Hate Speech Recognition

Krishil Rana
Computer Science Engineering, School of Technology
Institute of Technology, Nirma University
Ahmedabad, India
20BCE136@nirmauni.ac.in

Madhav Kanakhara
Computer Science Engineering, School of Technology
Institute of Technology, Nirma University
Ahmedabad, India
20BCE142@nirmauni.ac.in

Manav Vakharia
Computer Science Engineering, School of Technology
Institute of Technology, Nirma University
Ahmedabad, India
20BCE149@nirmauni.ac.in

Meet Shingala
Computer Science Engineering, School of Technology
Institute of Technology, Nirma University
Ahmedabad, India
20BCE158@nirmauni.ac.in

Megh Patel
Computer Science Engineering, School of Technology
Institute of Technology, Nirma University
Ahmedabad, India
20BCE159@nirmauni.ac.in

***Abstract*—The exponential growth of social media has led to a surge in hate speech and online abuse, leading to hate crimes, orchestrated riots and large-scale riots. The task is mainly to categorize text content into non-hate and hate. We've broken it down into three categories: Hate Speech, Offensive, and Neutral. All efforts to detect hate speech must be carried out with great care so as not to violate an individual's right to freedom of expression. Hate involves a long-term loss of empathy and a constant desire to hurt the target. Combined, these words have broader meanings such as insult, discrimination, inhumane treatment, demonization, and incitement to violence. Text classification organizes different types of text data into defined clusters or categories. The primary purpose of classifiers is to assign relevant documents to appropriate categories.**

***Index Terms*—Hate Speech Recognition, Text Classification**

I. Introduction

The exponential growth of Social Media; Major Players being Instagram, TikTok, Twitter etc. has resulted for them to be breeding of Hate Speech and Online Abuse resulting into Hate Crimes, Orchestrated Riots and Mass Dis-rest. Hate Speech is defined as " any kind of communication in speech that attacks or uses pejorative and discriminatory language with reference to a person or a group on the basis of who they are". The urgency of this matter is being released internationally and initiatives have been launched to assess the problem and counter-measures to it. The first step is to identify and track hate speech online. However, the process is very labour intensive, time consuming and not sustainable or scalable in reality. The task mainly involves classifying textual content into non-hateful and hateful. We have classified into three categories namely Hate Speech, Offensive and Neither. There is a tremendous amount of text data being produced every day in different parts of the world. If the information is not categorised, it will be quite difficult to manage this lengthy text documentary. Text data may be handled and used considerably more easily if they are arranged into similar categories. The term "text classification" refers to the process of placing text documents into a certain, pre-established category.

A. What is Hate Speech?

The definitions of literature identify various forms of hate speech, some of which may be more detrimental to its targets than others. In order to respect people's rights to free speech and to guard against abuse by those who would use it for improper purposes, any effort to identify hate speech must be carried out with extreme caution. We must comprehend the two key words in order to recognise. Hatred is a human emotion that is brought on by exposure to specific kinds of information. Hatred entails a pervasive dislike, a loss of empathy, and a constant desire to damage the person at whom the speech is directed. It is commonly accepted that various groups have a particular selection of members who might be recognised by their subsequent group identity:

- Citizenship
- Religious Practices
- Ethnic Groups
- Race
- Gender and Sex
- Age
- Sexual Orientation

Speech refers to any form of communication that proceeds through a variety of media, including words or other utterances, photos, texts, videos, and gestures. When the terms are used together, they often have a considerably broader meaning that encompasses instigation to violence, dehumanisation, demonization, and other negative concepts. Contrary to common assumption, the use of social media can persuade one's view

using their propaganda, but most individuals do not change their ideas on problems involving their essential values and beliefs that are already set.

II. The pipeline of text classification

Each classifier follows a certain set of criteria while classifying texts. A general method exists for classifying text data into predetermined groups. There were two distinct steps in the text classification process:

1) Training Phase:
   Each document undergoes training to fit within a particular, predetermined category. These categories' definitions are based on the information they include. It gives the unlabeled documents the appropriate category
2) Testing phase:
   In this stage, the features that were extracted to categorise the documents from the dataset that weren't used in the training phase are tested.

