**Dr B.R. Ambedkar National Institute of Technology Jalandhar**



**TWITTER SETIMENT ANALYSIS**

**Parth Chaudhary**

**(22103117)**

**Abstract:**

‘Sentiment analysis’ refers to **identifying as well as classifying the sentiments that are expressed in the text source**. Tweets are often useful in generating a vast amount of sentiment data upon analysis. These data are useful in understanding the opinion of the people about a variety of topics.

The objective of this task is to detect hate speech in tweets. For the sake of simplicity, we say a tweet contains hate speech if it has a racist or sexist sentiment associated with it. So, the task is to classify racist or sexist tweets from other tweets. For training the models, we have a labeled dataset of 31,962 tweets.

**INTRODUCTION:**

Machine learning: Machine learning is a subfield of artificial intelligence (AI). The goal of machine learning is to understand the structure of data and fit that data into models.

In traditional computing, algorithms are sets of explicitly programmed instructions used by computers to calculate or solving problems. Machine learning algorithms instead allow for computers to train on data inputs and use statistical analysis in order to output values that fall within a specific range.

Dataset: A data is a collection of data pieces that can be treated by a computer as a single unit for analytic and prediction purposes.

Data preprocessing: Data preprocessing refers to the technique of preparing (cleaning and organizing) the raw data to make it suitable for a building and training Machine Learning models.

Word Cloud :- A word cloud (also known as a tag cloud) is **a visual representation of words**. Cloud creators are used to highlight popular words and phrases based on frequency and relevance. They provide you with quick and simple visual insights that can lead to more in-depth analyses.

Classifiers: Classifier is a type of machine learning algorithm used to assign a class label to a data input.  
Example: Image recognition classifier to label an image (e.g., “car,” “truck,” or “person”).

Logistic Regression Classifier: Logistic regression is a supervised learning classification algorithm used to predict the probability of a target variable. The nature of target or dependent variable is dichotomous, which means there would be only two possible classes.

Accuracy: Accuracy is the fraction of predictions our model got right.

Accuracy = Number of correct predictions / Total number of predictions.

F1 score: The F1-score combines the precision and recall of a classifier into a single metric by taking their harmonic mean. It is primarily used to compare the performance of two classifiers.

**NLP : AN OVERVIEW**

Natural language processing (NLP) is the ability of a computer program to understand human language as it is spoken and written -- referred to as natural language. It is a component of artificial intelligence .

Tokenization :- Tokenization is breaking the raw text into small chunks. Tokenization breaks the raw text into words, sentences called tokens.

Normalization :- Normalization is helpful in  reducing the number of unique tokens present in the text , removing punctuations and also converting numbers into words.

Stemming:-  This divides words with inflection in them to root forms. Example: In the sentence, "fight”, ”fighter”, ”fighting”, the algorithm would be able to recognize the root of these words as "fight" .The algorithm can see that they are essentially the same word even though the letters are different.

hence, leading to a limitation that sometimes meaningful words lost its meaning , hence to overcome this lemmatization is used.

Lemmatization:- It is similar to stemming but it brings context to the words. For example it will convert “better”, ”best” into “good”.

Stop Words:- This is when common words (such as a ,the ,is ,etc)are removed from text so unique words that offer the most information about the text remain.

**OUR PROJECT**

Formally, given a training sample of tweets and labels, where label '1' denotes the tweet is racist/sexist and label '0' denotes the tweet is not racist/sexist, we have tried to predict the labels on the test dataset.

**LIBRARIES IMPORTED:**

**Import numpy as np:** NumPy is a Python library used for working with arrays. It also has functions for working in domain of linear algebra, fourier transform, and matrices.

**Import pandas as pd:** pandas is a library in Python for data manipulation and analysis (i.e. it offers data structures and operations for manipulating numerical tables and time series).

**Import matplotlib.pyplot as plt:** Matplotlib is a cross-platform, data visualization and graphical plotting library for Python and its numerical extension NumPy.

**Import seaborn as sns:** Seaborn is a library for making statistical graphics in Python. It builds on top of matplotlib and integrates closely with pandas data structures.

**Import nltk:** NLTK is a standard python library that provides a set of diverse algorithms for NLP. It is one of the most used libraries for NLP and Computational Linguistics.

**STEPS FOLLOWED:**

* For implementing the python model firstly we have imported all the required libraries like numpy , pandas ,etc.
* Then we have used Natural Language Preprocessing technique for cleaning the dataset.
* After that we trained the model with training data and then testing of data is done .
* We have used the LogisticRegressor classifier to predict the outcome.
* At last we checked the accuracy of these classifiers.

**WHY THIS MODEL**

Negativity spreads faster than positivity online, and news organizations at both ends of the political spectrum are leveraging this tendency on Twitter, according to a new study.

To test whether the broadcast news adage, “If it bleeds, it leads,” persists in the social media realm, Harvard Business School professor [Amit Goldenberg](https://www.hbs.edu/faculty/Pages/profile.aspx?facId=1194484) and colleagues Nathan Young and Andrea Bellovary of DePaul University analyzed 140,358 tweets posted by 44 news agencies in early 2020. An automated sentiment analysis tool confirmed their hunch: negativity was about 15 percent more prevalent than positivity, and negative tweets engaged more users.

“Although people produce much more positive content on social media in general, negative content is much more likely to spread,” says Goldenberg.

**CONCLUSION**

In this project , we successfully implemented model on “Twitter Sentiment Analysis” for indentifying hate speech and is also helpful to find a person spreading hate on twitter .

In the end , we conducted a thorough investigation into the efficacy of test modek in predicting hate tweets .Our results suggests that this test model can be successful in recognizing hate tweets .