Conceptual Clarity and Terminology of Hate-speech Intensity Scale

ABSTRACT

At the beginning time, emotion resolution or analysis is a method that is computing or expressing feelings, opinions, emotions, sentiments, and text subjectivity deeply. In modern times people are mostly used different type’s social media platforms for expressing themselves in front of others with a variety of techniques, views, and statements. Here, the main focusable contribution of this study is the detecting Hate and its four categories of hate emotions from English text data which is coming from after converting speech or voice to text using the Linear Support Vector Machine (SVM), the Long-Short Term Memory (LSTM), the Bidirectional Long-Short Term Memory (Bi-LSTM), the Decision Tree, and the Logistic Regression machine learning algorithms, or classifiers with various features Part-of-speech (POS) Tagging, Uni-gram, Bi-gram, Tri-gram, Term Frequency-Inverse Document Frequency (TF-IDF), and Word2Vec. So, using a labelled suitable dataset, the collection of the English text, or comments are curated for categorized four hate emotions before appreciating hate emotion from the text. Applying 20485 training datasets and 6455 testing datasets on five different approaches (e.g., Linear SVM, LSTM, Bi-LSTM, Decision Tree, and Logistic Regression), where it is shown the best accuracy that 87.07% from using the Bi-LSTM classifier whither SVM gives 86.18%, LSTM gives 82.09%, Decision Tree gives 72.23%, and Logistic Regression gives 85.16% accuracy.



KEYWORDS

Hate-speech, Toxic word, hate emotion, Natural Language Processing, Text classifications, Machine Learning, social media, Hate comment, Internet.

1 Introduction

The detection of Hate Speech from speech to text is becoming an interesting as well as an important part of Natural Language Processing in the Machine Learning Domain in recent decennia. In the past few years, researchers give their interest in detecting Sentiment or Emotion from the text. Here, emotion categorization is the process where it is indicated that studying a speaker's expression to determine if he or she has a positive or negative attitude about a certain issue. However, this concept has become important for researchers because they realize that the impression of this topic and its bright outcome. Working with this concept is often difficult for everyone when the person must have to reach an actual agreement that exactly to identify which type of speech acts under hate speech topic.



1. Hate Speech

Before describing the whole paper, one must have to be familiar with the word Hate Speech. “Hate Speech” is a pathological concept and a category of Emotions. There has not any universally adopted definition of it in International Human Rights Law. But some people are tried to define it according to their perception. Therefore, the definition of “Hate Speech” it might be a more complex concept which is composed of two basic concepts, they are Hate and Speech. Consequently, it can be split into two main components according to give basic concept:



**Hate:** Mainly, hate word indicates, the intense and irrational emotion of opprobrium, enmity, and detestation towards an individual or group, targeted because of their having certain – actual or perceived – protected characteristics. In additionally "Hate" is more than preference and must be so discriminative. In short, Hate is a gesture of an emotional state, opinion, or thought, and therefore it is separate from any manifests action or act.

**Speech:** Here, the Speech word indicates any expression which is imparting opinions, action, or ideas that bring subjective opinions, action, or ideas to an external audience, or listener. It can look, and take many forms such as written, non-verbal, visual, or artistic, and can be propagated through any kind of media, like the internet, print, radio, or television.

After going through these two basic elements, and to put it simply, one can say that “Hate Speech” is a type of emotional expression which indicates hate towards people.

1. Classes of Hate Speech

As this paper is built for describing detected hate, contempt, or toxic speech from voice to text which are tested by different methodology and algorithms, one must have to be familiar with what kinds of Hate Speech are detected. In shortly, receivers must be familiar with the classes of Hate Speech that are detected in this whole research. So, here four different hate emotions or speech classes such as stupidity, pathetic, greedy, and retarded are detected.

