

# Hate Speech Detection in Conventional Language on Social Media by using Machine Learning

Resmi Reghunathan  
PG student, VLSI & Embedded Systems,  
ECE Department  
TKM Institute of Technology, Karuvelil P.O, Kollam,  
Kerala-691505, India

Asha A S  
Assistant Professor,  
ECE Department  
TKM Institute of Technology, Karuvelil P.O, Kollam,  
Kerala-691505, India

**Abstract-** Social media networks (SMNs) are the fastest means of communication as messages are sent and received almost instantaneously. The rise of social media platforms has significantly changed the way our world communicates, and part of those changes includes a rise in inappropriate behaviors, such as the use of aggressive and hateful language online. Giving opinions which are harsh or rude to someone directly on face is a difficult task. People feel it is safe over internet to abuse or post something offensive to others. Hate Speech can hurt a person or a community. Detecting such content is crucial to filtering or blocking inappropriate content on the Web. It is time-consuming and difficult to manually process and classify massive quantities of text data. However, due to the huge amount of data posted every day, automatic methods are essential for identifying such type of content. I am using three machine learning models such as SVM, Logistic Regression, Random Forest. Because of the hasty growth of social media such as blogs and social net-working sites, where individuals put in freely their perspectives on different themes. Researchers prove that people find it comfortable to be opinionated in their mother tongue, be it verbal or written. So, I am developing an automatic hate speech detection technique on English by collecting a huge amount of data set from online platforms. After checking the automatic hate speech algorithm then go for Malayalam (conventional language) and compare the performance of the models with the help of a confusion matrix.

**Keywords:** Support Vector Machine (SVM), Social media networks (SMNs)

## I. INTRODUCTION

Social media networks (SMNs) are the fastest approach of communique as messages are sent and obtained nearly straight away. SMNs are the primary media for perpetrating hate speeches these days. In line with this, cyber-hate crime has grown significantly in the previous couple of years. More research is being conducted to cut down on the rising cases of hate speeches in social media (SM). Different calls had been made to SM companies to filter every comment before allowing it into the public domain. The impacts of hate crimes are already overwhelming due to sizable adoption of SM and the anonymity enjoyed via the online users. In this period of huge information, it is time-consuming and difficult to manually process and classify huge quantities of textual content records. Besides, the precision of the categorization of manual text can without difficulty be influenced by human elements, inclusive of exhaustion and competence. To attain extra accurate and much less subjective results, it's miles beneficial to apply

machine learning (ML) techniques to automate the textual content classification procedures ..

Social networks encourage the interactions between people to be more indirect and anonymous as a result presenting anonymity for some people making them feel more secure despite the fact that they express hate speech. It Can easily lead to disruptive anti-social outcomes if it remains unregulated and uncontrolled. Hate speech is therefore taken into consideration as a severe hassle internationally, and many countries and organizations resolutely resist it. The polarity detection of speech on structures is the first step and is essential to government departments, social protection services, law enforcement and social media companies which expect to remove offensive content from their websites. Compared with guide filtering which is very time consuming, computerized identification of hate speech will enable the platform to hit upon the hate speech and cast off them a great deal greater quickly and efficiently. The problem of on-line hate speech detection has raised a hobby in each the scientific community and the business world. There had been many studies efforts aimed toward automating the technique which is usually modeled as a supervised classification problem. Recently, device getting to know technique that may study the extraordinary institutions between pieces of textual content, and that a specific output is anticipated for a selected input by using pre-categorized examples as schooling statistics is popular in scientific studies for hate speech detection. Among various device learning methods, deep learning which is a subset of machine learning that could be very prominent in Natural Language Processing (NLP) to tackle the problem of textual content classification.

In recent times a potential and intense research awareness because of the hasty growth of social media which include blogs and social internet working websites, in which individuals installed freely their perspectives on one of a kind topics. Researchers prove that people find it snug to opinionated in their mother tongue, be it verbal or written. Given that now almost all social systems help most of the famous languages, the requirement to mine the emotions in numerous dialects is at the rise. However, not all statistics can be applicable; some may not have any impact on the end result and a few may have comparable meanings. A preprocessing phase is for this reason required to help make the dataset concise. Malayalam just like the different languages in the Dravidian own family exhibits the traits of

an agglutinative language. The preprocessing manner consists of cleansing the statistics, tokenization, stop word elimination, etc. In India, in which there are 22 officially diagnosed languages and more than 132 crore human beings out of which a good sized percentage are active Internet customers, the amount of records generated on a daily foundation is big.

