10.18653/v1/s19-2127.
[18] C. J. Hutto & E. Gilbert. (2014). VADER : A parsimonious rule-based model for sentiment analysis of social media text In: *Proceedings of the Eighth International AAAI Conference on Weblogs and Social Media*, pp. 216–225.
[19] N. D. Gitari, Z. Zuping, H. Damien, & J. Long. (2015). A lexicon-based approach for hate speech detection. *IJMUE*, *10*(4), 215–230.
[20] N. R. Fatahillah, P. Suryati, & C. Haryawan. (2018). Implementation of naive bayes classifier algorithm on social media (Twitter) to the teaching of Indonesian hate speech. In: *Proc. - 2017 Int. Conf. Sustain. Inf. Eng. Technol. SIET 2017*, pp. 128–131.
DOI: 10.1109/SIET.2017.8304122.
[21] M. A. Fauzi & A. Yuniarti. (2018). Ensemble method for indonesian twitter hate speech detection. *IJEECS*, 294–299. DOI: 10.11591/ijeecs.v11.i1.pp294-299.
[22] P. Sari & B. Ginting. (2019). *Hate speech detection on twitter using multinomial logistic regression classification method*, pp. 105–111.
[23] A. Briliani, B. Irawan, & C. Setianingsih. (2019). Hate speech detection in indonesian language on instagram comment section using K-nearest neighbor classification method. In: *Proc. - 2019 IEEE Int. Conf. Internet Things Intell. Syst. IoTaIS 2019*, pp. 98–104.
DOI: 10.1109/IoTaIS47347.2019.8980398.
[24] R. Zhao & K. Mao. (2016). *Cyberbullying detection based on semantic-enhanced marginalized denoising*.
DOI: 10.1109/TAFFC.2016.2531682.
[25] A. Rodriguez, C. Argueta, & Y. L. Chen. (2019). Automatic Detection of Hate Speech on Facebook Using Sentiment and Emotion Analysis. In: *1st Int. Conf. Artif. Intell. Inf. Commun. ICAIIC 2019*, pp. 169–174.
DOI: 10.1109/ICAIIC.2019.8669073.
[26] S. Jaki & T. De Smedt. (2018). *Right-wing German hate speech on twitter : Analysis and automatic detection*, pp. 1–31.
[27] M. Di Capua, E. Di Nardo, & A. Petrosino. (2016). *Unsupervised cyber bullying detection in social networks*, pp. 432–437.
[28] H. Rosa, D. Matos, L. Coheur, & P. Carvalho. (2018). A 'Deeper' look at detecting cyberbullying in social networks. In: *2018 Int. Jt. Conf. Neural Networks*, pp. 1–8.
DOI: 10.1109/IJCNN.2018.8489211.
[29] T. Van Huynh, D. Nguyen, K. Van Nguyen, N. L. Nguyen, & A. G. Nguyen. (2019). *Hate speech detection on vietnamese social media text using the Bi-GRU-LSTM-CNN Model*.
[30] B. Gambäck & U. K. Sikdar. (2017). Using convolutional neural networks to classify hate-speech. *No. 7491*, 85–90.
[31] E. W. Pamungkas, V. Patti, & D. Informatica. (2019). *Cross-domain and cross-lingual abusive language detection : A hybrid approach with deep learning and a multilingual lexicon*, pp. 363–370.
[32] S. Alsafari, S. Sadaoui, & M. Mouhoub. (2020). Hate and offensive speech detection on Arabic social media. *Online Soc. Networks Media*, *19*, 100096, 2020.
DOI: 10.1016/j.osnem.2020.100096.

www.ijemr.net ISSN (Online): 2250-0758 || ISSN (Print): 2394-6962

Publication Impact Factor (PIF) for 2020: 5.640

VANDANA PUBLICATIONS

International Journal of Engineering and Management Research

Ref No: IJEMR/V-11/I-2/16/2021 **(Section – A)** **Date:** 28-04-2021

Certificate of Publication

This is to certify that Research Paper title **"Automatic Hate Speech Detection: A Literature Review",** by **"Mohiyaddeen"** has been published with the "International Journal of Engineering and Management Research", Volume-11, Issue-2 of April 2021.