Simply described, text categorization is the process of grouping different kinds of text data into pre-established clusters or categories. The assignment of the relevant text material to the proper category is the main objective of a classifier. Every supervised machine learning issue requires an initial dataset for classification. Although a document may fall under more than one category, we are simply assigning each one a single category for the purposes of this study. It takes some time to divide text data into different groups. The text classification process consists of the following steps, which are listed below:

1) Documents Collection:
   Documents Collection: Text data is collected at the start of the process in a variety of formats, such as.doc and.txt files, html tags, website content, etc.
2) Pre-processing:
   Pre-processing: The process then involves converting each format to a word document when the papers have been successfully collected. Pre-processing steps include, among others, the following:
   - Tokenization:
     Tokenization is the conversion of each word in the document into a unique token. It breaks a sentence up into many tokens. Stop-words Prepositions and other stop words are dropped, including connectives. Take the following examples: "is," "am," "are," "the," "and," "but," "for," and "if."
   - Stemming:
     Stemming is the process of turning many words into their roots. An illustration might be: Examine au Exam.
3) Indexing:
   To diminish the intricacy of the archives, documentation portrayal is the one of the significant procedures of prehandling where the full text rendition of the report is changed to record vector. The most useful model for the portrayal of words is the vector space model.
4) Feature Selection:
   After pre-processing and ordering the significant composed literary example, the choice to join vector space is featured, working on the versatility, execution, and exactness of a substance classifier. The essential objective of element determination is to pick a subset of files' abilities.
5) Classification:
   Following element extraction, text documents were classified prior to being given on. To naturally order text reports into a specific class, an assortment of characterization methods are utilized, including the Innocent Bayes classifier, the SVM (Support Vector Machine) classifier, Neural Networks, decision trees, and so on.
6) Performance measure:
   The exactness of the more tasteful, F-measure, and accuracy of the more tasteful are utilized to dissect the running exhibition of the more tasteful.
   - Precision = $$\frac{TP}{(TP+FP)}$$
   - Recall = $$\frac{TP}{TP+FN}$$
   - F-measure = $$\frac{2*(Precision*Recall)}{(Precision+Recall)}$$
   - Accuracy = $$\frac{(TP+TN)}{(TP+TN+FP+FN)}$$

III. Results

A. Multinomial Naive Bayes Classifier

Naive Bayes is a probabilistic classifier that uses a group of calculations that all offer a typical characterization standard as opposed to a solitary calculation. This classifier produces separate extricated qualities from one another. The advantage of utilizing this classifier is that it performs well on both text based and numeric information, and it is additionally less complex to develop. This classifier's disadvantage is that it performs more awful when the removed elements are associated with one another. Popular in Normal Language Handling is the Bayesian learning strategy known as the Multinomial Credulous Bayes calculation (NLP). Utilizing the Bayes standard, the PC makes an informed expectation about the tag of a text, for example, an email or news story. It decides the probability of each tag for a specific example and results the tag with the most elevated probability

B. Support Vector Machine

A supervised learning technique called SVM can be applied to both classification and regression problems. Usually, it serves as a classifier. It is used in many different NLP text categorization jobs. Here, each piece of data is shown as an n-dimensional space, where n is the total number of extracted

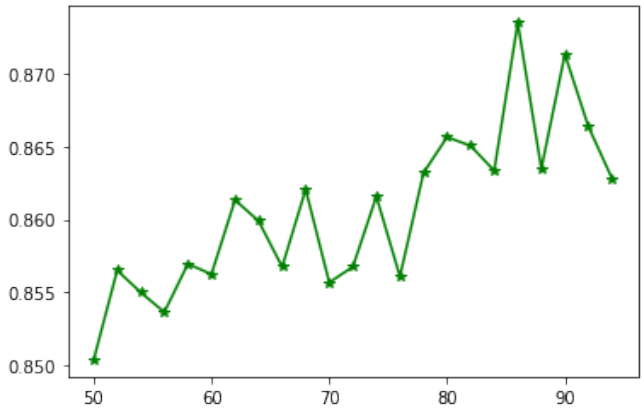

Fig. 1. Plot of Accuracy vs Training Dataset in %

features. The support for both positive and negative training sets is what makes SVM unique. The number of different keywords in each text document is represented as a vector with dimensions equal to that number. However, if the text document is large, there will be several dimensions in hyper-space, which could increase the computational expense of the process.