## MACHINE LEARNING

Machine learning (ML) is a field of inquiry devoted to understanding and building methods that 'learn', that is, methods that leverage data to improve performance on some set of tasks. It is seen as a part of artificial intelligence. Machine learning algorithms build a model Based on sample data, known as training data, in order to make predictions or decisions without being explicitly programmed to do so. Machine learning algorithms are used in a wide variety of applications, such as in medicine, email filtering, speech recognition, and computer vision, where it is difficult or unfeasible to develop conventional algorithms to perform the needed tasks. Machine learning approaches are traditionally divided into three broad categories, depending on the nature of the "signal" or "feedback" available to the learning system.

## SUPERVISED LEARNING

Supervised learning, as the name indicates, has the presence of a supervisor as a teacher. Basically supervised learning is when we teach or train the machine using data that is well labeled. Which means some data is already tagged with the correct answer. The supervised learning algorithm analyses the training data(set of training examples) and produces a correct outcome from labeled data. Supervised learning is classified into two categories of algorithms:

**Classification**-It is the process of finding or discovering a model or function which helps in separating the data into multiple categorical classes i.e. discrete values. In classification, data is categorized under different labels according to some parameters given in input and then the labels are predicted for the data. The classification process deals with the problems where the data can be divided into binary or multiple discrete labels. **Regression** -It is the process of finding a model or function for distinguishing the data into continuous real values instead of using classes or discrete values. It can also identify the distribution movement depending on the historical data. Because a regression predictive model predicts a quantity, therefore, the skill of the model must be reported as an error in those predictions

## HATE SPEECH

In antique instances, Hate Speech changed into restricted face to face conversations. But now because of the boom in social media systems the usage of hate speech is increasing. As human beings feel they're hidden on the net. Due to this, people feel secure to apply hate speech and it's human computive undertaking to identify hate speech on social

media so we need some automatic strategies to come across hate speech. On the other hand, people are more likely to share their views online, thereby leading to the dissemination of hate speech. Given that this type of prejudiced contact can be particularly un favorable to society, policymakers and social networking sites may also profit from monitoring and prevention gadgets Hate speech is typically described as any touch that distorts a character or network on the basis of traits such as coloration, ethnicity, gender, sexual preference, nationality or faith. According to Paula Fortuna and Sergia Nunes "Hate speech is language that assaults or diminishes, that incites violence or hate towards companies, primarily based on precise traits consisting of physical look, faith, descent, national or ethnic starting place, sexual orientation, gender identity or other. To penalize misclassification on minority classes weighted F1-rating is recommended as an evaluation measure. Nowadays with improvement in deep gaining knowledge of, CNN can be used for hate speech detection. Word-vector additionally acknowledged as phrase embedding may be educated on relevant corpus of the area. These pretrained phrase-vectors are used in CNN. Most gadget studying fashions makes use of bag-of phrases which fails to seize styles and sequences. Right here each phrase can be taken into consideration as hate speech but it's most possible that this sentence is hate speech. This sort of capabilities can't be treated with the aid of a bag of words which degrades the overall performance of traditional system learning algorithms.

## NATURAL LANGUAGE PROCESSING

Natural Language Processing or NLP (also referred to as Computational Linguistics) may be defined as the automated processing of human languages. As NLP is a huge and multidisciplinary field, but relatively a new location, there are many definitions accessible and practiced by way of one-of-a-kind human beings. One definition that could be part of any informed character's definition is "Natural Language Processing is a theoretically motivated range of computational techniques for analyzing and representing naturally occurring texts at one or more levels of linguistic analysis for the purpose of achieving human-like language processing for a range of tasks or applications." Thus, NLP is a discipline of computer technology and linguistics concerned with the interplay between computer systems and human (natural) languages. Moreover, it's miles pushed through advances in Machine Learning (ML) and also has an quintessential part of Artificial Intelligence (AI). The strategies of NLP are advanced in the sort of way that the instructions given in herbal language can be understood by the PC and additionally be capable of performing in step with it. It has to be mentioned that herbal language processing can be divided into two components, namely written and spoken language. Written languages play a less crucial position than speech in maximum activities, as the largest part of human linguistic communication occurs as speech. However, written language can be understood less complicated than spoken language, as spoken languages address a number of noise and ambiguities of the audio sign. Because there is a lot of ambiguity discovered in language,