(Editor-in-Chief)
International Journal of Engineering and Management Research (IJEMR)

Vandana Publications IJEMR

Improved Hate speech Detection System using Multi-Layers Hybrid Machine Learning Model

Mohiyaddeen[1], Sifatullah Siddiqi[2] and Faiyaz Ahmad[3]

[1]Integral University, Lucknow 226026, India
[2]Integral University, Lucknow 226026, India
[3]Integral University, Lucknow 226026, India
moinsheikhmt@gmail.com

Abstract. Hate speech and Cyber harassment have been a major concern on the internet for a long time. Furthermore, Social media platforms, particularly Twitter and Facebook, have elevated it to a worldwide platform on which hate speeches can spread much faster. Manually detecting hate speech from social media is a time-consuming process. Hence, several studies are still being conducted in this field. We build a two-layer hybrid machine learning model using existing machine learning algorithms. This Hybrid machine learning algorithm is capable of efficiently detecting hate speech from social media texts. The hybrid approach combines nine different machine learning algorithms to make one hybrid machine learning model.
Additionally, we used the Bag of Words and TFIDF techniques with the two-gram approach to extract the features. Significant experiments are carried out on the Hate speech dataset. The accuracy gained by the hybrid machine learning model is much higher than that of available conventional machine learning models.

Keywords: Natural Language Processing, Hybrid Machine learning, Classification Algorithm.

1 Introduction

The easiest way to connect to people is through social networking sites. However, since social networking sites have become global, people found an illegitimate and unethical way to use these platforms. And the most popular and dangerous misuse of online social media platforms is Hate speech. We can describe hate speech as abusive language, hatefulness, threats, racism, cyberbullying, aggression, insults, provocation, personal attacks, or sexism. These are some main threats to an online social media platform.

An online study conducted by cybersecurity solutions provider Norton by Symantec found that eight out of 10 individuals in India have encountered some online abuse at some point in their lives, including 41 percent of women who experienced sexual harassment on social networking sites [1].

All kinds of Tweets and posts on social media are in the form of text data called unstructured data. Most of the social media platforms provide their APIs (Application

Programming Interfaces) to enable programmers or researchers to collect their public data. Some examples are Twitter API, Facebook Graph API, YouTube API, and Reddit API. Another way of collecting valuable data from the social media platform is Web Scraping. Web Scraping is the technique to extract a large amount of data from websites and save it locally in a database.

We use Natural Language Processing (NLP) and machine learning to extract the important words from unstructured text data. We all know that machine learning algorithms cannot understand texts or characters, so it is very important to convert them into a machine-understandable format (such as numbers or binary) to perform any kind of analysis on text data. The ability to make machines understand and interpret text data is termed natural language processing [2].

This paper has implemented a hybrid machine learning model that can classify the text into two categories: "Hate speech" and "Normal speech" using Machine learning libraries and Natural language processing.

This paper has implemented a hybrid machine learning model that can classify the text into two categories: "Hate speech" and "Normal speech" using Machine learning libraries and Natural language processing.

The proposed hybrid machine learning algorithm combined six distinct machine learning algorithms to create a hybrid machine learning model. We also employed the Bag of Words and TFIDF techniques with the two-gram approach. Extensive tests are conducted on the Hate speech dataset, and the accuracy attained by the multi-layer hybrid machine learning model is much greater than that of standard machine learning models

2 Related Works

Research interest in Natural language Processing and emotion analysis has grown significantly due to the increased usage of social media websites. Recently, several types of research have been published which classify the tweets and texts. To identify hate speech in texts, they employed a variety of machine learning algorithms.

[3] Proposed an approach that collected dataset from Kaggle, and then they divided it into two subsets: a training dataset, which contains 31962 records, and a test dataset, which includes 49159 records. In order to process the data, tweets are converted into lowercase and then removed unnecessary content such as punctuations, numbers, special characters, hashtags, and stop words. They have also used stemming techniques to convert the terms into root words. To extract the features from the tweets, they used TF-IDF, and the Bag of words feature engineering methodologies. After performing data processing, logistic regression classifier is used to categorize the tweets into hate speech or not hate speech. With the Bag of words features, the logistic regression model obtained 94.11% accuracy, whereas the logistic regression with the TF-IDF feature gives an accuracy of 94.62%. And the result shows that logistic regression classifiers with Bag of words and TF-IDF provide almost the same accuracy.