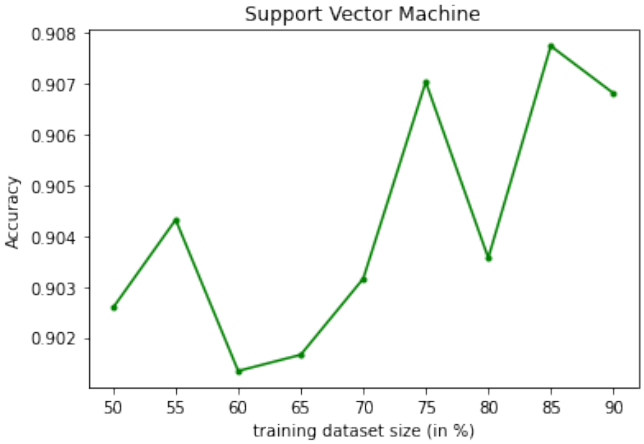

Fig. 2. Plot of Accuracy vs Training Dataset in %

C. Decision Tree Classifier

The decision tree is another classifier technique that is widely employed for classification. It operates in accordance with a set of requirements and test questions. It is modelled as a tree, with the leaves representing different ”classes” and the branches representing ”weight.” Decision-tree is effective at learning disjunction expressions and can handle noisy data. A decision-tree development process, however, might be pricey. The entire sub-tree will be defective and the entire structure may be deemed invalid if there is an issue at a very high level.

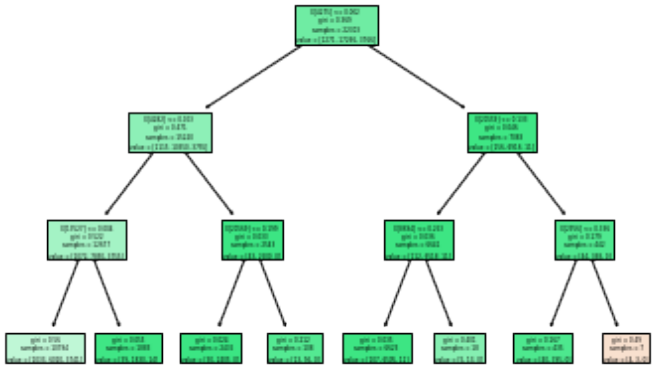

Fig. 3. Structure of Decision Tree (Gini Index)

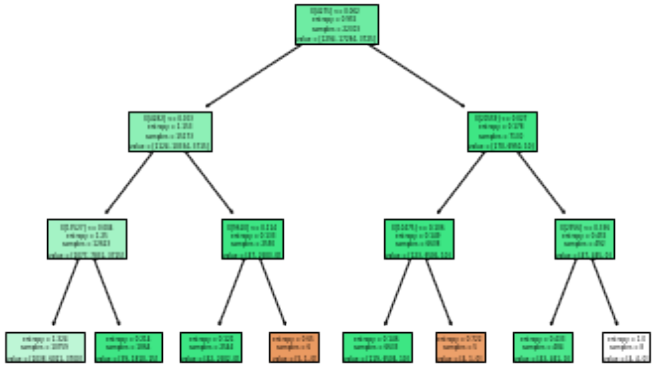

Fig. 4. Structure of Decision Tree (Entropy)

