PUNE INSTITUTE OF COMPUTER TECHNOLOGY

Department of Computer Engineering

(2022- 2023)

Information Retrieval

Batch: - R4

Twitter Sentiment Analysis

Group members:

41443 – Anuj Mutha

41457 – Umesh Sawant

Guided by: - Prof. L.A. Pawar

IR Mini Project

Title: Twitter Sentiment Analysis.

Problem Statement: Develop a Sentiment Analysis model to categorize a tweet as Positive or Negative.

Date: 08/11/22

Objectives:

- To per-process the tweets.
- To analyse the data.
- To develop a ML model to analyse the sentiment of a tweet and classify it into positive or negative.
- And finally evaluate and compare the ML models.

Theory:

- **Required Libraries**

1. **NumPy:** NumPy is the fundamental package for scientific computing in Python. It is a Python library that provides a multidimensional array object, various derived objects (such as masked arrays and matrices), and an assortment of routines for fast operations on arrays, including mathematical, logical, shape manipulation, sorting, selecting, I/O, discrete Fourier transforms, basic linear algebra, basic statistical operations, random simulation and much more.

2. **Pandas:** pandas is a software library written for the Python programming language for data manipulation and analysis. In particular, it offers data structures and operations for manipulating numerical tables and time series. It is free software released under the three-clause BSD license. The name is derived from the term "panel data", an econometrics term for data sets that include observations over multiple time periods for the same individuals

3. **Matplotlib:** Matplotlib is an amazing visualization library in Python for 2D plots of arrays. Matplotlib is a multi-platform data visualization library built on NumPy arrays and designed to work with the broader SciPy stack. It was introduced by John Hunter in the year 2002**.**

4. **Scikit-learn (Sklearn):** Scikit-learn (Sklearn) is the most useful and robust library for machine learning in Python. It provides a selection of efficient tools for machine learning and statistical modelling including classification, regression, clustering and dimensionality reduction via a consistence interface in Python.

5. **NLTK (Natural Language Toolkit)**: is a leading platform for building Python programs to work with human language data. It provides easy-to-use interfaces to many corpora and lexical resources. Also, it contains a suite of text processing libraries for classification, tokenization, stemming, tagging, parsing, and semantic reasoning. Best of all, NLTK is a free, open source, community-driven project.

6. **TextBlob:** Text Blob is a python library for Natural Language Processing (NLP).TextBlob actively used Natural Language ToolKit (NLTK) to achieve its tasks. NLTK is a library which gives an easy access to a lot of lexical resources and allows users to work with categorization, classification and many other tasks. TextBlob is a simple library which supports complex analysis and operations on textual data.

7. **Tweepy:**
   Tweepy is an open source Python package that gives you a very convenient way to access the Twitter API with Python. Tweepy includes a set of classes and methods that represent Twitter's models and API endpoints, and it transparently handles various implementation details, such as: Data encoding and decoding.

8. **Streamlit**
   Streamlit is an open source app framework in Python language. It helps us create web apps for data science and machine learning in a short time. It is compatible with major Python libraries such as scikit-learn, Keras, PyTorch, SymPy(latex), NumPy, pandas, Matplotlib etc.

- **Important Points**

- **Natural Language Processing (NLP):**

The discipline of computer science, artificial intelligence and linguistics that is concerned with the creation of computational models that process and understand natural language. These include: making the computer understand the semantic grouping of words (e.g. cat and dog are semantically more similar than cat and spoon), text to speech, language translation and many more

- **Sentiment Analysis:** It is the interpretation and classification of emotions (positive, negative and neutral) within text data using text analysis techniques. Sentiment analysis allows organizations to identify public sentiment towards certain words or topics.
  Sentiment analysis, also referred to as opinion mining, is a sub machine learning task where we want to determine which is the general sentiment of a given document. Using machine learning techniques and natural language processing we can extract the subjective information of a document and try to classify it according to its polarity such as positive, neutral or negative.It is a really useful analysis since we could possibly determine the overall opinion about a selling objects, or predict stock markets for a given company like, if most people think positive about it,possibly its stock markets will increase, and so on. Sentiment analysis is actually far from to be solved since the language is very complex (objectivity/subjectivity, negation, vocabulary,grammar,...) but it is also why it is very interesting to working on.In this project I choose to try to classify tweets from Twitter into “positive” or “negative”sentiment by building a model based on probabilities. Twitter is a microblogging website where people can share their feelings quickly and spontaneously by sending tweets limited by 140 characters. You can directly address a tweet to someone by adding the target sign “@” or participate to a topic by adding an hashtag “#” to your tweet. Because of the usage of Twitter, it is a perfect source of data to determine the current overall opinion about anything.