In [4], they worked on hate speech detection in the Indonesian language by implementing four machine learning algorithms, including Naive Bayes, SVM, BLR, and

RFDT. They used Twitter data as the source of the dataset and collected the tweets using Twitter Streaming API. Only 1,100 tweets were collected and labeled manually as hate speech (HS) or not hate speech (non-HS). The dataset was annotated manually by 30 volunteers. For the pre-processing step, they (1) removed retweeted text, (2) removed unnecessary words, (3) converted the text into lowercase, (4) corrected the spelling mistakes, and (5) removed the stop words. In this paper, they simply extracted N-gram features from the tweets, and then these features are vectorized according to their Bag of words (BOW). Weka platform is used to conduct all experiments. Based on the experimental results, they found an n-gram feature with RFDT (93.5%), BLR (91.5%), NB (90.2%), and SVM (86.5%) of accuracy. 10- fold cross-validation has been performed on test data, and the result shows that RFDT performed best with 93.5% accuracy.

[5] Proposed a systematic approach to detect hate speech in the Bangla language. Web Scraping is a traditional way to extract data from social media and websites. In this paper, the author has used the Web Scraping technique to extract the data from Facebook. Data collected from Facebook is in the form of text. To make this data useful, they labeled the data into two categories. One category was general speech, and another category was hate speech. In the pre-processing phase, they removed emojis, spelling mistakes, and other unnecessary words. To evaluate the training data set, they have performed data analysis to balance hate speech and normal speech. After pre-processing phase, they have used count vectorizer and TF-IDF feature extraction techniques to extract the features. For classification, they have used two algorithms, Naive Bayes and Support Vector Machine. Applying the SVM gives them a 70% accuracy, while the Naive Bayes has a 72% accuracy. The result shows that Naive Bayes performed best with 72% accuracy.

[6] Proposed an approach to classify Anti-social comments based on the kNN algorithm. At first, they used text based ASB Corpus [Anti-social Behavior corpus for harmful language detection], which consists of a collection of 148 documents that have been concluded to be abusive, offensive, and aggressive. After data collection, pre-processing phase requires (1) to remove stop words, (2) from converting the text in lower case, (3) to Stemming words. In order to obtain the features from the pre-processed data, they implemented a Term Document Matrix. The author found the most common terms via the term-document matrix and eliminated the Sparse terms from the matrix. The resulting matrix is given input into the kNN algorithm. The region space is drawn and divided according to the type of the data set. The result shows that the dark Point in kNN Area Space reflects Non-Anti-Social Texts While the lighter Point represents the Anti-social text.

3 Proposed Approach

This research aims to identify whether the given social media text is hateful or not. We have built a hybrid machine learning algorithm using nine different machine-learning algorithms to accomplish the objective.

3.1 Data Collection

To achieve the objective of this work, we have collected the dataset from two sources Crowd Flower dataset [7] and the HateSpeechData dataset [8]. The collected dataset included duplicate tuples along with multiple features. We modified this dataset into two features and filtered out the duplicate tuples. Dataset is separated into two subsets to perform classification algorithms: Label and Tweet. Label feature has two classes, 0 and 1. where 1 indicates hateful texts and 0 represents texts that are not hateful. Initially, we performed pre-processing on the 55591 rows of the dataset and completed training on 14000 rows of the dataset.

3.2 Imbalance Dataset

Imbalance Dataset is a key challenge in classification problems when the labeled classes in the training dataset are not evenly distributed. Since we have two classes in our dataset, we maintained a balance of around 50 percent for the Not-Hate speech class and 50 percent for the Hate speech class. The distribution of hate speech in the dataset is illustrated in Figure 1.

