

PROJECT NAME

## MALIGNANT COMMENT CLASSIFICATION

Submitted by : -

Anirudh Bhardwaj

## INTRODUCTION

* People Use social Media as a platform to express their opinions and views.
* They do this freely and without any reluctance.
* With it comes the risk of cyber abuse and harassment, which can put an end

to people expressing themselves and giving up on seeking opinions.

* To protect the user from cyberbullying every organization should have an automated system in place to identify such malicious comments and suitable actions against the same.





# OBJECTIVE

We have undertaken this project to classify the comments made by different users on social media platforms, like the tweets from Twitter in this case, to build a model that is capable of detecting and classifying these tweets into various classes like threat, obscene, insult and identity-based hate.

DATA EXTRACTION

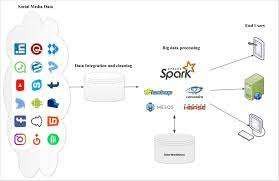
⯈ We started out by using the online social media platform for example twitter, facebook, insta etc. in Python to extract the comments .

⯈ We collected the comments containing the cuss words using the

track parameter to gather relevant data to perform our analysis.

⯈ We saved the data gather from online social media platform in a

CSV File.





## PREPROCESSING OF THE DATA

⯈ We have performed the following pre-processing on the data:

* Removed Punctuation
* Removed the stop words
* Stemming and lemmatization
* Applied counter vectorizer

## DATA MODELLING

Problem Transformation method

* + We used the Multinomial Naïve Bayes classifier for classification
  + We used this classifier, since it is suitable for classification with the

discrete features.

* + The algorithm estimates the conditional probability of a particular word given a class as the relative frequency of term in documents belonging to class.
  + The variation takes into account the number of occurrences of term in training documents from class include multiple occurrences.

## DATA MODELLING

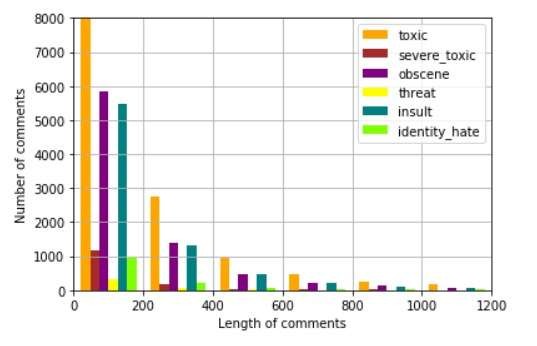
Adaptation Algorithms

⯈ We used the Backpropagation for Multilabel Learning algorithm.

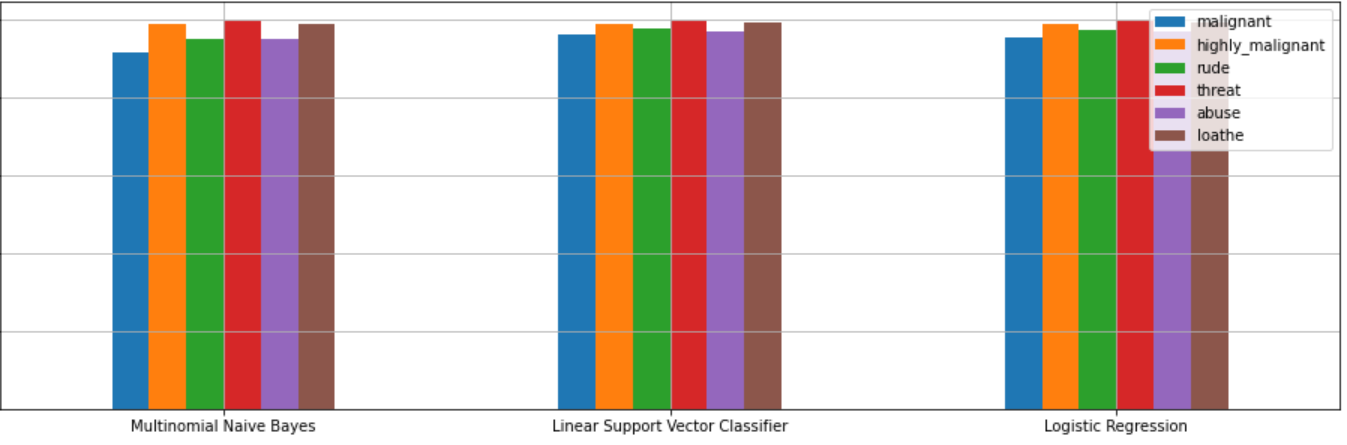
⯈ The BPMLL algorithm is a back-propagation neural network algorithm adapted for multi-label classification by having multiple binary outputs as the label variables.

⯈ It helped to classify the tweets in real-time as the algorithm changes its behavior at the time it is run, based on information available and on a *priori* defined criterion.

## DATA VISUALIZATIONS



ANALYSIS OF RESULTS



## CHALLENGES

Multiple labels – For Binary classification we have many models available but for multilabel classification the selection of models decreases drastically.



Collection of USABLE DATA – although we extracted over 100000 records, only 14000 records were usable among them.



Cleaning the data – Due to the character limit tweets contain a lot of short forms which are different to classify using bag of words model.



TWITTER RATE LIMIT – Due to the

twitter rate limit we were getting timed out every 15 mins due to which the data collection process took over 8 hours.



## FUTURE SCOPE

⯈ The current project predicts the type or toxic in the comment.

⯈ We are planning to add the following features in the futures.

⯈ Analyze which age group is being toxic towards a particular group or brand.

⯈ Add feature to automatically sensitize words which are classified as toxic.

⯈ Automatically send alerts to the concerned authority if threats are classified

as severe.

⯈ Build a feedback loop to further increase the efficiency of the model.

⯈ Handle mistakes and short forms of words to get better accuracy of the result



# THANK YOU