Hate Speech Detection

Sonam Singh[1], Shivani Shinde[2], Subodh Nikumbh[3], Prof. Dhiraj Amin[4]
[1.2.3]*Member, Pillai College of Engineering, New Panvel, Maharashtra – 410206, India*
[4]*Guide, Pillai College of Engineering, New Panvel, Maharashtra – 410206, India*

***Abstract*- As online content continues to grow, so does the spread of hate speech. We identify and examine challenges faced by online automatic approaches for hate speech detection in text. The increasing propagation on social media and the urgent need for effective countermeasures have drawn significant investment from governments, companies, and researchers. Machine Learning and predictive analytics now help companies to focus on important areas, anticipating problems before they happen, reducing costs, and increasing revenue.**

***Index terms*- — Hate Speech, Classification, NLP, SVM, Naive Bayes, Bag of Words (BOW).**

I.INTRODUCTION

What constitutes hate speech and when does it differ from offensive language? No formal definition exists but there is a consensus that it is speech that targets disadvantaged social groups in a manner that is potentially harmful to them. Hate speech is a particular form of offensive language that makes use of stereotypes to express an ideology of hate. People have an increased willingness to express their opinions online thus contributing to the propagation of hate speech. Research on safety and security in social media has grown substantially in the last decade. A particularly relevant aspect of this work is detecting and preventing the use of various forms of abusive language in blogs, micro-blogs, and social networks. A number of recent studies have been published on this issue such as the work by Xu et al.(2012) on identifying cyber-bullying, the detection of hate speech (Burnap and Williams, 2015) which was the topic of a recent survey (Schmidt and Wiegand, 2017), and the detection of racism (Tulkens et al., 2016) in user generated content. Hate speech (Schmidt and Wiegand, 2017) is an act of offending, insulting or threatening a person or a group of similar people on the basis of religion, race, caste, sexual orientation, gender or belongingness to a specific stereotyped community. Abusive speech categorically differs from hate speech because of its casual motive to hurt using general slurs composed of demeaning words. Both abusive as well as hate speech are sub-categories of offensive speech.

Twitter is also actively in an ongoing process to enforce new guidelines related to how it handles hateful conduct and abusive behavior, by users, taking place on its platform. In addition to threatening violence or physical harm, they also want to look for accounts affiliated with certain respective groups that promote violence against citizens to move further in their hateful intentions.

II. LITERATURE SURVEY

In this the relevant techniques in literature is reviewed. It describes various techniques used in the work. Identify the current literature on related domain problem. Identify the techniques that have been developed and present the various advantages and limitation of these methods used extensively in literature.

A literature review is an objective, critical summary of published research literature relevant to a topic under consideration for research. Its purpose is to create familiarity with current thinking and research on a particular topic, and may justify future research into a previously overlooked or understudied area.

In this paper they have addressed the problem of hate speech detection in online user comments showing the use of paragraph2vec in a two-step method with the help of continuous bag of words (CBOW) neutral language model. After using this approach they found out that by using paragraph2vec, some non-obvious swear words were also detected and this approach obtained higher Area under the curve (AUC) than other BOW models.

Another paper that we research suggested that classification of hateful speech on the basis of degree of hateful intent. They pre-processed the dataset by removing hash tags, URLs and Usernames. IN processing, they used three methods with random forest classifier giving the best result (76.42%) followed by Naive Bayes (73.42%) and lastly SVM with linear function kernel (71.71%). They have concluded saying that the paper focuses on classification and detection of harmful speech with a large future scope.

In this paper they have addressed the classification of the dataset in three categories: Hate speech, offensive language and neither. They tested five types of models out of which Linear SVM and Logistic regression gave best results. In conclusion they have said that there is a difference between hate speech and offensive language which can hinder or help the accurate classification in relation to its context.

Another paper that we research suggested that they examined methods to detect hate speech and differentiate it from general profanity, they have categorised dataset into three categories: Hate speech, offensive and OK. Showing that with the increase in training data, the accuracy increases in SVM. They used two surface features viz. Surface n-grams and word skip-grams, out of which the best was obtained by character 4-gram model achieving 78% accuracy and having multiple future directions to their given research.

In this paper they have addressed that they have detected hate speech and death threats of violence using machine learning by using BOW on a dataset of YouTube comments. They have specially chosen two words which for text mining which improved the prediction result by 5%. The problem they faced was due to use of slangs, bad grammar; the prediction result was affected in a negative way.

Another research paper that we research suggested that they have used Naive Bayes Algorithm for detecting hate speech on twitter. They first collected the database then preprocessed the dataset. Later they assigned two classes to each tweet, one describing the sentiment of the tweet and one describing the subject matter discussed in the tweet, which later detected if the tweet was positive or negative. They also used scikit-learn using python language. They got 80% accurate results and faced two problems i.e. They could not include video and emoticons and they did not have sufficient dataset due to the working of twitter.