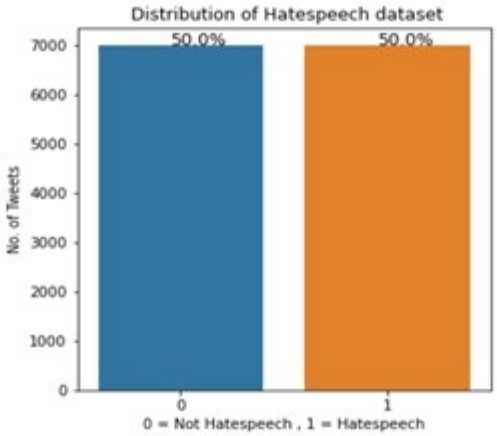


Fig. 1. Balanced dataset

3.3 Data Preprocessing

Collected raw data includes unnecessary, incomplete, inconsistent, and duplicate data that are unable to provide higher accuracy, hence diminishing its performance. It is an important step that must be completed prior to supplying the dataset to the machine learning algorithm. Figure 2 is representing the various pre-processing procedures.

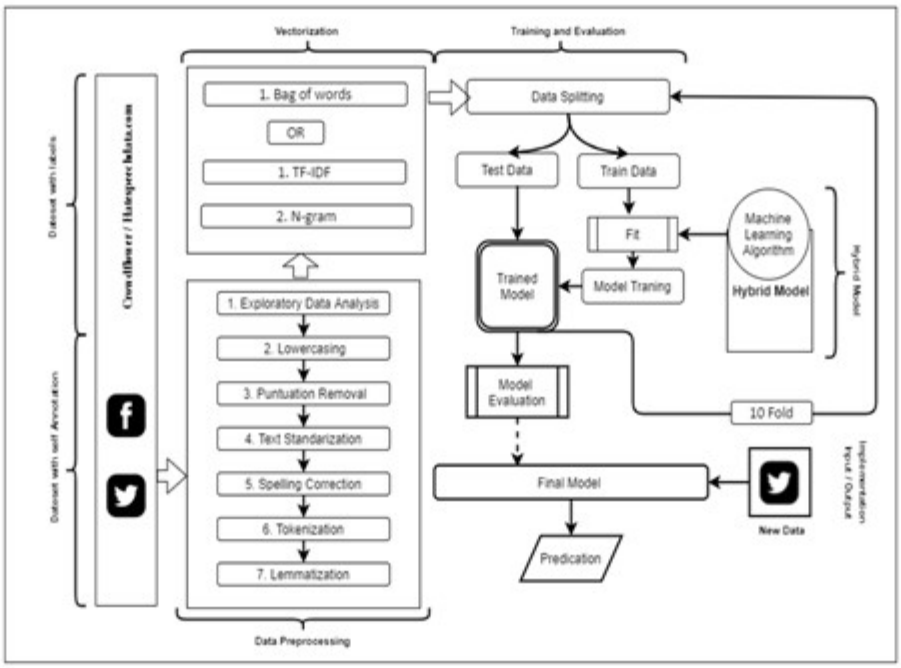


Fig. 2. The proposed framework of Hate speech Detection

Pre-processing of the dataset is performed in the following manner.

Label Encoding. We have two classes in our hate speech dataset: Hate speech and Normal speech. During the label encoding procedure, we converted the Hate speech class into 1, whereas the Normal speech converted into 0.

Lowercasing Texts. In this process, we transformed the words into lowercase. This phase is required to vectorize the dataset, and it must be completed at the beginning of the pre-processing phase.

Punctuation. Hashtags (#) and User tags (@) do not provide any information. Therefore, I eliminated all unnecessary special characters, numbers, and punctuations.

Accented Characters. Accented characters and emojis do not convey vital information. Thus, I removed any characters that were associated with emojis or accented characters from the dataset.

Stop words. Stop words are words that do not contribute any sense to a sentence. They may be easily omitted without altering the meaning of the sentence's overall structure. We are using Spacy Library to remove Stop words because Spacy supports 326 stop words. We have also created a custom set of stop words that is capable of removing unnecessary words not related to this research.

Tokenization. In this phase, we divided the chunk of text into distinct words depending on the presence or absence of a certain delimiter.

Lemmatization. We have employed lemmatization for retrieving a root word by considering vocabulary. For example, the words "bad," "worse," and "worst" are lemmatized into the word "bad.".

Word frequency. Word frequency is the way to visualize the top occurring words in the dataset. Figure 3 shows the top 20 most frequent words in our Hate speech dataset.