Fig. 5. Plot of Accuracy vs Training Dataset in %

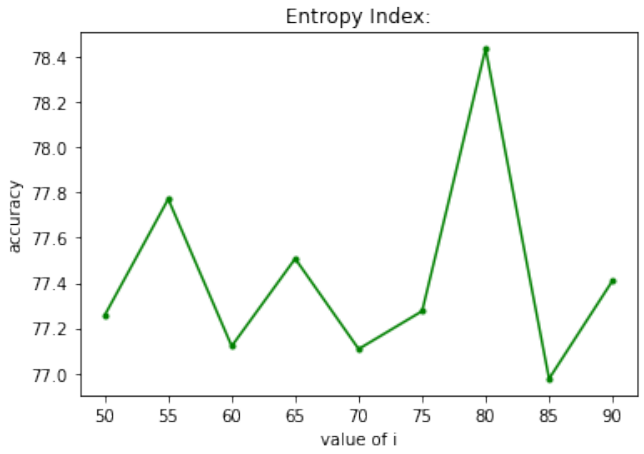

Fig. 6. Plot of Accuracy vs Training Dataset in %

D. K-Nearest Neighbour

KNN is among the least difficult and most clear AI calculations. KNN groups text archives by estimating the distance among them and their neighbors; neighbors with equivalent classes are probably going to have a place with that class. In any case, KNN is a lazy learning calculation, and it is genuinely difficult to sort out example closeness.

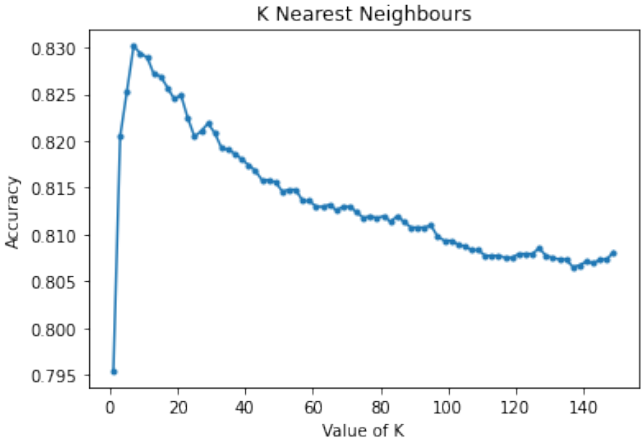

Fig. 7. Plot of Accuracy vs Training Dataset in %

E. Multilayer perceptron classifier

MLP is the contraction for multi-facet insight. It is comprised of thick, totally associated layers that might change any information aspect into the ideal aspect. A brain network with various layers is alluded to as a multi-facet insight. To construct a brain organization, we consolidate neurons so a portion of their results are likewise their bits of feedbacks.

IV. DISCUSSION

Throughout the duration of running our models for the dataset we used which is used for text classification and classifying the dataset into three classes which are hate speech

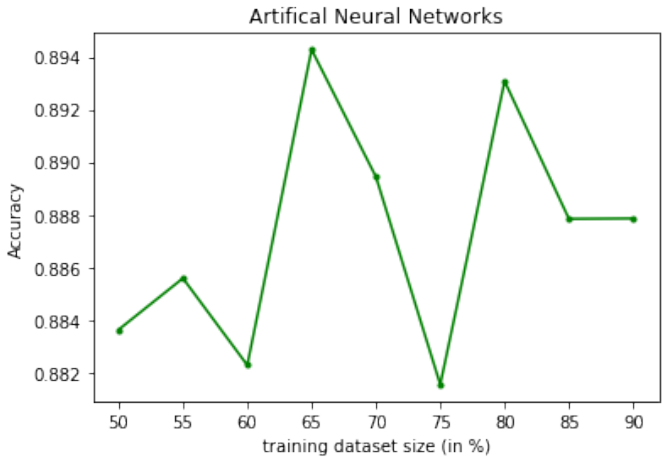

Fig. 8. Plot of Accuracy vs. Training Dataset in %

Models	Accuracy in % (without GridSearch)	Accuracy in % (with GridSearch)
Multinomial Naive Bayes	80.56	83.96 (alpha = 0.1)
Support Vector Machine	90.77	91.11 (C= 1000, Gamma = 0.001, kernel = rbf)
Artifical Neural Networks	89.42	88.017 (Activation = 'logistic', Hidden_Layer_Sizes = (6,) Solver = 'lbfgs')
K Nearest Neighbours	83.01	82.94 (K = 5)
Decision Trees (Gini Index)	78.45	89.28 (criterion = 'entropy')
Decision Trees (Entropy)	78.43	