- **Data Visualization**
  1. **Piechart**
     A pie chart (or a circle chart) is a circular statistical graphic, which is divided into slices to illustrate numerical proportion. In a pie chart, the arc length of each slice (and consequently its central angle and area) is proportional to the quantity it represents. While it is named for its resemblance to a pie which has been sliced, there are variations on the way it can be presented.

  2. **Countplot**

A count plot can be thought of as a histogram across a categorical, instead of quantitative, variable. The basic API and options are identical to those for barplot(), so you can compare counts across nested variables.

3. **Wordcloud**
   Word Cloud is a data visualization technique used for representing text data in which the size of each word indicates its frequency or importance. Significant textual data points can be highlighted using a word cloud. Word clouds are widely used for analyzing data from social network websites.

The Basics of NLP for Text Analytics

1. Sentence Tokenization
2. Word Tokenization
3. Text Lemmatization and Stemming
4. Stop Words
5. Regex
6. TF-IDF

1. **Sentence Tokenization**

Sentence tokenization (also called sentence segmentation) is the problem of dividing a string of written language into its component sentences. The idea here looks very simple. In English and some other languages, we can split apart the sentences whenever we see a

punctuation mark.

2. **Word Tokenization**

Word tokenization (also called word segmentation) is the problem of dividing a string of written language into its component words. In English and many other languages using some form of Latin alphabet, space is a good approximation of a word divider.

3.**Text Lemmatization and Stemming**

For grammatical reasons, documents can contain different forms of a word such as drive, drives, driving. Also, sometimes we have related words with a similar meaning, such as nation, national, nationality.

The goal of both stemming and lemmatization is to reduce inflectional forms and sometimes derivationally related forms of a word to a common base form.

Stemming usually refers to a crude heuristic process that chops off the ends of words in the hope of achieving this goal correctly most of the time, and often includes the removal of derivational affixes.
Lemmatization usually refers to doing things properly with the use of a vocabulary and morphological analysis of words, normally aiming to remove inflectional endings only and to return the base or dictionary form of a word, which is known as the lemma.

4.**Stop words**

Stop words are words which are filtered out before or after processing of text. When applying machine learning to text, these words can add a lot of noise. That's why we want to remove these irrelevant words.

Stop words usually refer to the most common words such as "and", "the", "a" in a language, but there is no single universal list of stopwords. The list of the stop words can change depending on your application.

The NLTK tool has a predefined list of stopwords that refers to the most common words. If you use it for your first time, you need to download the stop words using this code: nltk.download("stopwords"). Once we complete the downloading, we can load the stopwords package from the nltk.corpus and use it to load the stop words.

5.CountVectorizer

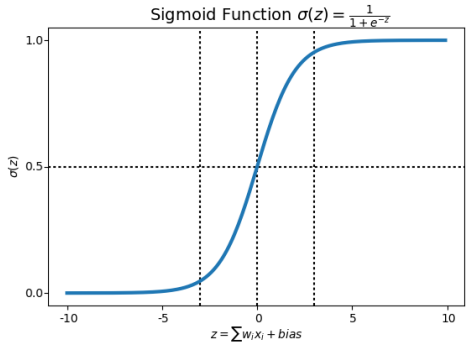
In order to use textual data for predictive modeling, the text must be parsed to remove certain words – this process is called tokenization. These words need to then be encoded as integers, or floating-point values, for use as inputs in machine learning algorithms. This process is called feature extraction (or vectorization).

Scikit-learn’s CountVectorizer is used to convert a collection of text documents to a vector of term/token counts. It also enables the pre-processing of text data prior to generating the vector representation. This functionality makes it a highly flexible feature representation module for text.

