DETECTION AND CLASSIFICATION OF HATE SPEECH

Girish V P, Namratha Bhat, Bhavani B S, Abhin K V, Mrs. Sahana V
ISE, JSSATE-B

Abstract— **The challenges that are to be faced while handling with hate speech is not a new thing. From thepast few years due to the boosted usage of internet, hateful activities across social media is increasing rapidly. Improved technology has made it possible to create a platform where people can feel free to share their opinions and experiences.it wouldn't be a problem if this is just the case. but we can also see hateful comments running throughout the social media targeting a person or a community. Hate speech is the statement that targets a person or community of people discriminating based on caste, creed, nationality etc.**

Our project aims at resolving the above problem by using Machine Learning techniques to automatically detect hate speech and classify them into various classes such as extremely positive, positive neutral etc. We have used classifier that works based on the lexicons and finally compare it with other classifiers that doesn't use lexicons. Aimed beneficiaries of this model are the people who are being targeted on social media. Based on the results they can calculate intensity of the comments.

Keywords—: Machine learning Techniques, Rule-based approach.

I. INTRODUCTION

Hate Speech manages content that advances viciousness against or has the main role of prompting disdain against people or gatherings dependent on specific traits, for example, race or ethnic inception, religion, inability, sex, age, veteran status, sexual direction/sex personality. Despise discourse covers numerous types of articulations which spread, induce, advance or legitimize contempt, savagery and oppression an individual or gathering of people for an assortment of reasons. It presents grave perils for the attachment of a vote-based society, the insurance of human rights and the standard of law. Whenever left unaddressed, it can prompt demonstrations of brutality and strife on a more extensive scale.

In this sense abhor discourse is an extraordinarytype of bigotry which adds to detect wrong doing. The manners by which focused networks experience despise discourse is a significant, however frequently ignored, segment of the discussion over the authenticity of abhor discourse laws. 'Detest discourse' is a term broadly utilized, yet deficient with regards to a solitaryimportance. Despise is the same old thing and lamentably is ubiquitous in our general public, subsequently the fight against online detest discourse and loathe all in all gets a great deal of consideration. While numerous sciences study theimpact of web on our conduct and abhor discourse, administrators attempt to get up to speed by making structures comprising of improved guidelines and various ways to deal with compel the spread of detest. In any case, while the goals are acceptable, the genuine center of the conversation is lost enroute the issues concerning the idea, the need andexplanation behind an additional structure for online loathe discourse, and the troubles of this new sort of interchanges.

Natural language processing is of two types:

- Rule based techniques.
- Machine Learning techniques.

- Rule based Techniques.

Rule based assumption investigation alludes to examine directed by the language specialists. Rule based is otherwise called lexical or slant vocabulary. This is investigation of set of rules where in which the words are arranged are either positive or negative alongside their relating power measure.

a) Text Blob

Text Blob is a well-known Python library for handling literary information. It is based on NLTK, another famous Natural Language Processing tool compartment for Python. Text Blob utilizes an assumption vocabulary (comprising of predefined words) to relegate scores for each word, which are then arrived at the midpoint of out utilizing a weighted normal to give a general sentence notion score.

a) Vader

Valence Aware Dictionary and sentiment Reasoner is another mainstream rule-based library for notion investigation. Like Text Blob, it utilizes a supposition vocabulary that contains power measures for each word dependent on human-commented on names.A key contrast in any case, is that VADER was structured with an emphasis via web-based networking media writings. This implies it puts a ton of accentuation on decides that catch the pith of content normally observed via web-based networking media.

- Machine Learning Techniques.

ML-based approaches are the ones that don't depend on the manually created rules instead, they use machine learning techniques. This method relies on the algorithms that are capable of learning without the explicit programming. It considers several factors other than input and also handles different types of variables which are independent of each other. This method sometimes treats the algorithm as black box. Machinelearning models play a very important role in tasks such as Classification, Clustering, Natural Language Processing Etc. Using machine learning for natural language processing requires large dataset in order to give accurate results.

a) Logistic Regression

Moving ahead from rule-based methodologies, the following technique endeavored is a strategic relapse among the most regularly utilized managed learning calculations for characterization. Strategic relapse is a straight model prepared on named information — the term direct is significant on the grounds that it implies the calculation just uses straight blends (for example entireties and not results) of data sources and parameters to create a class forecast.

b) Support Vector Machine

Support Vector Machines are fundamentally the same as strategic relapse as far as how they enhance a misfortune capacity to produce a choice limit between information focuses. The essential distinction, be that as it may, is the utilization of "portion capacities", for example capacities that change a mind boggling, nonlinear choice space to one that has higher dimensionality, with the goal that a fitting hyperplane isolating the information focuses can be found. The SVM classifier hopes to expand the separation of everydatum point from this hyperplane utilizing "bolster vectors" that portray each separation as a vector.