Fig. 9. Results

(indexed - 0), offensive language (indexed - 1) and neither of them (indexed - 2). We used multiple models and here are our observations. The models which we implemented are Multinomial Naive Bayes, Support Vector Machine, Decision Trees, K nearest neighbour and Artificial Neural Networks (used MLPClassifier of sklearn). We ran the models for our dataset. Some of our models took quite some time. We did two approaches: One is without grid search wherein we plotted a graph of accuracy vs. training size% and the one with gridsearch wherein we gave in different values of various parameters into the gridsearch and ran it. The program with gridsearch took a long time, that is hours. The program without gridsearch was put under observation and there was an increase in the accuracy of model prediction with the increase of training size of dataset that is train data%. We observed a peak when training data is set around 80%. Amongst the various Models, we observed that SVM started giving the maximum accuracy with 90.89%. The reason is that SVM has overfit protection which is it doesn't depend on the number of features and that is why it performs so well.

V. CONCLUSION

Identifying defamatory rhetoric can be difficult as it requires processing content and evaluating specific situations. Information collections containing derogatory discourse are usually not in their original state, so pre-processing is essential

before detecting derogatory discourse in the characterization computation. The benefits of AI models vary, and they outperform other AI models in certain efforts, such as discriminating discourse. Some models are more capable, others are more accurate. To find the ideal model for discovering the discourse of disdain, we need to look at several different model exhibits. Recently, preparation strategies have gained prominence, so it is important to examine whether they complement algorithms for identifying disparaging discourse. For our dataset, Support Vector Machine is the best model, giving an expected accuracy of 91%. About 88% of the data is used for training and the rest for testing. After hyper boundary tuning and model tuning, we found that SVM gives the best results, but takes the longest.

VI. Scope for future Work

Future work can zero in on fostering a more dependable disdain discourse order engineering to address troubles with discontinuity, versatility, class irregularity information, and conventional metadata plan. It is feasible to do research to foster a perception screen for classifying disdain discourse feeling. Later on, directed learning might be utilized to gauge the feeling of another tweet utilizing dataset with opinion marks. At long last, greater opinion classes can be obliged by extending the fluffy rationale phonetic factors in the laid out conventional metadata engineering for further developed order.

VII. Abbreviations and Acronyms

- SVM - Support Vector Machines
- NLP - Natural Language Processing
- KNN - K-Nearest Neighbours
- MNB - Multinomial Naive Bayes
- TP - True Positive
- TN - True Negative
- FP - False Positive
- FN - False Negative

VIII. Acknowledgment

We would like to thank our peers, faculties,families and our organization to provide this opportunity and the constant support.

References

1) https://www.un.org/en/hate-speech/understanding-hate-speech/what-is-hate-speech
2) Hate Speech Detection: A Solved Problem? The Challenging Case of Long Tail on Twitter arVix Link
3) Rudnicki, Konrad & Steiger, Stefan. (2020). Online hate speech - Introduction into motivational causes, effects and regulatory contexts.
4) Femi Emmanuel Ayo, Olusegun Folorunso, Friday Thomas Ibharalu, Idowu Ademola Osinuga, Machine learning techniques for hate speech classification of twitter data: State-of-the-art, future challenges and research directions,Computer Science Review
5) Brett Drury, Samuel Morais Drury, Md Arafatur Rahman, Ihsan Ullah, A social network of crime: A review of the use of social networks for crime and the detection of crime, Online Social Networks and Media
6) N. A. Setyadi, M. Nasrun and C. Setianingsih, "Text Analysis For Hate Speech Detection Using Backpropagation Neural Network," 2018 International Conference on Control, Electronics, Renewable Energy and Communications (ICCEREC), 2018, pp. 159-165, doi: 10.1109/ICCEREC.2018.8712109.