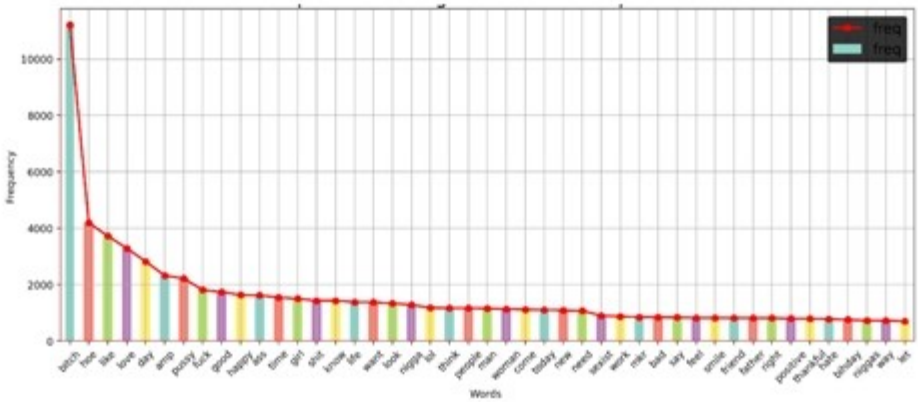


Fig. 3. Most frequently occurring words in the Hate speech dataset

Word cloud. After applying all the above pre-processing steps, we can visualize the tokenized dataset into a cloud form. Figure 3 is showing the most used word in Word Cloud view.

Fig. 4. Word Cloud view of the pre-processed dataset

3.4 Feature Extraction

Collected raw data includes unnecessary, incomplete, inconsistent, and duplicate data that are unable to provide higher accuracy, hence diminishing its performance. It is an important step that must be completed prior to supplying the dataset to the machine learning algorithm. Figure 2 is representing the various pre-processing procedures.

Bag of words. Bags of words is a technique to extract the essential features from the set of documents or corpus. The term corpus is a collection of documents, and the document is referred to a group of words. Each word or text in the document is called a token.

In the Bag of words technique, each tuple is considered as a separate document and tries to determine how often each unique word is used in all documents or corpus. This technique focuses on the frequency of a word's appearance inside a document, but it does not give any information wherein the document the term appears. So, if two tuples have similar words, then it declares both documents equal.

In this work, we employed the Bag of words technique to compare the result with the TFIDF technique, which is another method for vectorization.

Term Frequency-Inverse Document Frequency. TFIDF is another more efficient approach to distribute a weightage to each word in the corpus. It determines the significance of a term ti in the given document Dj.

Term Frequency. TF measures how often a term occurs in the given document. We can formulate the equation as follows.

$$tf(t,d) = Number\ of\ t\ in\ d\ /\ Total\ count\ of\ the\ words\ in\ d \quad (1)$$

Here, t is a term (word), and d is a document/set of words.

Inverse Document Frequency. It decreases the weightage of most frequent words and increases the weightage of the rare words in the given document.

$$idf(t,d) = log\ (|D|\ /\ |\{d{:}t_i \in d\}|) \quad (2)$$

Here, D is the total set of documents, t is a term/word, and d is a document.
We get the TFIDF equation by multiplying the equation (1) and (2)

$$tfidf(t,d) = tf(t,d) * idf(t,d) \quad (3)$$

equation (3) can evaluate the weight of any term/word based on the importance of that word/term in the whole corpus.

3.5 Proposed Model

This work aims to improve the performance and accuracy of the hate speech detection model using the Hybrid Machine Learning Algorithm. We are using 2 Layers of a Hybrid Machine Learning algorithm to train our model. I have stacked various ML models in the first layer, including Support vector machine, Logistic regression, K nearest neighbor, Naive Bayes, Multinomial Naïve Bayes, Decision Trees, Random Forest, MLP Classifier, Ada Boost Classifier. In the second layer, I have selected Logistic regression as a Meta classifier to improve the prediction. The following is a list of the numerous stages that were taken during the project.

Formation of Hybrid ML Algorithm. In this research, we are using two layers hybrid algorithm to improve the model's accuracy. To implement this algorithm, we have categorized the various machine learning models into two layers: Base and Meta layer. To build the base layer, we employed nine Machine learning Algorithms, including Support Vector Machine, Decision Tree, Random Forest, Logistic regression, Bernoulli Naïve Bayes, AdaBoost Classifier, MLP Classifier, Multinomial Naïve Bayes, and KNeighbour Classifier. In the second layer, we have employed Logistic regression as a meta-layer algorithm.