NLP is seen as a difficult problem in computer science. Research in herbal language processing has been going on since the overdue 1940s. Machine translation (MT) was one of the first laptop-based software related to herbal language. Cambria and White (2014) feedback that NLP research has evolved from the technology of punch cards and batch processing, wherein the evaluation of a sentence may want to absorb to 7 mins, to the generation of Google and the likes of it, wherein thousands and thousands of web pages can be processed in less than a 2d. The most explanatory technique for offering what honestly happens inside a Natural Language Processing gadget is by means of the 'levels of language' approach. These ranges are utilized by humans to extract the meaning from text or spoken languages. This is because language processing particularly is predicated on formal models or representation of know-how related to those stages. Moreover, language processing packages distinguish themselves from fact processing systems with the aid of the usage of the expertise of language. The analysis of herbal language processing has the subsequent stages: Phonology, Morphology, Lexical, Syntactic, Semantic, Discourse and Pragmatic.

## II. RELATED WORK

There are many approaches for detection of hate speech. But they differ from each other based on the output they obtained in Ref. 8 hate speech was classified into three classes race, nationality and religion. Ref. 8 uses sentiment analysis technique for detection of hate speech but just not detecting but they also classified into one of the three classes and also rate the polarity of speech. We found two survey papers for automatic hate speech detection [6],[14]. In Ref. 6 motivation for hate speech detection is shown and why it became necessary to develop more robust and accurate models for automatic hate speech detection. The problem of hate speech detection is more often researcher keep data private while collecting it and there are less open source code available which make it difficult for comparative study [6]. This degrades the progress in this field. Different features related to hate speech are described in Ref. 14, like simple surface feature which includes bag of words, unigrams or n-grams. Both training set and testing set need to have same predictive word but it is problem as detection of hate speech is applied on very small piece of text so to overcome this issue word generalization is applied [14]. Knowledge of annotator for hate speech was examined in Ref. 15. Authors produce some very good results in amateur annotation in comparison to expert annotations. Also, Waseem provide its own dataset and its evaluation. To penalize misclassification on minority classes weighted F1-score is suggested as an evaluation measure. Nowadays with development in deep learning, CNN can be used for hate speech detection [2],[1]. Word-vector also known as word embedding can be trained on relevant corpus of the domain. This pre trained word-vectors are used in CNN [2]. Most of machine learning models uses bag-of words which fails to capture patterns and sequences. It can be understood by the example in Ref. 2. "if a tweet ends saying" if you know what I mean" here each word can be considered as hate speech but it is most likely that this sentence is hate speech. This type

of features cannot be handled by a bag of words which degrades the performance of traditional machine learning algorithms.

## III. METHODOLOGY

Methodology explains the proposed gadget which has been hired to classify speech into two special classes particularly, "hate speech, clean speech". Fig. 1 suggests the complete studies method. As proven on this figure, the studies method is contained in six key steps particularly, Data set collection, preprocessing, Feature extraction, Model Training, Perform evaluation, and Model testing..