Algorithm Used in Project

1.Logistic Regression

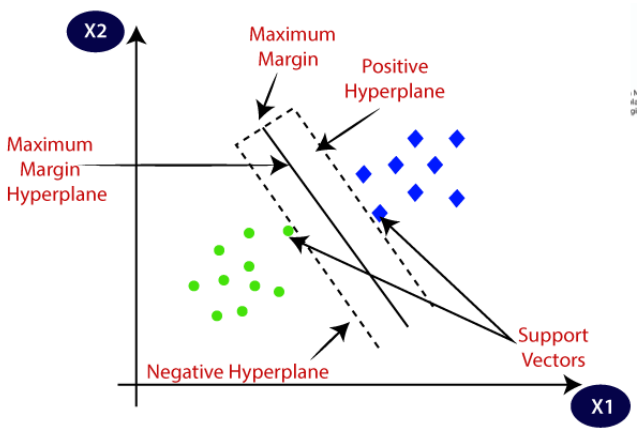
In statistics, the (binary) logistic model (or logit model) is a statistical model that models the probability of one event (out of two alternatives) taking place by having the log-odds (the logarithm of the odds) for the event be a linear combination of one or more independent variables ("predictors"). In regression analysis, logistic regression[1] (or logit regression) is estimating the parameters of a logistic model (the coefficients in the linear combination). Formally, in binary logistic regression there is a single binary dependent variable, coded by a indicator variable, where the two values are labeled "0" and "1", while the independent variables can each be a binary variable (two classes, coded by an indicator variable) or a continuous variable (any real value). The corresponding probability of the value labeled "1" can vary between 0 (certainly the value "0") and 1 (certainly the value "1"), hence the labeling;[2] the function that converts log-odds to probability is the logistic function, hence the name. The unit of measurement for the log-odds scale is called a logit, from logistic unit, hence the alternative names.



3. **Linear Support Vector Classification.**

Similar to SVC with parameter kernel='linear', but implemented in terms of liblinear rather than libsvm, so it has more flexibility in the choice of penalties and loss functions and should scale better to large numbers of samples.

This class supports both dense and sparse input and the multiclass support is handled according to a one-vs-the-rest scheme



System Architecture:

1. Jupyter Notebook

2. Python 3.9

3. Windows OS

4. 8GB RAM

5. I5 Processor 10th gen

Methodology/Steps:

1. Importing Dependencies
2. Importing Dataset
3. Preprocessing Text
   - 3.1. Lower Casing
   - 3.2. Replacing URLs
   - 3.3. Replacing Emojis
   - 3.4. Replacing Username

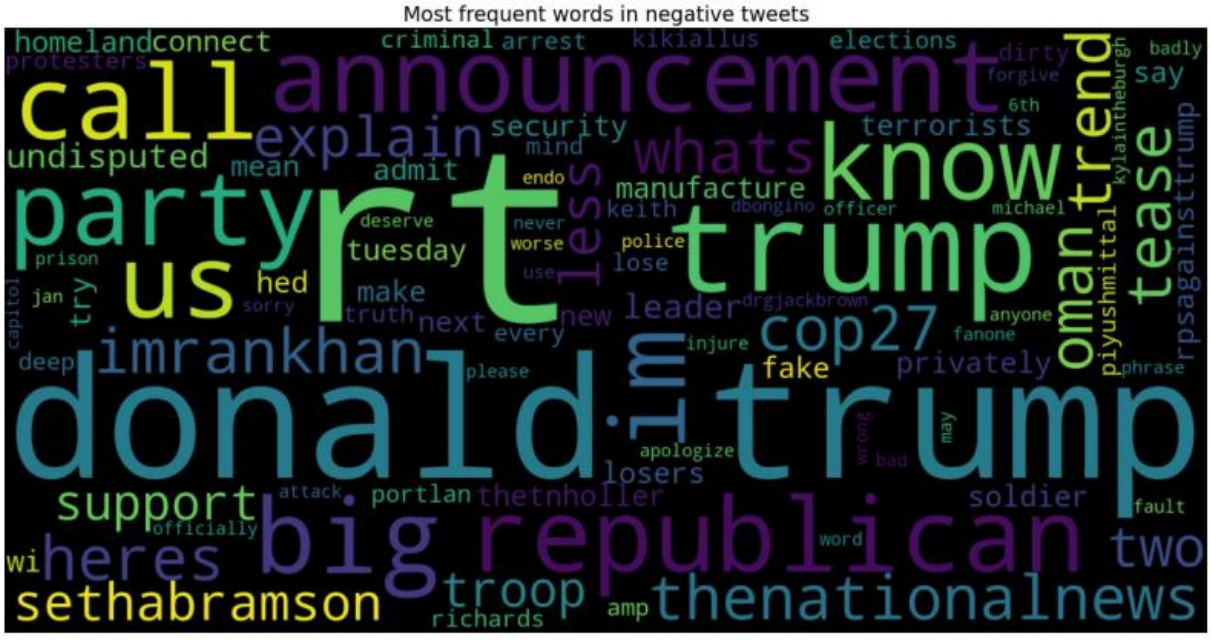
3.5. Removing Non-Alphabets
3.6. Removing Short Words
3.7. Removing Stop Words
3.8. Lemmatizing

4. Get polarity of tweets using TextBlob
5. Append a column sentiment and categoriesd tweets in negative positive and neutral tweets