This paper will be shown four types of Hate Speech (stupidity, pathetic, greedy, and retarded) from a text, line, or paragraph with previously used methodology and new suggest algorithms. This process can automate the analysis of a user's reaction towards a specific emotion like newspaper, book, or social media. With more and more people expressing their antithetic opinions openly on social networking sites, analyzing the sentiment of comments made about a specific text, movie and advertisement indicate how they feel. This paper will be accentuation some objectives like:

* This research proposes the detection of the four hate emotions namely stupidity, pathetic, greedy, and retarded for English text or speech.
* Supervised and unsupervised learning methodology can easily be applied to English text for hate or detest emotion analysis or opinion mining for the specific domain of the text.



* Ambiguous and unambiguous despise emotions for English text can be handled using the semi-supervised methodology.



* One more direction is, based on English text, the identification of overlapped comments like user's comments on each other about a subject and dual emotion detection as like the same comment is stupidity for one and greedy for another people.
* The project can score multiple dislike emotions at a time with a percentage.
* It can also work to remove punctuation with emoticons, sentence splits.

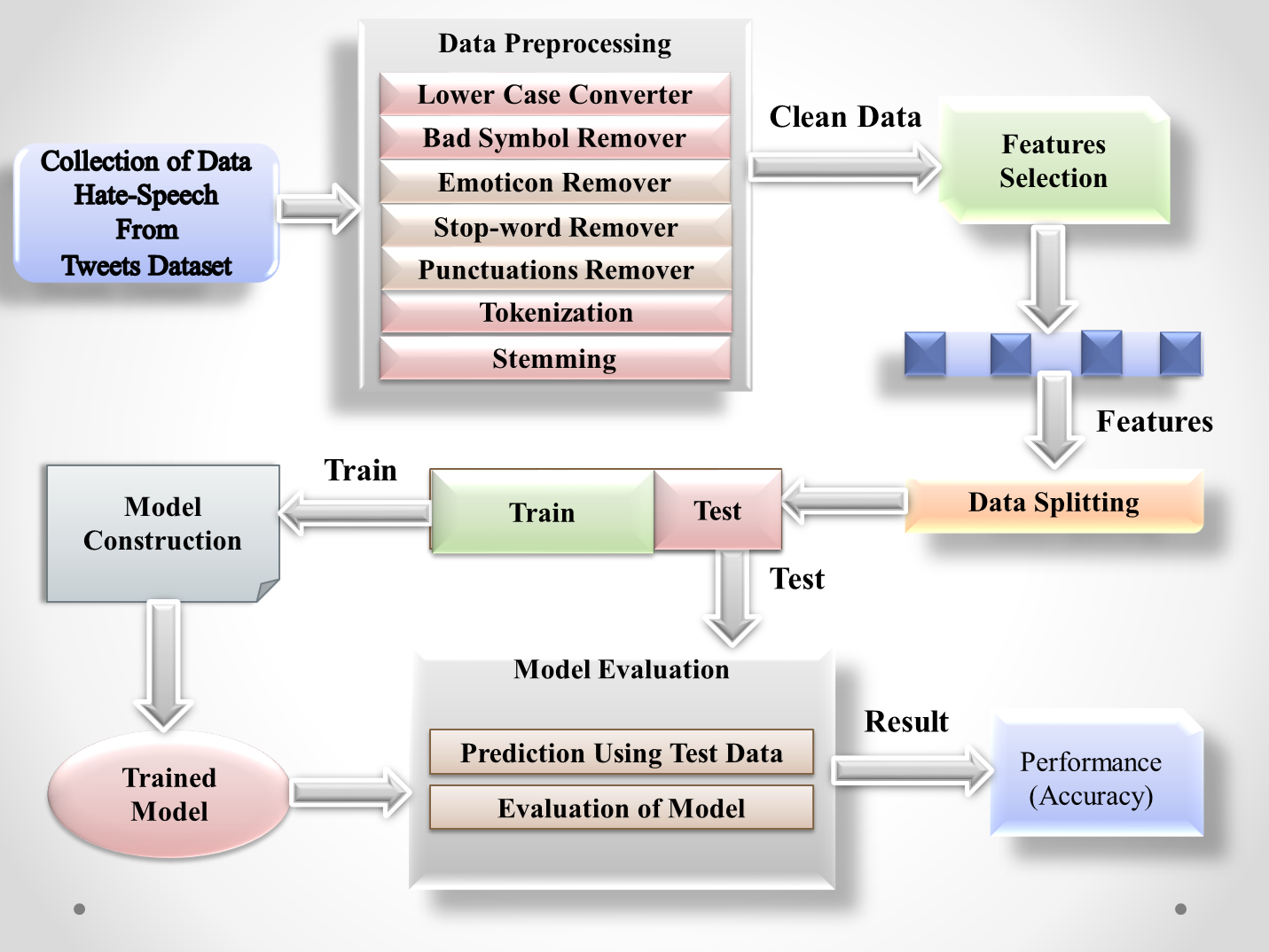
2 Literature review

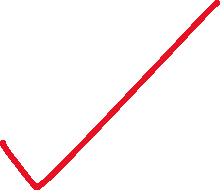
This concept is largely motivated by numerous articles that attempted to categorize various groups, such as aggression and emotions. And it is one of the important works that have been following. For highly resourced languages like English, Arabic, and other European languages, emotion analysis or hate-speech detection from text is a well-studied research topic. Recent research has placed a greater emphasis on acknowledgment. A tremendous rise in the internet and social networks usages has been seen for a couple of years. But unfortunately, the increase of the dark side in toxic speech is greater than the absolute side. Toxic or hate speech can be defined as feisty and uncertain, abusive, or threatening speech or writing that expresses against a particular group based on gender discrimination, religion, nation, denomination, and so on. The interest of research on hate speech begins in 2014 (from 2013 to 2018) which is revealed in the volume of Web of Science (WoS) – indexed production [1]. In this time another author uses semantic content analysis using Natural Language Processing (NLP) and Deep Neural Networks (DNN) methodologies for detecting non-toxic and toxic words [2]. Using those methodologies, they found their result 88% and 78% (for non-toxic using NPL and DNN) and 6.5% and 10.9% (for toxic using NPL and DNN) with 200 datasets [2]. Observing 104 articles a group of author’s works on racism and hate speech around 500 words as a result they find 40.38% results by using qualitative methods and 35.58% results by using quantitative methods in different geographic regions from social media platforms [3]. Another crew works to discover hate speech and international legislation in four countries (Egypt, Kenya, Serbia, and South Africa) with BERT fine-tuning, fastText embedding, and BERT embedding methodologies. As a result, they find accuracy nearly 78.0%, 75.7% and 76.2% [4]. After that, we find some other researchers used BOW and SVM methodologies for clearly outperforming results such as 0.3140% and 0.7941%. Where SVM gives, the highest level of accuracy about 0.4201% [5]. Afterward, another troupe authors are seen to grant interest to find out hate speech from tweeter data with 860 tweets using HASOC (only) and HASOC (offensEval) which gives result 0.7945% and 0.7976% in Macro F1-score. On other hand in Weighted F1-score gives by using HASOC (only) and HASOC (offensEval) which gives results of 0.8426 % and 0.8504% [6]. From social network sites other group searcher works on Recurrent Neural Network named Long Short-Term Memory (LSTM) and Support Vector Machine (SVM) where SVM is given strong performance rather than LSTM. In this observation, they find 410 weak hate, 130 strong hate from 540 words data set [7]. From Twitter and Wikipedia, a couple of researchers used word-based, sub word-based, joint word & character and embedding methodology for detecting hate speech. As a result, they find a good result with 83.0%, 71.2%, 64.1% and 71.5% [8]. On 17 June 2021, a fellowship investigator is used CNN-GRU model, Long Short-Term Memory (LSTM) network, Gradient-Boosted Decision Tree (GBDT) to compare their result from others [9]. After that a few, researchers are continuing this type of research from serval social networks such as Yahoo (950K), Twitter (160K), news (112K), Wikipedia (115K), Facebook (300K). But they only work on detecting three types of emotion such as threat, identity hates, and insult with 78%, 88%, 43% using CNN, LSTM, and GRU methodology [10].