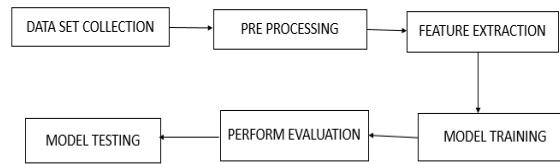


Figure1: Block diagram

## DATA SET COLLECTION

Most machine learning algorithms require data to be formatted in a very specific way, so datasets generally require some amount of preparation before they can yield useful insights. Some datasets have values that are missing, invalid, or otherwise difficult for an algorithm to process. If data is missing, the algorithm can't use it. If data is invalid, the algorithm produces less accurate or even misleading outcomes. Some datasets are relatively clean but need to be shaped (e.g., aggregated or pivoted) and many datasets are just lacking useful business context (e.g., poorly defined ID values), hence the need for feature enrichment. Good data preparation produces clean and well-curated data which leads to more practical, accurate model outcomes. Mainly using two types of data set 1) English 2) Malayalam , the English data set can be downloaded from online social media platforms like GitHub. The Malayalam data set is created. The data set mainly consists of two classes a) Class0 b) Class1 where class0 indicates hate speech and class 1 indicates clean speech. In this 30% of data were used for testing purposes and the remaining 70% used for training purposes. Following images showing some examples of data sets. For improving the accuracy data balancing operations are to be done.

```
2176 r1llamckilla: @joeDelanoITPACK &#128514;&#128514;&#128514;&#128514;&#128514; this  
2177 apawkey: Lol rt. @Cameron5833: If yo bitch don't read books or eat fruits... She a hoe. &#8221;&#128514;&#12851  
2178 aynlga: Charlie Strong WILL have Texas back as a power&#221; better power of big 12 is anybody's now  
2179 iechatfruth: Trap keep will fly that's why I love him &#10084; http://t.co/AlexGC2414&#8221; ayyyy you got tha  
2180 illasfvck: all these bitches getting mommy money&#8221; niggas too.  
2181 rue_east: That &#8226; God&#8221; by Drake is so trash, dawg.&#8221; Thank you ass water !!!&#128686;&#12868  
2182 ruuuchains: If you already out of school, fuck you I hate you and I hope you trip on the stage at graduation h  
2183 evy_8: i want affection 24/7 but i also want to be left alone 24/7... do you see the issue here?&#8221;fag  
2184 fflovel: show a lotta love to my sisters tho but these bitches so predictable.&#8221; Im in trouble  
2185 DTIAeatTHEass: Niggas swear they love a honest bitch tho http://t.co/vastMFJphu&#8221; &#128514;&#128514;&#12851  
2186 HIGheatness : &#8220;@SGKo71: @urhIGheatness &#128514; nah lemme chll i love the bitches&#8221;&#128557;&#128557  
2187 rkannat : &#8220;@what_evaittakes: Nasty pay the bills bitch&#128540;&#8221; &#128524;&#8221;&#128514;  
2188 erkannat_ : He told the bitch "stop texting my phone my shawty crazy" &#128524;&#128528;&#8221;glad he know &
```

Figure2: English data set sample

Fig3: Malayalam data set sample

## PRE-PROCESSING

The process of converting data to something a computer can understand is referred to as pre-processing. One of the major forms of pre-processing is to filter out useless data. Stop Words: A stop word is a commonly used word (such as “the”, “a”, “an”, “in”) that a search engine has been programmed to ignore, both when indexing entries for searching and when retrieving them as the result of a search query. In the dataset, carried out different preprocessing-techniques to clear out noisy and non-informative functions. In preprocessing, change the given data into a lower case. Also, removed all the URLs, usernames, white areas, hashtags, punctuations and stop-words using sample matching strategies from the collected speech. Besides this, also accomplished tokenization and stemming from preprocessed speech. The tokenization, converts every single speech into tokens or words, then the porter stemmer converts phrases to their root forms, such as angry to offend using porter stemmer. Various methods for machine learning models are used to achieve higher evaluation measures. Methods Used are explained below.

## TOKENIZING

Tokenizing is the process in which each sentence is divided into words. It can be done to either separate words or sentences. If the text is split into words using some separation technique it is called word tokenization and the same separation done for sentences is called sentence tokenization. It is breaking the raw text into small chunks. Tokenization breaks the raw text into words, sentences called tokens. These tokens help in understanding the context or developing the model for the NLP. The tokenization helps in interpreting the meaning of the text by analyzing the sequence of the words. For example, the text “It is raining” can be tokenized into ‘It’, ‘is’, ‘raining’. This method is used to create vocabulary for a dataset. This vocabulary is used to represent each speech in the dataset representation based on a choice of method like TF-IDF or bag of words.

## STOP WORDS

Stop words are those words which have less meaning or are useless for e.g, a, the, of which often occurs in most sentences. We would not want these words to take up space in our database, or take up valuable processing time. So it is required to remove these stop words otherwise it will cause misclassification. For this, we can remove them easily, by storing a list of words that you consider to stop words. NLTK(Natural Language Toolkit) in python has a list of stop words stored in 16 different languages.

## STEMMING

Stemming is the process of producing morphological variants of a root/base word. It removes the prefix or suffix of a word. Stemming programs are commonly referred to as stemming algorithms or stemmers. A stemming algorithm reduces the words “chocolates”, “chocolatey”, “choco” to the root word, “chocolate” and “retrieval”, “retrieved”, “retrieves” reduce to the stem “retrieve”. Stemming is an important part of the pipelining process in Natural language processing. The input to the stemmer is tokenized words.