We can formulate the hybrid machine learning algorithm as follows:

$$\sum_{j=1}^{n} \sum_{i=1}^{j} \left[\left(ML_i + LM_j \right) + GM \right] \quad (4)$$

Where ML_i = Various types of Machine Learning Algorithm
LM_i = Respective Local Meta Classifiers Algorithm

GM = Global Meta Classifier Algorithm
The previously described formula can be represented as a flowchart, as shown in Figure 6.

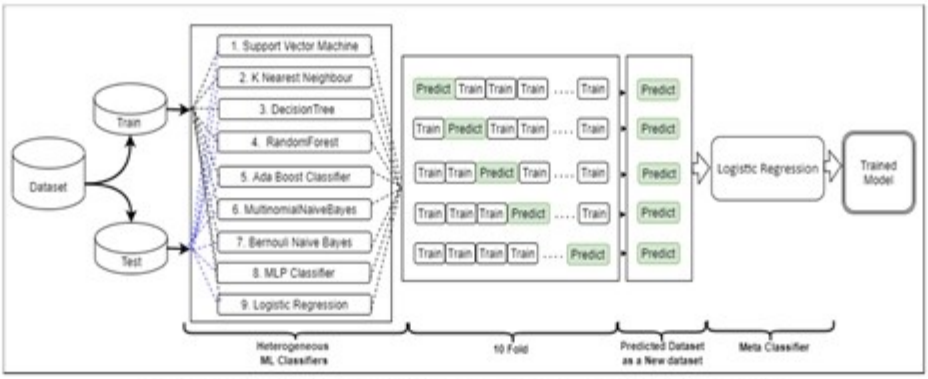


Fig. 5. Working Flowchart of Hybrid Machine Learning Algorithm

We split the main dataset into train and test data using the 10-fold cross-validation technique in the base layer. 10-fold cross-validation is a method to divide the dataset into ten subsets. Thus, the process is repeated ten times. Moreover, one of the ten subsets is selected during each iteration as a test dataset, and the other (10-1 = 9) subsets are combined to create a training dataset. This procedure implements on each machine learning model.

We then collect all predicted datasets generated by the machine learning model in the base layer and created a new dataset to feed into the meta classifier. Here we have employed Logistic Regression as a meta classifier. This phase also follows the 10-fold cross-validation technique to generate the final model.

Performance of Conventional ML Model After pre-processing, we trained our model using nine different Machine Learning algorithms, and the results were as follows.: Support vector machine (94.17%), Logistic regression (94.16%), K nearest neighbor (88.15%), Bernoulli Naïve Bayes (92.59%), Multinomial Naïve Bayes (91.62%), Decision Trees (91.52%), Random Forest (93.39%), MLP Classifier (91.23%), AdaBoost Classifier.(92.36%). The performance of nine machine learning models on our dataset is shown in Figure 5. Support vector machine is the top performer, with a score of 94.17 percent.

Fig. 6.

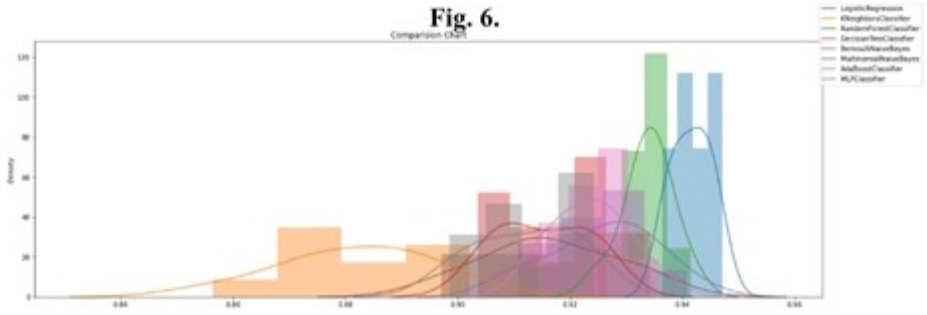


Fig. 7. Comparison Chart of various Machine learning models on Hate speech dataset

Performance of Hybrid ML Model We have implemented our pre-processed Hate speech dataset with the TFIDF vectorization technique to our two-layer hybrid model. After predicting with the test dataset, we have found that the Two layers ML Model Approach performed much better than the other 9 Machine learning models. Figure 7 shows a comparison plot between the two-layer ML model with other ML models.