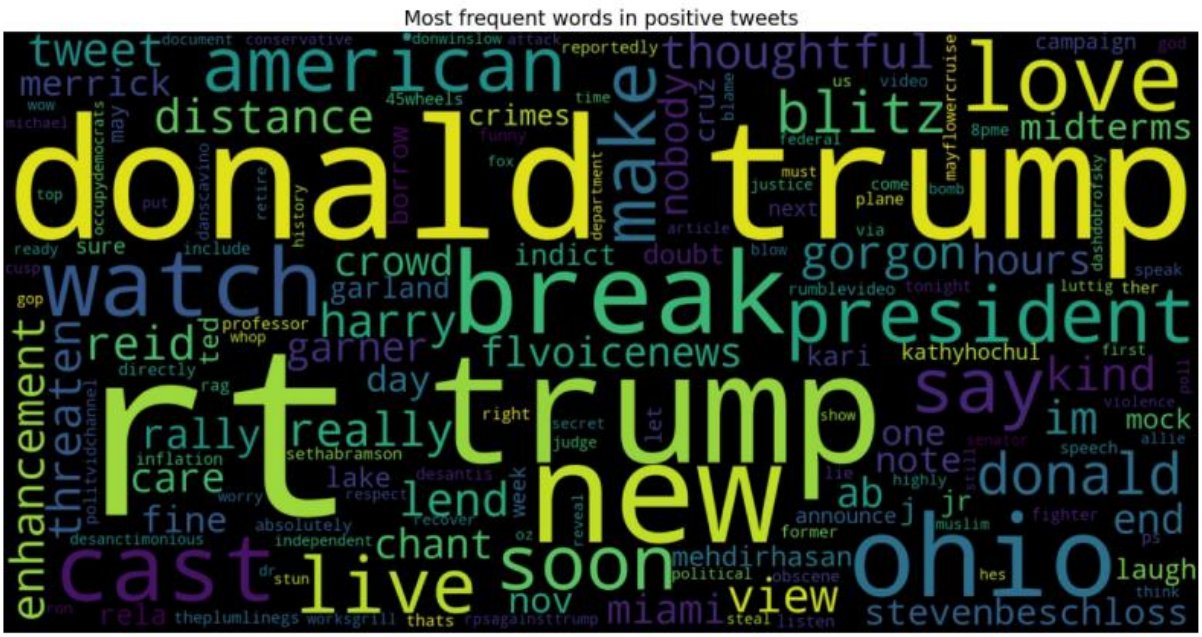
4. Analyzing Data
   4.1. Word Cloud for negative tweets
   4.2. Word Cloud for positive tweets
5. Splitting the data into training and testing
6. Count Vectorizer
7. Transforming the dataset
8. Creating and evaluating models
   8.1. Logistic Regression (LR)

Results:

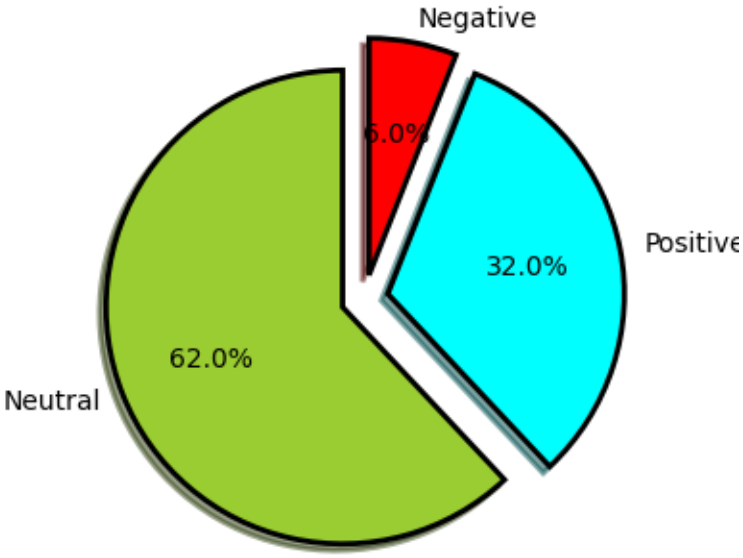
Word Cloud for negative tweets :



Word Cloud for positive tweets :



Analysis



Accuracy of ML algorithms used:-

Sr no.	ML Algorithm	Accuracy
1.	LinearSVC	0.78
2.	Logistic Regression	0.77

Conclusion:

In this mini project we analysis the tweets in domain of data science for that we do first Text preprocessing using NLTK library then using TextBlob we find polarity of tweets and the do sentiment analysis for that we use Logistic regression and SVM machine learning algorithm.
We find that the SVM give more accuracy than logistic regression. Hence SVM is more suitable for Tweets Analysis.