II. OBJECTIVES

- Lexicon based approaches are used to predict the probabilities of the results
- Class labels or Intensities are used for Classification.
- Finally, the results of different classifiers are compared to find the best one.

III. METHODOLOGY

Analysis of the sentiment is the method of extracting knowledge from a user's opinion. Each individual shares their opinion, their individual interests and data on social media like Facebook, Instagram, twitter etc. and we get connected with many other people's thoughts. It is the NLTK process, in other words Natural Language processing task. Sentiment analysis relates to identifying similarities in data and inferring the mood of the provided piece of knowledge that could be grouped into one of these classifications:

1) Negative.
2) Neutral.
3) Positive.

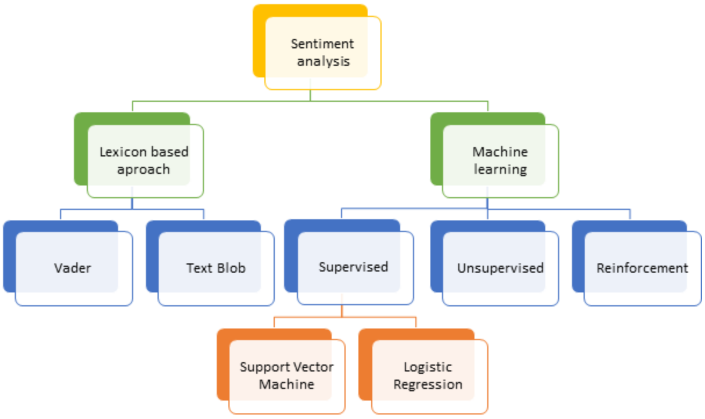


Fig 1. Sentimental Analysis Technique.

- **Lexicon Based Approach**

Development of a lexicon has been one of the two key methods to the study of feelings and includes measuring the meaning from the linguistic direction of the term or sentences in a document. In general terms a fragment of text messaging is described as a bag of words in lexicon- based techniques. A mixing feature, such as total or average, is implemented to render the final calculation concerning the message's overall sentiment. With the exception of a meaning of emotion, typically the dimension of a word's local context is taken into account, such as negation or entrenchment.

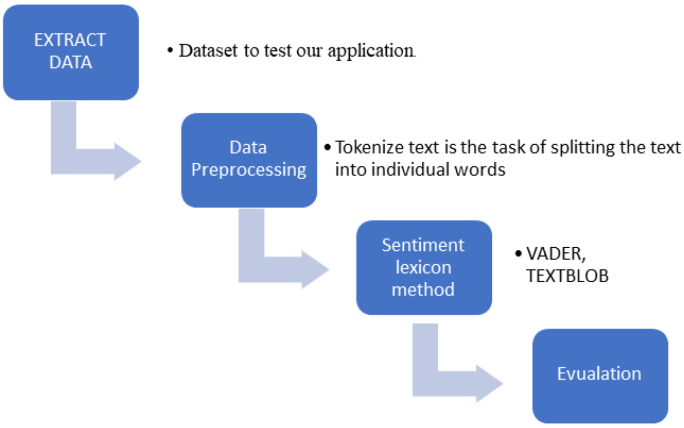

EXTRACT DATA
• Dataset to test our application.
Data Preprocessing
• Tokenize text is the task of splitting the text into individual words
Sentiment lexicon method
• VADER, TEXTBLOB
Evualation

Fig.2 Block diagram for lexicon-based approach.

- Extract Data - A collection of data from (usually unstructured or improperly structured) sources of data for more data analysis or data storage (data migration) or operation. Thus, the input into the intermediate extraction method is typically accompanied by data translation and probably the inclusion of metadata in the application workflow before export to another level.

- Data pre-processing - Cleaning the data byremoving stop words punctuation, html tags andother unwanted and meaningless words shows example of data pre-processing.