## LEMMATIZATION

It is the process of grouping together the different inflected forms of a word so they can be analyzed as a single item. It usually refers to doing things properly with the use of a vocabulary and morphological analysis of words. It normally aims to remove inflectional endings only and to return the base or dictionary form of a word, which is known as the lemma , which is a root word rather than root stem, the output of stemming. After lemmatization I got a valid word that means the same thing. Basically it is similar to stemming but it brings context to the words. So it links words with similar meanings to one word. Text preprocessing includes both Stemming as well as Lemmatization. Many times people find these two terms confusing. Some treat these two as the same. Actually, lemmatization is preferred over Stemming because lemmatization does morphological analysis of the words. Applications of lemmatization are: Used in comprehensive retrieval systems like search engines, used in compact indexing.

## FEATURE EXTRACTION

The ML algorithms cannot understand the classification rules from the raw text. These algorithms need numerical features to understand classification rules. Hence, in text classification one of the key steps is feature engineering. This step is used for extracting the key features from raw text and representing the extracted features in numerical form. In this I have performed TFIDF technique.

## MODEL TRAINING

According to the “no free lunch theorem”, there is no single classifier which best performs on all kinds of datasets. Therefore, it is recommended to apply several different classifiers on a master feature vector to observe which one achieves better results. Here use three different machine learning models: Support Vector machine (SVM), Logistic regression (LR), Random forest (RF).

## SUPPORT VECTOR MACHINE (SVM)

Support Vector Machine or SVM is one of the most popular Supervised Learning algorithms, which is used for Classification as well as Regression problems. However, primarily, it is used for Classification problems in Machine Learning. Support vector machine is extremely favored by many as it produces notable correctness with less computation power .The goal of the SVM algorithm is to create the best line or decision boundary that can segregate n-dimensional space into classes so that we can easily put

the new data point in the correct category in the future. This best decision boundary is called a hyperplane. In two-dimensional space, this hyperplane is a line splitting a plane into two parts where each class lies on either side.

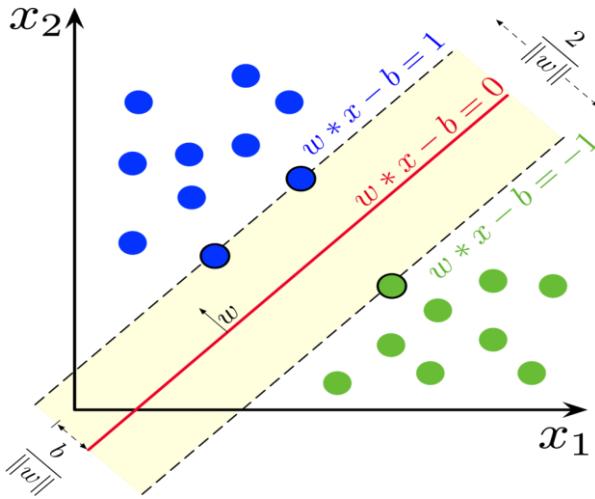


Fig4: SVM

#### LOGISTIC REGRESSION (LR)

Logistic regression is one of the most popular Machine Learning algorithms, which comes under the Supervised Learning technique. It is used for predicting the categorical dependent variable using a given set of independent variables. It predicts the output of a categorical dependent variable. Therefore the outcome must be a categorical or discrete value. It can be either Yes or No, 0 or 1, true or False, etc. but instead of giving the exact value as 0 and 1, it gives the probabilistic values which lie between 0 and 1.

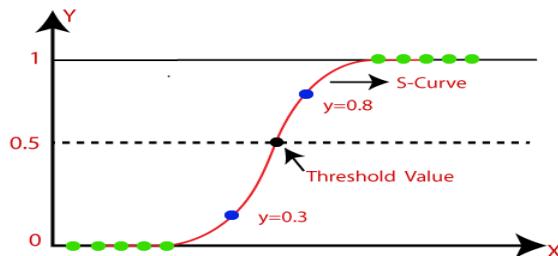


Fig5: Sigmoid function

$$f(n) = \frac{1}{1+e^{-n}} \quad (1)$$

$$n = \beta_0 + \beta_1 x_1 + \dots + \beta_n x_n \quad (2)$$

$\beta$  is the random value it may be 1.  $x_1, x_2, \dots, x_n$  will be the each TF-IDF value in row wise. The output value will be in range [0,1]. Based on the range we can classify the speech.

$$F(n) = \frac{1}{1+e^{-n}} - (\beta_0 + \beta_1 x_1 + \dots + \beta_n x_n) \quad (3)$$

It is the probability of the datapoint belonging to any class. Then cutoff point (mostly 0.5) is used to divide between the two classes. Cutoff point is not fixed; it can be changed according to the dataset.