In this paper they have addressed that a dataset is formed for Italian texts using data crawling technique in which the comments from Facebook public pages like artist or groups is taken. There are two classifiers used which are based on different algorithms one is based on SVM (Support vector machines) and the other based on neural network named as LSTM (Long short term memory). Two different classification experiments were conducted, the first one considering three different categories of hate and the second one considering two different categories of hate.

Another paper that we research suggested that three CNN based models are used to classify sexist and racist abusive language. Two English datasets that are waseem and hovy 2016 and Waseem 2016 are used which contains tweets withs sexist and racist comments. The two datasets were concatenated into a single dataset and then divided into three datasets for one and two step classification. One step dataset is segmentation for multi class classification. For two step classification the sexismm and racism labels were merged together. They created one more dataset to experiment a second classifier to distinguish between sexist and racist comments. All the results obtained are average results.

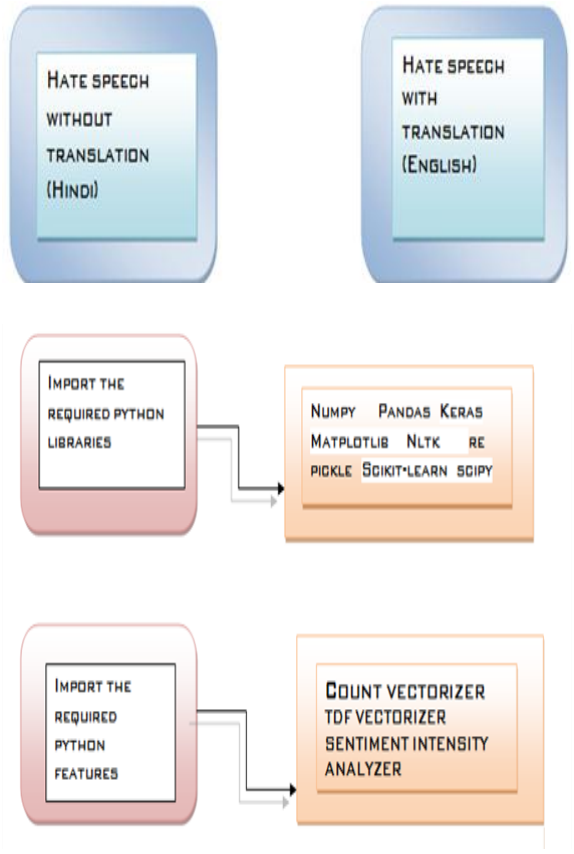
In this research paper they have addressed that for this system, they used a supervised classification which uses NLP features that measures different aspects of user comments. They have specifically used Vowpal Wabbit's regression model with a standard rate of 28. The features are divided into four classes N grams, linguistic, Syntactic and Distributional Semantics. First three classes are used for mild preprocessing to transform some of the noise into data. It needs a large amount of training data to fare with the current hate detection system.

Another research paper that we research suggested an effective approach to detect cyber bullying messages from social media through a SVM classifier algorithm.. Web links are being filtered using a page ranking algorithm. They concluded by saying that comments are classified into positive or negative, if positive comments display on the site or if negative comments are deleted from the site. Overall with help of SVM classifier and predator this project achieves 87% accuracy. The future work of this paper analyzes

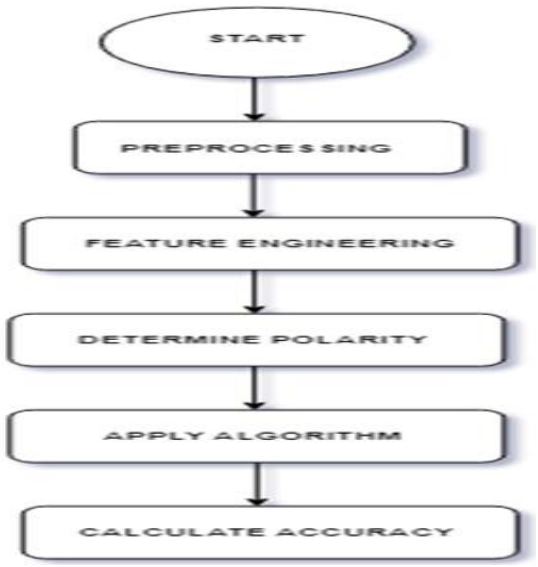
the video comments and detects the positive video or negative video. If in case negative video is avoided from social media.