- Sentiment lexicon - The built emotion lexicon comprises around 6300 terms. It was automatically converted with the SentiWordNet program as a baseline. -- term in the lexicon is given a meaning from −100 (most negative) to 100 (most positive), which reflects emotion. In this project, we have used two types of sentiment lexicon that is Vader and Text blob.

- Evaluation - After the successful testing ofthe model, it is time for it to be deployed and it starts making predictions. It's time to evaluate the sentiment lexicon model for prediction.

- **Machine Learning Method**

Machine learning strategies for classification of sentiment are increasing in popularity due to their potential to model several features and in doing so, capture meaning, their simpler adaptability to adjust data, and the likelihood of calculating the degree of ambiguity from which a classification is produced. A subset of Artificial Intelligence is Machine Learning. This area applies to teaching robots in, over time, executing such activities that they get stronger at.

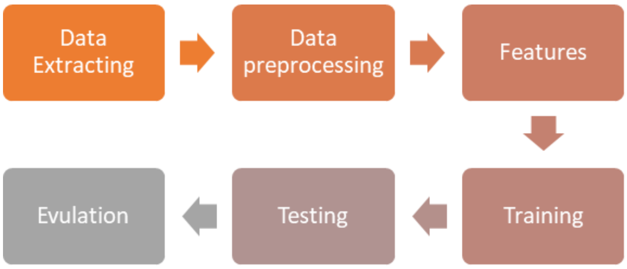


Fig 3. Steps for Text classification.

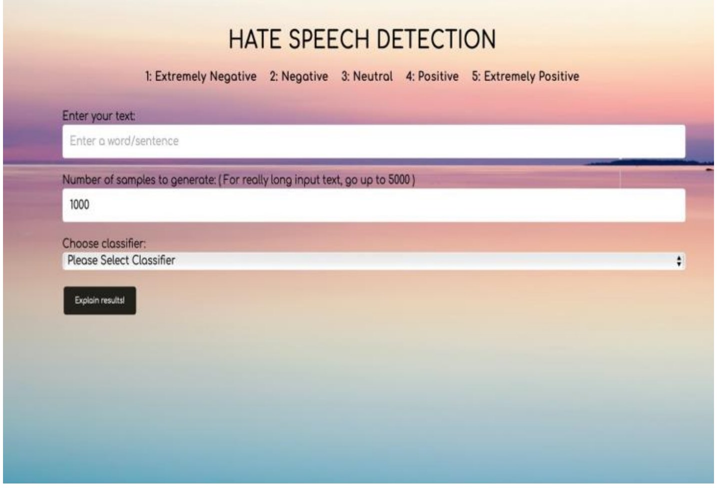
- Data Extraction - Getting the data corpus that will be used to train the model i.e., the classifier and test it.
- Data pre-processing- Cleaning the data byremoving stop words punctuation, html tags and other unwanted and meaningless words.
- Features - Converting the terminology to its numerical expression this is also known as vectorizing. Here, in the work we use count vectorizer and TF-IDF vectorizer. count vectorizer is used to transform a text corpus intoa term / token count vector. It also offers the option to pre-process the text data before creating the vector representation making it a highly scalable framework for text representation of applications. TF_IDF (Term Frequency– Inverse Document Frequency)this tokenizes documents, learns new vocabulary and reverse document frequencyweightings, and allows new documents tobe encoded.
- Training - The data here is split into training andtesting sets. The ML classifiers are trained by the training dataset by providing both features and labels as input. The efficiency of the given classifier also depends on how the vectors in question werefirst created i.e., in the previous step. In this work, wewill be using Logistic regression and support vector machines.
- Testing - After the training is completed, the testing data set is tested. the data set is divided into 80:20that is 80% for the training set and 20% for testing depending upon the requirement.
- Evaluation - Once the testing dataset is successfully tested the model, is then deployed and it starts tomake predictions when the input is served.

IV. REQUIREMENTSPACKAGES

- pytreebank==0.2.6
- tqdm==4.33.0
- cython==0.29.3
- pandas==0.25.0
- nltk==3.4.5
- textblob==0.15.3
- lime==0.1.1.36
- spacy==2.1.8

V. RESULTS

Using Flask along with HTML and CSS, webpage is created. Webpage indicates a text box where; a sentence or word is entered and a sample is set for generating texts and later we need to choose one classifier among the four classifiers for obtaining the particular results of that classifier.