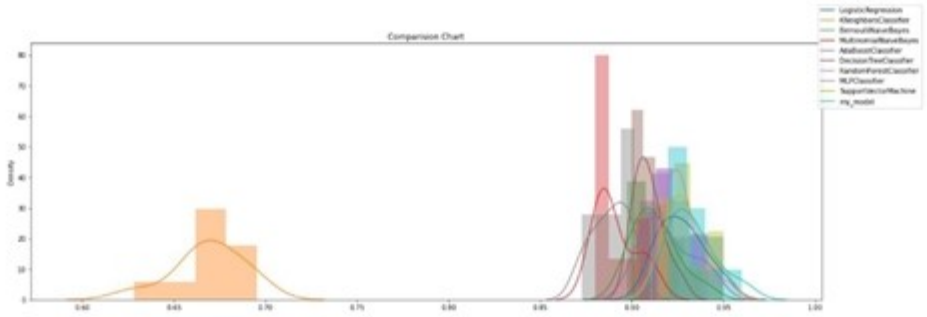


Fig. 8. Comparison between Hybrid ML model with other nine ML Models

To evaluate the performance, we used a confusion matrix to determine the accuracy of our classification model, and the results were positive. As shown in Figure 8, only 3.55% of the dataset lies on False Negative, and 1.78% of the dataset lies on False Positive portion. The confusion matrix demonstrates that our performance is excellent, with the least amount of error. The model's accuracy is 94.68 %, and the F1 score is 94.59 %, which shows that the model is highly accurate.

Fig. 9. Confusion Matrix of Two-layer ML Model

4 Results

The proposed model is implemented on hate speech datasets which are publicly available on Crowd flower. The dataset was imbalanced. Hence, we balanced it and employed various pre-processing steps to clean the dataset. We are using the TFIDF technique to vectorize and extract the features. 2-gram has been used to extract more accurate features. Using a traditional machine learning algorithm, we have built a hybrid machine learning model to implement on hate speech dataset. We applied six machine learning algorithms in the first layer to train our dataset through a 10-fold cross-validation technique. The 10-fold technique is used here to reduce the overfitting problem. The predicted dataset generated by the first layer is then given to our second layer, which is logistic regression in this case. The accuracy and f-score of the final model are far better than other conventional machine learning algorithms, as shown in table 1. This paper proposed a more effective technique for detecting hate speech text using a hybrid machine.

Table 1. A comparison between 9 different machine learning models with our proposed Hybrid Model

Sr. No.	Model Name	Precision	Recall	Accuracy	F_score
1	Logistic Regression	97.17	90.99	94.16	93.97
2	RandomForestClassifier	94.02	92.67	93.39	93.34
3	BernoulliNaiveBayes	91.64	93.74	92.59	92.68
4	MultinomialNaiveBayes	89.22	94.70	91.62	91.87
5	AdaBoostClassifier	97.71	86.74	92.36	91.90
6	SupportVectorMachine	97.80	90.37	94.17	93.94
7	Decision TreeClassifier	90.36	92.96	91.52	91.64
8	MLP Classifier	90.86	91.69	91.23	91.27
9	KNeighborsClassifier	95.56	80.01	88.15	87.10
10	My_Hybrid Model	96.31	93.00	94.78	94.59

The approach used in [1] was a logistic regression with the TFIDF technique, and they trained their dataset with 81121 rows, but they achieved only 94.62 % accuracy. On the other hand, our proposed approach is based on a multi-layer hybrid machine learning model with a 10-fold cross-validation technique. We have taken 14000 datasets only and achieved 94.78 % accuracy. Which is better than other recently published work in the same field.