#### RANDOM FOREST (RF)

A series of trees is referred to as forest. Similarly, Random forest is a collection of decision trees and it's far called random due to the fact it's far from a group of tremendously uncorrelated bushes working as a single version. Random forest is a supervised learning algorithm which is used for both classification as well as regression. But it is mainly used for classification problems. As we know that a forest is made up of trees and more trees means more robust forest. Similarly, a random forest algorithm creates decision trees on data samples and then gets the prediction from each of them and finally selects the best solution by means of voting. It is an ensemble method which is better than a single decision tree because it reduces the over-fitting by averaging the result

#### PERFORMANCE EVALUATION

In this step, the constructed classifier predicts the elegance of unlabeled textual content (i.e. "hate speech, clean speech") using a test set. The classifier performance is evaluated with the aid of calculating true negatives (TN), false positives (FP), false negatives (FN) and true positives (TP). These four numbers constitute a confusion matrix. Different performance metrics are used to assess the overall performance of the built classifier. Some not unusual overall performance measures in textual content categorization are mentioned briefly underneath.

**Precision:** Precision is also known as the positive predicted value. It is the proportion of predictive positives which are actually positive. Or it can be defined as the number of correct outputs provided by the model or out of all positive classes that have predicted correctly by the model, how many of them were actually true.

$$Pr = TP / (TP + FP) \quad (4)$$

**Recall:** It is the proportion of actual positives which are predicted positive.

$$R_c = TP / (TP + FN) \quad (5)$$

**F-Measure:** It is the harmonic mean of precision and recall. The standard F-measure gives equal importance to precision and recall. If two models have low precision and high recall or vice versa, it is difficult to compare these models. So, for this purpose, we can use an F-score. This score helps us to evaluate the recall and precision at the same time. The F-score is maximum if the recall is equal to the precision

$$F = 2 * (Pr + R_c) / (Pr + R_c) \quad (6)$$

**Accuracy:** It is one of the important parameters to determine the accuracy of the classification problems. It defines how

often the model predicts the correct output. It can be calculated as the ratio of the number of correct predictions made by the classifier to all number of predictions made by the classifiers. It is the number of correctly classified instances (true positives and true negatives).

$$A = (TP+TN)/(TP+FP+FN+TN) \quad (7)$$

TP is True Positive. FP is False Positive. This refers to non-hate speeches that were classified as hate speech. FN is False Negative. This refers to those hate speeches that were not identified by the model as hate speech. It identified as non hate speech. TN is True Negative

#### IV. RESULT AND DISCUSSION

Most machine learning algorithms require data to be formatted in a very specific way, so datasets generally require some amount of preparation before they can yield useful insights. Good data preparation produces clean and well-curated data which leads to more practical, accurate model outcomes. Mainly using two types of data set 1) English 2) Malayalam , the English data set downloaded from online social media platforms like GitHub. The Malayalam data set is created. In this 30% of data were used for testing purposes and the remaining 70% used for training purposes. Performed the preprocessing operations for the data. The ML algorithms cannot understand the classification rules from the raw text. These algorithms need numerical features to understand classification rules . Hence, in text classification one of the key steps is feature engineering. This step is used for extracting the key features from raw text and representing the extracted features in numerical form. In this I have performed TFIDF technique. SVM, logistic regression and random forest are different machine learning algorithms that are used. Confusion matrix helps to find the performance evaluation parameters of different machine learning models. A confusion matrix is a matrix used to determine the performance of the classification models for a given set of test data . The order of the matrix depends upon the number of classes. Each row of the matrix represents the instances in an actual class while each column represents the instances in a predicted class, or vice versa . It compares the actual class of data and the predicted output of each model of the same data Confusion matrix of each model is shown below.