HATE SPEECH DETECTION

1: Extremely Negative 2: Negative 3: Neutral 4: Positive 5: Extremely Positive

Enter your text:

Enter a word/sentence

Number of samples to generate: (For really long input text, go up to 5000)

1000

Choose classifier:

Please Select Classifier

Fig 4. Webpage Details

- **Accuracy & F1 score for all four classifiers**

Classifier	Accuracy	F1score
LR	40.6	35.3
SVM	41.40	38.2
Textblob	28.3	24.6
Vader	31.5	31.2

- **Confusion Matrix for all four classifiers**

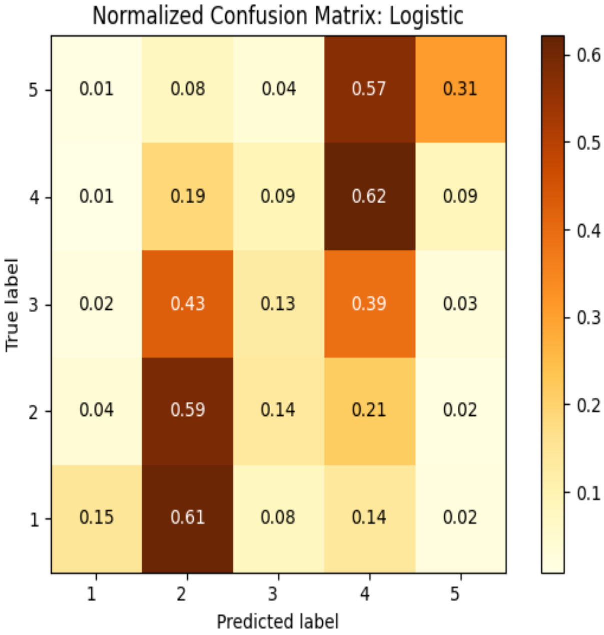


Fig 5. Confusion Matrix for Logistic Regression

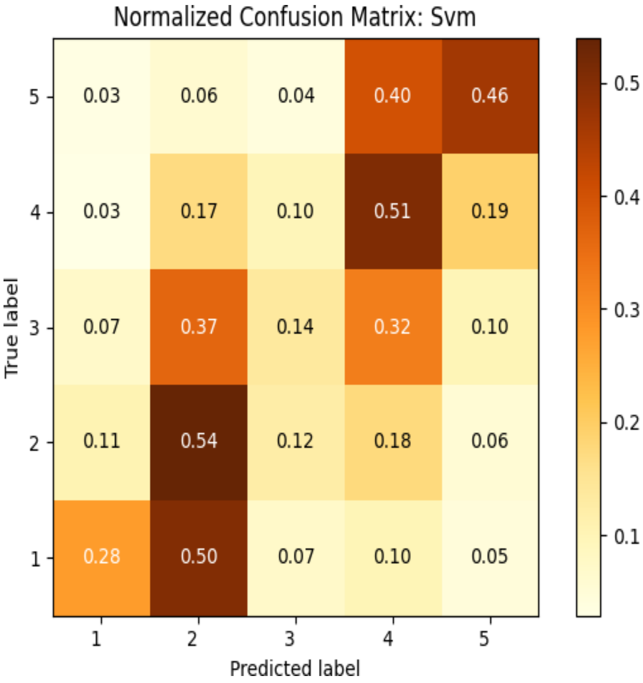


Fig 6. Confusion Matrix for SVM

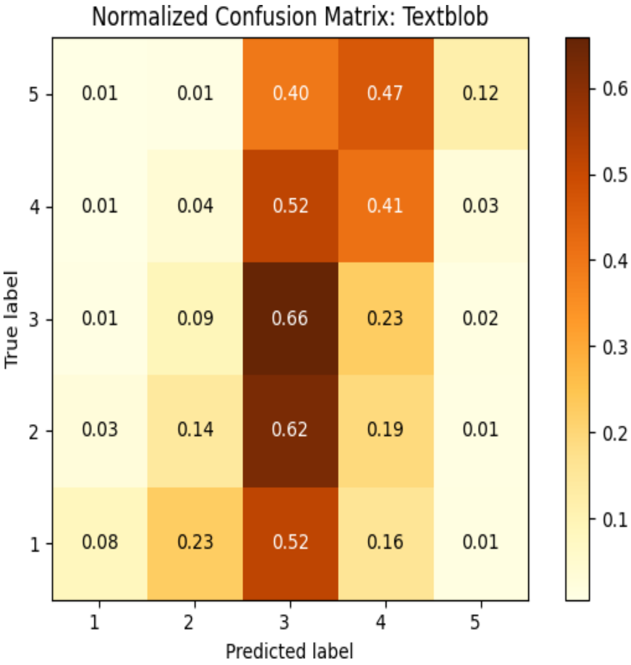


Fig 7. Confusion Matrix for Textblob

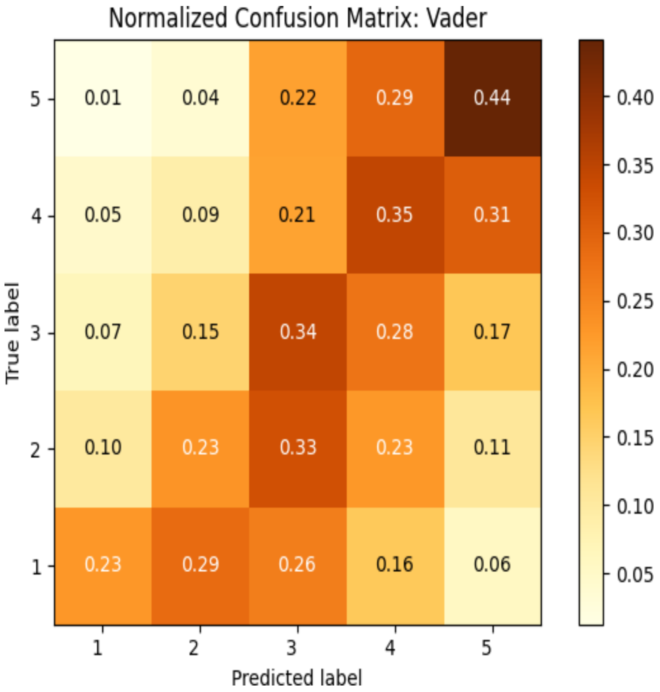


Fig 8. Confusion Matrix for Vader

VI. CONCLUSION

In our project we have used four different NLP classifiers for different class predictions using our dataset. By working with complex models, we were able to find the accuracy and F1 score nearly around 41% and 38% respectively. Machine learning techniques are trained and later the models are tested so the performance is better and accuracy is up to the mark. In the other hand, rule-based methods are better in class prediction and accuracy is low compare to machine learning techniques. A normalized confusion matrix is plotted for all the classifiers, which gives us better insights on how the class predictions and accuracy are obtained for machine learning based techniques and rule-based methods.