5 Conclusion

In this paper, we proposed a more effective technique for detecting hate speech text using a hybrid machine learning algorithm with the help of the TFIDF and Bag of words approach. We employed nine heterogeneous machine learning models to create the base layer of the algorithm, and logistic regression is used to build a meta classifier for the algorithm. We are using a 10-fold cross-validation technique to split our dataset into train and test. Using Bag of word with 2-gram and Hybrid machine learning algorithm, we obtained 94.72% accuracy, while TFIDF with 2-gram gives an accuracy of 94.78 %. They have nearly the same accuracy. We have provided a final comparison chart between other machine learning models and our Hybrid machine learning model. The performance of our model is much better than other machine learning models.

Recently published work on hate speech detection is focused on conventional machine learning models, and the accuracy gained by them varies between 70% to 90%. This research is based on a multi-layer hybrid machine learning model with a 10-fold cross-validation approach, ensuring the accuracy and f-score of more than 94 percent.

References

1. Bhargava, Y.: 8 out of 10 Indians have faced online harassment - The Hindu, https://www.thehindu.com/news/national/8-out-of-10-indians-have-faced-online-harassment/article19798215.ece, last accessed 2021/02/18.
2. Kulkarni, A., Shivananda, A.: Natural Language Processing Recipes: Unlocking Text Data with Machine Learning and Deep Learning using Python. Apress (2019).Author, F., Author, S., Author, T.: Book title. 2nd edn. Publisher, Location (1999).
3. Koushik, G., Rajeswari, K., Muthusamy, SK: Automated hate speech detection on Twitter. Proc. - 2019 5th Int. Conf. Comput. Commun. Control Autom. ICCUBEA 2019. (2019).
4. Alfina, I., Mulia, R., Fanany, M.I., Ekanata, Y.: Hate speech detection in the Indonesian language: A dataset and preliminary study. 2017 Int. Conf. Adv. Comput. Sci. Inf. Syst. ICACSIS 2017. 2018-January, 233–237 (2018).
5. Ahammed, S., Rahman, M., Niloy, M.H., Chowdhury, SMMH: Implementation of Machine Learning to Detect Hate Speech in Bangla Language. Proc. 2019 8th Int. Conf. Syst. Model. Adv. Res. Trends, SMART 2019. 317–320 (2020).
6. Chandra, N., Khatri, S.K., Som, S.: Anti social comment classification based on kNN algorithm. 2017 6th Int. Conf. Reliab. Infocom Technol. Optim. Trends Futur. Dir. ICRITO 2017. 2018-Janua, 348–354 (2018).
7. Hate Speech Identification - dataset by crowdflower | data.world, https://data.world/crowdflower/hate-speech-identification, last accessed 2021/06/22.

Paper accepted to publish in Springer Book Series, 'Algorithms for Intelligence System'

CVR 2021 notification for paper 116: Accepted

CVR 2021 <cvr2021@easychair.org> Thu, Jul 15, 2021 at 11:40 AM
To: "Mr. Mohiyaddeen" <moinsheikhmt@gmail.com>

Dear Mr. Mohiyaddeen,

Greetings!

Thank you for submitting your research article to the International Conference on Computer Vision and Robotics (CVR 2021).

We are pleased to inform you that based on reviewers' comments your paper titled "Improved Hate Speech Detection System using Multi-Layers Hybrid Machine Learning Model" has been accepted for presentation during the CVR 2021 and publication in the proceedings to be published in **Springer Book Series, 'Algorithms for Intelligent Systems'** subject to the condition that you submit a revised version as per comments. The reviewer's comments have been sent to the corresponding author. It is also required that you prepare a response to each comment from the reviewer and upload it as a separate file along with the paper. The similarity index in the final paper must be less than 20%. Please note that the high plagiarism and any kind of multiple submissions of this paper to other conferences or journals will lead to rejection at any stage.

Please carry out the steps to submit the camera-ready paper and online registration as per the instructions available at https://www.cvr2021.scrs.in/page/camera-ready-paper-submission

Please note that the Last date for submission of the camera-ready paper, payment of registration fee, and online registration is July 25, 2021.

Feel free to write to the "General Chairs, CVR 2021" at cvr.scrs@gmail.com should you have any questions or concerns. Please remember to always include your Paper ID- 116, whenever inquiring about your paper.

Looking forward to meeting you online during the conference.

With Regards

Team CVR 2021