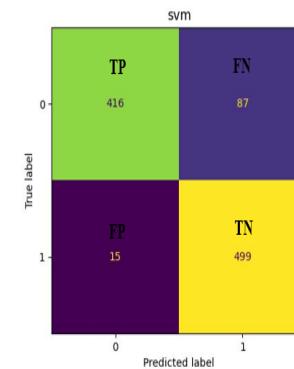


Fig 6: Confusion matrix of SVM for English dataset

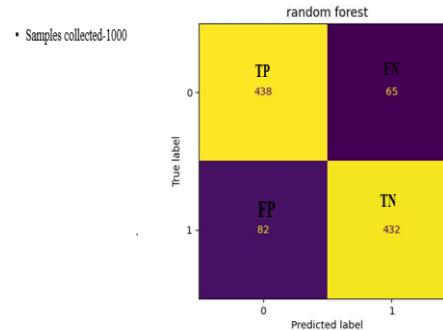


Fig 7: Confusion matrix of random forest for English dataset

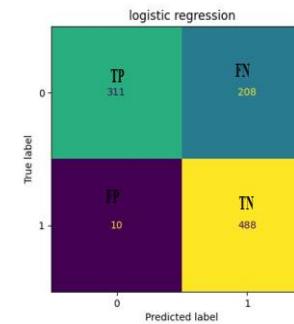


Fig 8 : Confusion matrix of Logistic regression for English dataset

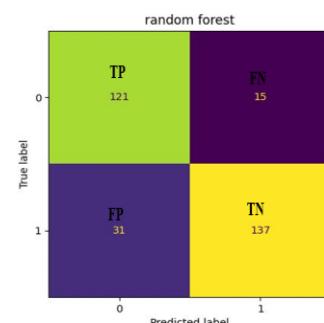


Fig 9 : Confusion matrix of Random forest for Malayalam dataset

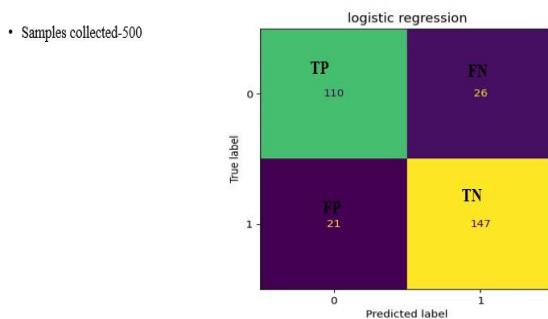


Fig 10 : Confusion matrix of Logistic regression for Malayalam dataset

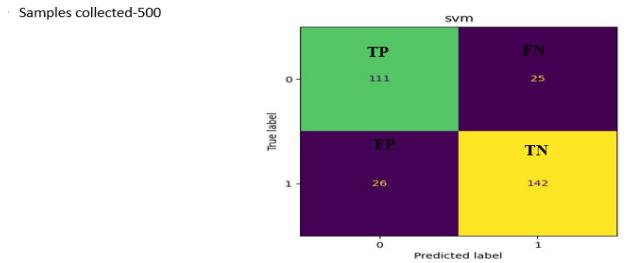


Fig 11 : Confusion matrix of SVM for Malayalam dataset

By analyzing the confusion matrix obtained the performance evaluation characters of different model

Table4.1: Different models result

SPECIFICATIONS	SVM		LOGISTIC REGRESSION		RANDOM FOREST	
	ENGLISH	MALAYALAM	ENGLISH	MALAYALAM	ENGLISH	MALAYALAM
ACCURACY	90	93.75	81	90	86	92
PRECISION	91	94	85	90	86	92
RECALL	90	94	80	91	86	93
F1SCORE	90	94	80	90	86	92

## V. CONCLUSION

In routine life, as the usage of social media is increased everyone seems to think like they can speak or write anything they want. Due to this thinking, hate speech has increased. Hate speech can hurt a person or a community. But it is very difficult to identify the hate speech manually. So it becomes necessary to automate the process of classifying the hate speech data. To simplify the process of classifying hate speech I have used a machine learning approach to detect hate speech from the speech. Most machine learning algorithms require data to be formatted in a very specific way, so datasets generally require some amount of preparation before they can yield useful insights. Good data preparation produces clean and well-curated data which leads to more practical, accurate model outcomes. Mainly using two types of data set 1)English 2) Malayalam, the English data set downloaded from online social media platforms like GitHub. The Malayalam data set is created. In this 30% of data were used for testing purposes and the remaining 70% used for training purposes. The ML algorithms cannot understand the classification rules from the raw text. These algorithms need numerical features to understand classification rules .For this use TF-IDF and bag of words methods to extract features from the speech. To classify hate speech implemented machine learning algorithms like SVM, Logistic Regression and Random

Forest. After preprocessing with TF-IDF calculated the performance evaluation characters of different machine learning models such as SVM, Logistic Regression, Random Forest with the help of confusion matrix.. Found that SVM gives best performance with 90% Accuracy Score for English dataset and 94% accuracy for Malayalam dataset. Developed automatic hate speech detection techniques in English, Malayalam.