III. PROPOSED SYSTEM

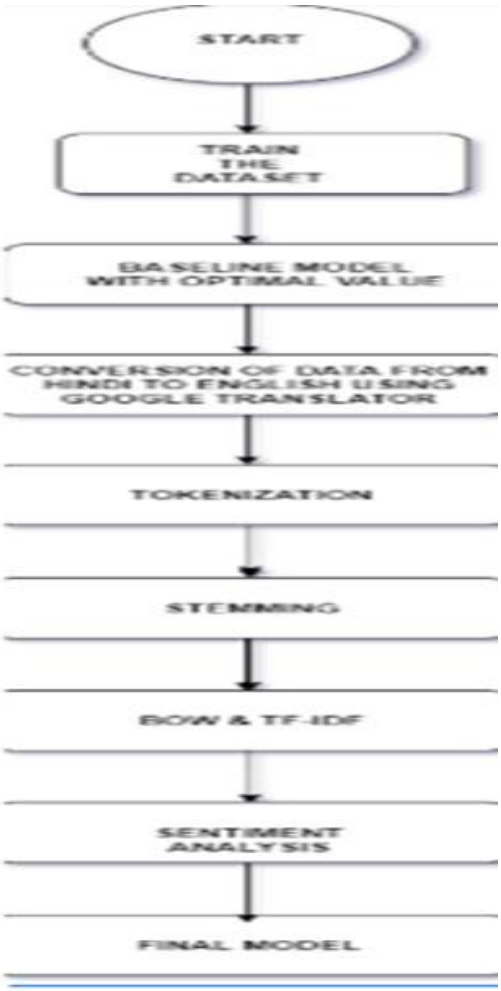
In this, we analyze the problem of hate speech detection in code-mixed texts and present a Hindi-English code-mixed dataset. We also propose a supervised classification system for detecting hate speech in the text using various character levels, word level. Due to their large-scale availability. However, in multilingual societies like India, usage of code-mixed languages (among which Hindi-English is most prominent) is quite common for conveying opinions online. We have divided the sections into two different categories viz. Hate speech without translation and Hate speech with translation.



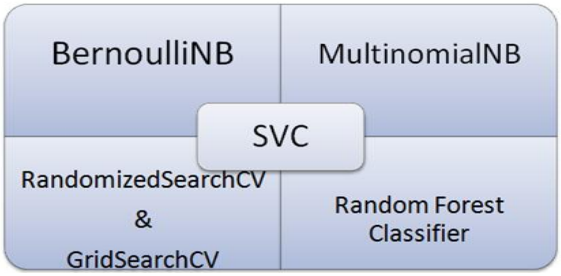
IV. FLOWCHART FOR HATE SPEECH WITHOUT TRANSLATION



V. FLOWCHART FOR HATE SPEECH WITH TRANSLATION



VI. ALGORITHM USED



VII. HARDWARE AND SOFTWARE SPECIFICATIONS

The experiment setup is carried out on a computer system which has the different hardware and software specifications as given in Table 3.2 and Table 3.3 respectively.

Processor	2 GHz Intel
HDD	180 GB
RAM	2 GB

Hardware details

Operating System	Windows 10
Programming Language	Python
Database	MySql

Software details

REFERENCES

[1] Hate speech detection using comment embeddings. Nemanja Djuric, Jing Zhou, Robin Morris, Mihajlo Grbovic, Vladan Radosavljevic, Narayan Bhamidipati

[2] Degree based Classification of Harmful Speech using Twitter Data SanjanSharma LTRC, KCIS IIIT Hyderabad, India

[3] Automated Hate Speech Detection and the Problem of Offensive Language Thomas Davidson,1 Dana Warmsley,2 Michael Macy,1,3 Ingmar Weber

[4] Detecting Hate Speech in Social Media Shervin Malmasi Marcos Zampieri

[5] Automatic detection of hateful comments in online discussion Hugo Lewi Hammer

[6] Using Naïve Bayes Algorithm in detection of Hate Tweets.Kelvin Kiema Kiilu, George Okeyo, Richard Rimiru, Kennedy Ogada

[7] Hate me, hate me not: Hate speech detection on Facebook Fabio Del Vigna12, Andrea Cimino 23, Felice Dell'Orletta 3 ,Marinella Petrocchi 1, and Maurizio Tesconi 1

[8] One-step and Two-step Classification for Abusive Language Detection on Twitter Ji Ho Park and Pascale Fung

[9] VADER: A Parsimonious Rule-based Model for Sentiment Analysis of Social Media Text C.J. Hutto Eric Gilbert

[10] Abusive Language Detection in Online User Content Chikashi Nobata Joel Tetreault Achint Thomas Yi Chang

[11] Classification of Hate Tweets and Their Reasons using SVM Natalia Tarasova