3 Methods

Some advanced models are selected for detecting the Hate-Speech from text or comments and divides those comments or text into four Hate emotion categories (e.g., stupidity, pathetic, greedy, and retarded). To detect the Hate-Speech from text or comments, so many steps have too across. The whole process is shortly shown in figure: 1. each of the steps is discussed with details in the approximate sections.

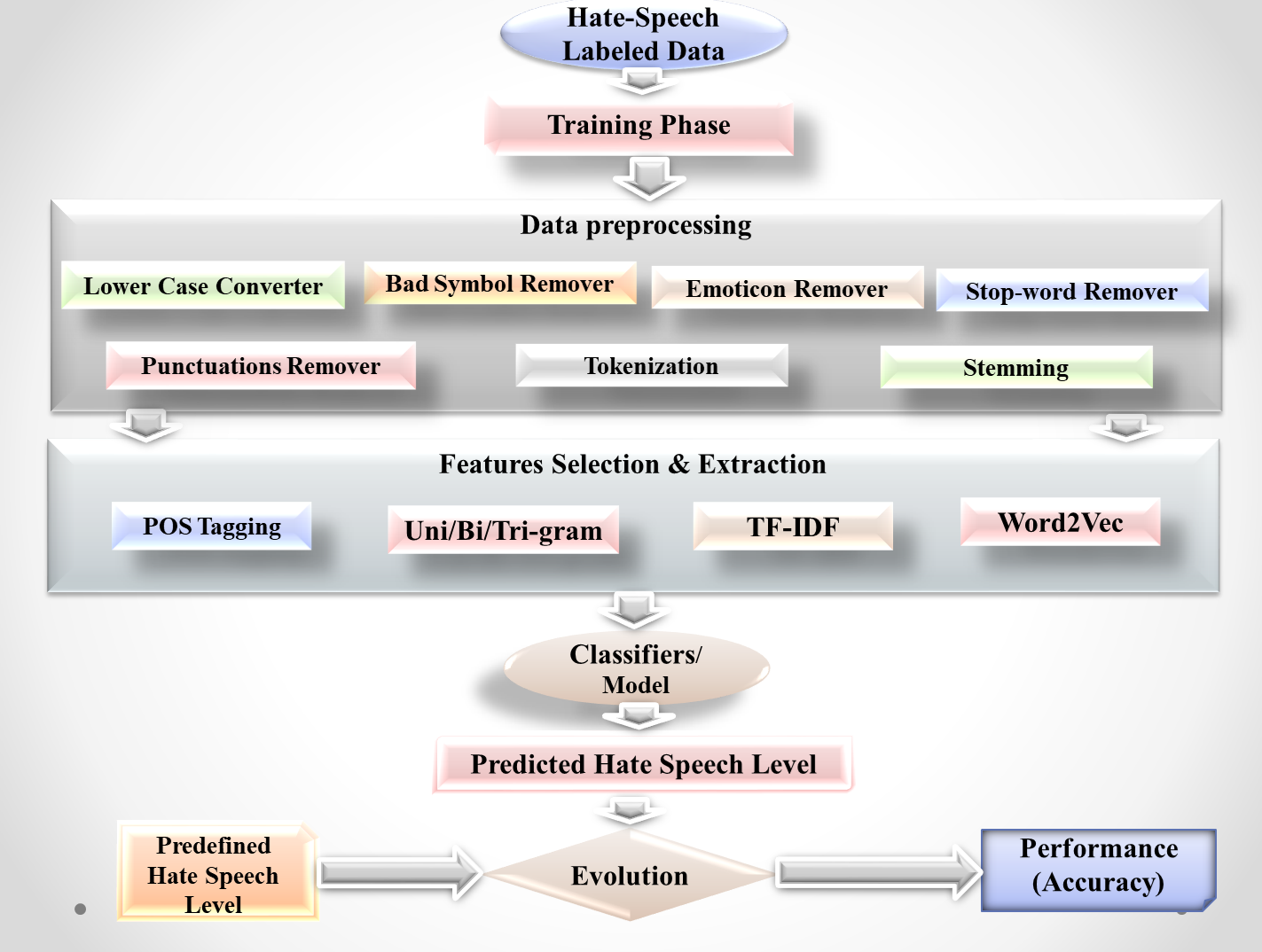






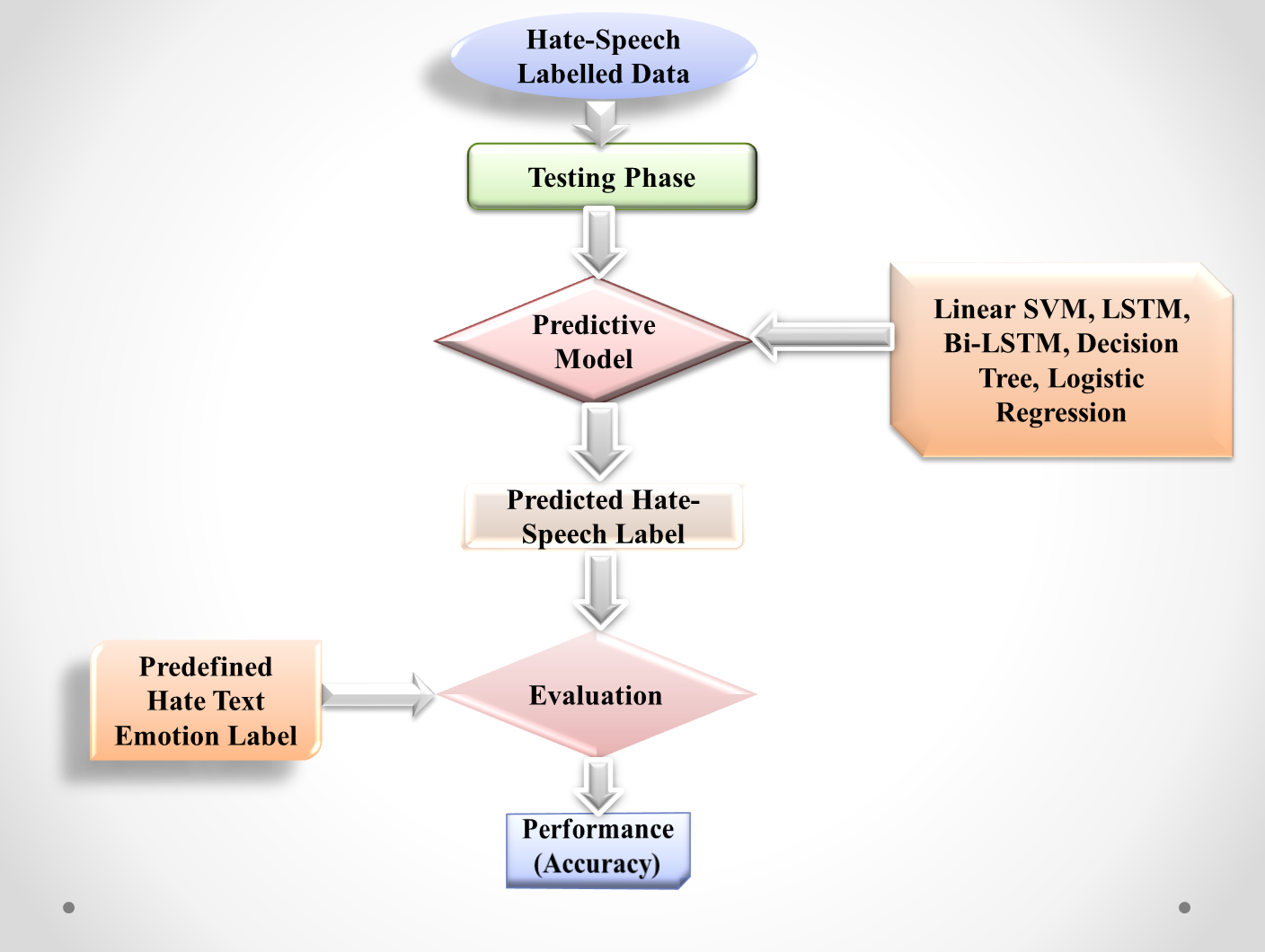
**Fig. 1** System Overview

Therefore, the system can be separated into two phases: training and testing to understand the whole process spontaneously and clearly. For this reason, at first, the entire dataset was first sorted into training and test data. Then before feeding the training set to the classifier, it is gone through so many pre-processing phases, and then several feature selections approaches are performed. After that, the test data run through the same process and predicting possible hate speech with emotion labels for each document. So, the system's efficiency is then evaluated and can compare with the predefined emotion labels to the predicted ones.





**Fig. 2** Architecture of Training Phase



**Fig. 3** Architecture of Testing Phase

The proposed method can be divided into four sequential phases:

**3.1** **Dataset** **Preparation**

For progressing this work a suitable dataset have to find and use by labeling that was not used by previous researchers who were not accrued desire of output. Therefore, this potion is divided into two parts named Data Collection and Data Filtering. In the data collection, the part around 26554 sentences or lines is collected by converting speech to text from Twitter and other social media. Then, into the data filtering part there 20485 train datasets and 6455 test datasets are found. So, the class distribution of the used dataset is given below,



|  |  |  |
| --- | --- | --- |
| Label | Train Dataset | Test Dataset |
| Stupidity | 10033 | 3202 |
| Pathetic | 3825 | 1210 |
| Greedy | 5412 | 1238 |
| Retarded | 1215 | 805 |
| Total | 20485 | 6455 |

Table 1: Class distribution in the Dataset

**3.2** **Pre**-**processing**

Pre-processing or Data pre-processing is a process that is prepared the raw data and make it suitable for Machine Learning models. So, in this paper, seven different approaches are used for this process that prepared suitable raw data such as lower case converter, bad symbol remover, emoticon remover, stop-word remover, punctuations remover, tokenization, and stemming for using Machine Learning model.

**3.3** **Feature** **Selection** **and** **Extraction**

After the Pre-processing, Text pre-processing, or Data pre-processing Feature Selection and Extraction step to come in some way that affects the whole result of this paper or research. Here, four different Feature Selection and Extraction are used named Part-of-speech (POS) tagging, uni/bi/tri-gram, term frequency-inverse document frequency (TF-IDF), and word2vec.

**3.4** **Classifier**/**Model**

To get a wonderful outcome for this research five different and suitable advanced from previous used classifier or model is used properly. The five different classifiers or models are Linear Support Vector Machine (Linear SVM), Long-Short Term Memory (LSTM), Bidirectional Long Short-Term Memory (Bi-LSTM), Decision Tree, and Logistic Regression. All approaches are the best and well know approaches in the Machine Learning domain.

4 Results

Afterward gone through the Method part, result part comes and plays a noticeable part to detect or this work because for this part the research can find its desired destination easily. So, after applying Linear Support Vector Machine (SVM), Long-Short Term Memory (LSTM), Bidirectional Long-Short Term Memory (Bi-LSTM), Decision Tree, and Logistic Regression into the implementing part, a good and noticeable number of accuracy or results come that are comparable. In the implementing part, both train and test datasets are also

used. Furthermore, precision, recall,

and F-1 score for each of the hate emotion classes or square as well as the total acquiring that are average accuracy is measured with its efficacy are given into this proposed paper in its method section. Here, the detected hate emotions categories are such as stupidity, pathetic, greedy, and retarded. It must say that this paper can consider those four types of hate emotions categories, all the list with their prediction and accuracy will be shown that how to find and computable outcome according to each or a sentence.