## REFERENCES

- [1]. Pinkesh Badjatiya, Shashank Gupta, Manish Gupta, and Vasudeva Varma. Deep learning for hate speech detection in tweets. In Proceedings of the 26th International Conference on World Wide Web Companion, pages 759–760, 2017.
- [2]. Md Abul Bashar and Richi Nayak. Qutnocturnal@ hasoc’19: Cnn for hate speech and offensive content identification in hindi language. In Proceedings of the 11th annual meeting of the Forum for Information Retrieval Evaluation (December 2019), 2019.
- [3]. Pete Burnap and Matthew L Williams. Cyber hate speech on twitter: An application of machine classification and statistical modeling for policy and decision making. *Policy & Internet*, 7(2):223–242, 2015.
- [4]. Fabio Del Vigna12, Andrea Cimino23, Felice Dell’Orletta, Marinella Petrocchi, and Maurizio Tesconi. Hate me, hate me not: Hate speech detection on facebook. In Proceedings of the First Italian Conference on Cybersecurity (ITASEC17), pages 86–95, 2017.
- [5]. Shimaa M Abd El-Salam, Mohamed M Ezz, Somaya Hashem, Wafaa Elakel, Rabab Salama, Hesham ElMakhzangy, and Mahmoud ElHefnawi. Performance of machine learning approaches on prediction of esophageal varices for egyptian chronic hepatitis c patients. *Informatics in Medicine Unlocked*, 17:100267, 2019.

- [6]. Shimaa M Abd El-Salam, Mohamed M Ezz, Somaya Hashem, Wafaa Elakel, Rabab Salama, Hesham ElMakhzangy, and Mahmoud ElHefnawi. Performance of machine learning approaches on prediction of esophageal varices for egyptian chronic hepatitis c patients. *Informatics in Medicine Unlocked*, 17:100267, 2019.
- [7]. Paula Fortuna and S'ergio Nunes. A survey on automatic detection of hate speech in text. *ACM Computing Surveys (CSUR)*, 51(4):1–30, 2018.
- [8]. Purnama Sari Br Ginting, Budhi Irawan, and Casi Setianingsih. Hate speech detection on twitter using multinomial logistic regression classification method. In 2019 IEEE International Conference on Internet of Things and Intelligence System (IoTaIS), pages 105–111. IEEE, 2019.
- [9]. Njagi Dennis Gitari, Zhang Zuping, Hanyurwimfura Damien, and Jun Long. A lexicon-based approach for hate speech detection. *International Journal of Multimedia and Ubiquitous Engineering*, 10(4):215–230, 2015.
- [10]. Ammar Ismael Kadhim. Term weighting for feature extraction on twitter: A comparison between bm25 and tf-idf. In 2019 International Conference on Advanced Science and Engineering (ICOASE), pages 124–128. IEEE, 2019.
- [11]. Harpreet Kaur, Veenu Mangat, and Nidhi Krail. Dictionary-based sentiment analysis of hinglish text and comparison with machine learning algorithms. *International Journal of Metadata, Semantics and Ontologies*, 12(2- 3):90–102, 2017.
- [12]. Joni Salminen, Maximilian Hopf, Shammur A Chowdhury, Soon-gyo Jung, Hind Almerekhi, and Bernard J Jansen. Developing an online hate classifier for multiple social media platforms. *Human-centric Computing and Information Sciences*, 10(1):1, 2020.
- [13]. TYSS Santosh and KVS Aravind. Hate speech detection in hindie-english code-mixed social media text. In Proceedings of the ACM India Joint International Conference on Data Science and Management of Data, pages 310–313, 2019.
- [14]. Anna Schmidt and Michael Wiegand. A survey on hate speech detection using natural language processing. In Proceedings of the Fifth International Workshop on Natural Language Processing for Social Media, pages 1–10, 2017.
- [15]. Zeerak Waseem. Are you a racist or am i seeing things? annotator influence on hate speech detection on twitter. In Proceedings of the first workshop on NLP and computational social science, pages 138–142, 2016.
- [16]. Tingxi Wen and Zhongnan Zhang. Effective and extensible feature extraction method using genetic algorithm-based frequency-domain feature search for epileptic eeg multiclassification. *Medicine*, 96(19), 2017
- [17]. Abro, S., Sarang Shaikh, Z. A., Khan, S., Mujtaba, G., & Khand, Z. H. Automatic Hate Speech Detection using Machine Learning: A Comparative Study. *Machine Learning*, 10, 6,2020.