VII. REFERENCES

- Pang and L. Lee. (2008) Opinion mining and sentiment analysis. Foundations and Trends in Information Retrieval,2(1-2):1–135.
- Tom Mitchell Publisher: McGraw- Hill Science/Engineering/Math; (March 1, 1997) Machine Learning ISBN: 0070428077
- Dr. J.G.R. Sathiaseelan in (2017) Mining: Survey on Techniques and Applications by M Uma Maheshwari.
- Kshama Singh, SitaramPatel in (2019) A Survey Paper on Text Mining on Web Data Using Machine Learning Technique.
- Anna Schmidt, Michael Wiegandin (2018) A Survey on Hate Speech Detection using Natural Language Processing.
- Thomas Davidson, Dana Warmsley in (2018) Automated Hate Speech Detection and the Problem of Offensive Language.
- Shanita biere in (2018) Hate speech Detection using Natural Language ProcessingTechniques.
- Saud Alashri, Sultan Alzahrani, Muneera Alhoshan in (2019) Lexi-Augmenter: Lexicon-based Model for Tweets SentimentAnalysis.
- Elvira, Budhi Irwanin (2019) Hate Speech Detection on Instagram comment section using Maximum Entropy Classification Method.
- Shervin Malmasi and Mrcos Zampieri in (2019) Detecting Hate Speech in Social Media.
- Susmita Ray (2019) A Quick Review of Machine Learning Algorithms. DOI:10.1109/COMITCon.2019.8862451.
- Sheena Angra (2017) Machine learning and its applications. **DOI:** 10.1109/ICBDACI.2017.8070809