|  |  |
| --- | --- |
| Hate Emotions | Probability |
| Stupidity | 0.4898 |
| Pathetic | 0.1867 |
| Greedy | 0.2642 |
| Retarded | 0.0593 |

**Table 2:** Detailed evaluation using best model

Here the probability comes by using a rule that is,

P (Hate Emotion = Stupidity) = Total stupidity data / Total train data

The outcome of this paper using five different classifiers are;

|  |  |
| --- | --- |
| Classifiers | Accuracy |
| Linear SVM | 0.8618 |
| LSTM | 0.8209 |
| Bi-LSTM | 0.8707 |
| Decision Tree | 0.7223 |
| Logistic Regression | 0.8516 |

### Table 3: Classifiers

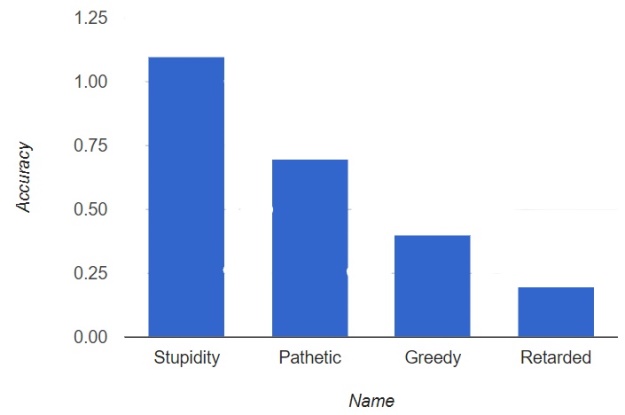
Whenever any sentence of the used dataset is considered, the sentence split is shown with its accuracy for each word of the sentence is like.

|  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- |
| it | aint | nothing | to | cut | a | bitch | off |
| 0.004 | 0.002 | 0.402 | 0.0023 | 0.37 | 0.003 | 0.66 | 0.0012 |

**Table 4:** Sentence split and accuracy

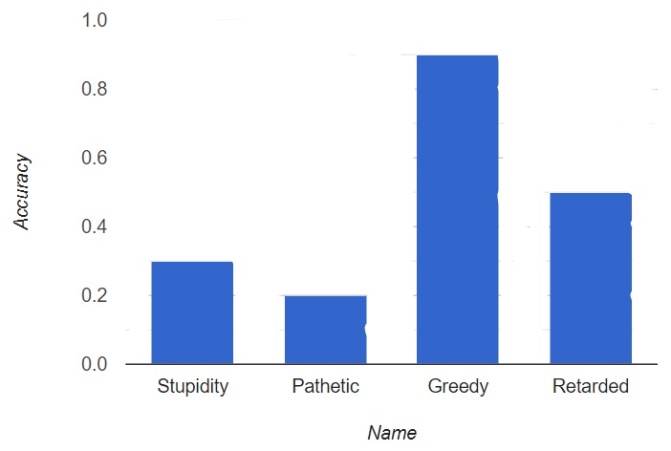


There, if any random sentence or comment is choosing for detected hate emotion or speech then it will be defined smoothly.





**Fig. 4:** Detect Hate emotions (Stupidity)



**Fig. 5:** Detect Hate emotions (Greedy)

The possible Accuracy =100 \* correct \_prediction/total sentences

5 Discussion

### After completing the methodology and result section, this research gives outstanding outcomes for all the searchers and this developing society. Here, this paper represents a wonderful result by using five different approaches with their features converting speech or voice to text. There the noticeable thing is that the four Hate emotions named Stupidity, Pathetic, Greedy, and Retarded are detected frequently who is the classification of Hate-speech. Using the Linear Support Vector Machine (SVM) gives 86.18% accuracy, the Long-Short Term Memory (LSTM) gives 82.09% accuracy, the Bidirectional Long-Short Term Memory (Bi-LSTM) gives 87.07% accuracy, the Decision Tree gives 72.23% accuracy, and the Logistic Regression gives 85.16% accuracy. Therefore, seeing that, accuracy the Bidirectional Long-Short Term Memory (Bi-LSTM) gives the best accuracy among them with the 20485-train dataset. It is indicated that according to using this huge dataset the Bidirectional Long-Short Term Memory (Bi-LSTM) performance is best and most preferable for this research.

