Page :
Date : / /

Title : Twitter Data Analytics

Problem Statement : Use twitter data for sentiment analysis. The dataset is 3MB in size and has 31,962 tweets. Identify the tweets which hate tweets and which are not.

Objective : To apply sentiment analysis technique on twitter data and classify the tweets as hate or not.

Outcome : Students will learn to apply proper preprocessing on twitter data and also learn the basics of sentiment analysis.

Theory :

- A large amount of data that is generated today is unstructured, which requires processing to generate insights eg., data on news articles, posts on social media etc
- The process of analyzing natural language and making sense out of it falls under the field of Natural Language Processing (NLP).
- Sentiment analysis is a common task which involves classifying texts or parts of texts into a pre-defined sentiment.

P.T.O

Page :
Date : / /

## Dataset

- The dataset contains 31,962 labelled tweets.

Different steps involved are:

1. Loading the data
2. Tokenizing the data.
3. Normalizing the data
4. Determining word density
5. Model training

1. Loading the data: The data is provided as a csv file, we use 'pandas library' to load the data.

2. Tokenizing the data:

- Language in its original form cannot be accurately processed by a machine, so you need to make it easier for machine to understand.
- The first part of making sense of data is through a process called tokenization or splitting strings into smaller parts called tokens.

3. Normalizing the data:

- Words have different forms - for instance, 'ran', 'runs' and 'running' are various forms of the same verb 'run'.
- Normalization in NLP is the process of converting a word to its canonical form.

Two popular techniques of normalization are- stemming and lemmatization.

4. Determining word density.

- The most basic form of analysis on textual data is to take out the word frequency.
- A single tweet is too small of an entity to find out the distribution of words, hence, the analysis of the frequency of words would be done on all positive tweets.

5. Model training.

- We use the NaiveBayesClassifier to build the model.

Analysis:

The train dataset was split to create a validation dataset.

The accuracy on the validation data is 83.25% using the NaiveBayes classifier.

Conclusion: We have successfully classified the twitter data into hate/not hate label with 83.25% accuracy.