6 Conclusion

In the perfecting time, to detect or identify the hate emotion and its classification from English speech, which is converted into text data for suitably uses, in the Machine Learning- based approaches namely the Linear Support Vector Machine (SVM), the Long-Short Term Memory (LSTM), the Bidirectional Long-Short Term Memory (Bi-LSTM), the Decision Tree, and the Logistic Regression algorithms are implemented frequently. After discovering the required or desirable corpus datasets, it gives 20485 comments, text or word from the training dataset and 6455 comments, text or words from the testing dataset. Stupidity (), Pathetic (), Greedy (), and Retarded () are the four detecting hate emotion categories from datasets. Every unwanted symbol (e.g., commas, dots, hyphens, emoticons, and so on) are separated during the text sectionalization period for using the dataset suitably. There, retarded has a dataset of only (). Afterward, the Command-Line string split function is used in this paper to tokenize all the words or characters in statements. Moreover, for finding hate speech from text with four hate emotion classes, precision, recall, and F1-score are utilized smoothly for achieving the desired and preferable accuracy. The obtained accuracy of the Linear Support Vector Machine (SVM) is 86.18 percent, the Long-Short Term Memory (LSTM) is 82.09 percent, the Bidirectional Long-Short Term Memory (Bi-LSTM) is 87.07 percent, the Decision Tree is 72.23 percent, and the Logistic Regression is 85.26 percent when the whole procedures are completed. Here, with 87.07 percent accuracy, the Bidirectional Long-Short Term Memory (Bi-LSTM) gives the best outcome.

REFERENCES

[1] J. M.-D. María Antonia Paz1, Hate Speech: A Systematized Review, Universidad International de La Rioja (UNIR), October-December 2020: 1–.

[2] I. I. D. F. Ashwin Geet D'Sa, Towards Non-Toxic Landscapes: Automatic Toxic Comment Detection Using DNN, F-54000 Nancy, France: University de Lorraine, CNRS, Inria, LORIA.

[3] A. M.-F. a. J. Farkas2, Racism, Hate Speech, and social media: A Systematic Review and Critique, Vol. 22(2) 205–224: Television & New Media, 2021.

[4] N. S.-B. L. M. T. V. H. M. H. L. Sergio Andr´es Casta˜no-Pulgarína, Internet, social media and online hate speech. A systematic review, International Journal of Communication, Received 14 July 2020; Received in revised form 26 January 2021; Accepted 23 March 2021.

[5] J. H. T. K. M. N. a. R. P. 6. Salla-Maaria Laaksonen1\*, The Datafication of Hate: Expectations and Challenges in Automated Hate Speech Monitoring, Jyväskylä, Finland: Department of Language and Communication Studies, University of Jyväskylä, February 2020 | Volume 3.

[6] Hate Speech Detection using Transformer Ensembles on the HASOC dataset? Szeged, Hungary: MTA-SZTE Research Group on Artificial Intelligence, 2019-02996.

[7] A. C. F. D. M. P. a. M. T. Fabio Del Vigna12, hate me, hate me not: Hate speech detection on Facebook, Istituto di Informatica e Telematica, CNR, Pisa, Italy., October 23, 2016.

[8] S. B. S. G. K. B. Y. Al-Onaizan, NeuralWord Decomposition Models for Abusive Language Detection, Amazon AWS, USA, August 2019.

[9] W. Y. a. A. Zubiaga, Towards generalizable hate speech detection: a review on obstacles and solutions, United Kingdom: School of Electronic Engineering and Computer Science, Queen Mary University of London, London, 17 June 2021.

[10] J. R. a. R. Krestel, Toxic Comment Detection in Online Discussions, Hasso Plattner Institute, University of Potsdam.