KENT
STATE
UNIVERSITY®

Twitter Sentiment Analysis

Presented By: Team # 6
1.Priyanka Yakkateela
2.Vanaja Karpurapu
3.Alekha Polavarapu

Contents:

- Introduction
- Problem Statement
- Literature Review
- Data set
- Data Processing & Analysis
- Methodology
- Results
- Conclusion

Introduction

- Sentiment Analysis refers to identifying and classifying the sentiments expressed in the text source
- Here we are going to classify the tweet data using machine learning techniques to determine whether the considered tweet is positive or negative

Problem Statement

- The process of identifying and classifying the emotional tone of a text is generally referred to as "sentiment analysis."
- The aim of this project is to create a social media evaluation and analysis . Here we are considering text data extracted from twitter API which contains sentiments positive and negative (which is Twitter Sentiment Analysis)

Literature review

a) Sentiment analysis involves categorizing attitudes in texts into "positive," "negative," or "neutral" categories.

b) For performance they have used machine learning algorithms. Classifiers that are used are Naive Bayes, Logistic Regression, Stochastic gradient descent and linear SVC.

c) For preprocessing they used maximum entropy, and support vector machine algorithms.

d) In few papers they used clustering techniques that are Supervised clustering and unsupervised clustering.

e) They have used datasets those are available from kaggle /Twitter API.

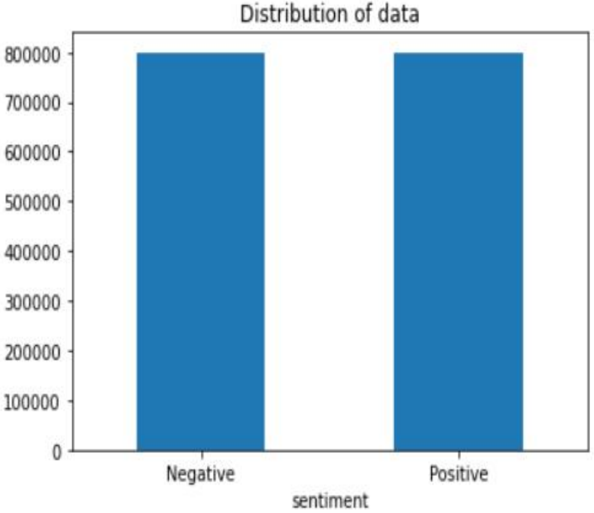
f) In order to evaluate, different machine learning like SVM, DTC, ABC, RFC are compared with deep learning models like CNN, LSTM, RNN. Accuracy, recall, precision, and specificity are evaluated against each model.

DataSet

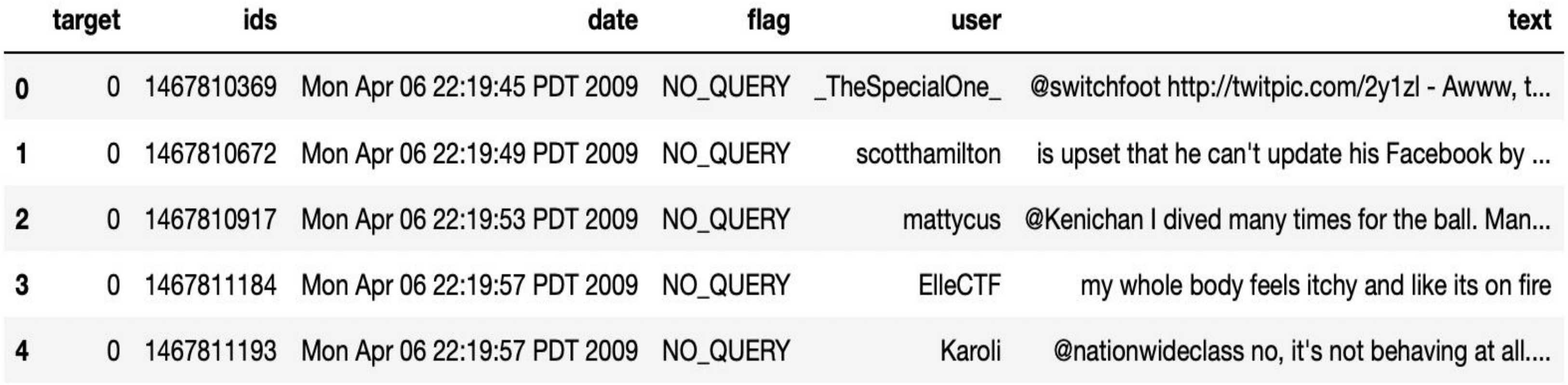
Open source data

Link : http://cs.stanford.edu/people/alecmgo/trainingandtestdata.zip

- Contains 1.6 million tweets
- 50% positive tweets & 50% negative tweets
- Contains 6 fields (sentiment, id , date, flag , user, text)



Dataset



	target	ids	date	flag	user	text
0	0	1467810369	Mon Apr 06 22:19:45 PDT 2009	NO_QUERY	_TheSpecialOne_	@switchfoot http://twitpic.com/2y1zl - Awww, t...
1	0	1467810672	Mon Apr 06 22:19:49 PDT 2009	NO_QUERY	scotthamilton	is upset that he can't update his Facebook by ...
2	0	1467810917	Mon Apr 06 22:19:53 PDT 2009	NO_QUERY	mattycus	@Kenichan I dived many times for the ball. Man...
3	0	1467811184	Mon Apr 06 22:19:57 PDT 2009	NO_QUERY	ElleCTF	my whole body feels itchy and like its on fire
4	0	1467811193	Mon Apr 06 22:19:57 PDT 2009	NO_QUERY	Karoli	@nationwideclass no, it's not behaving at all....

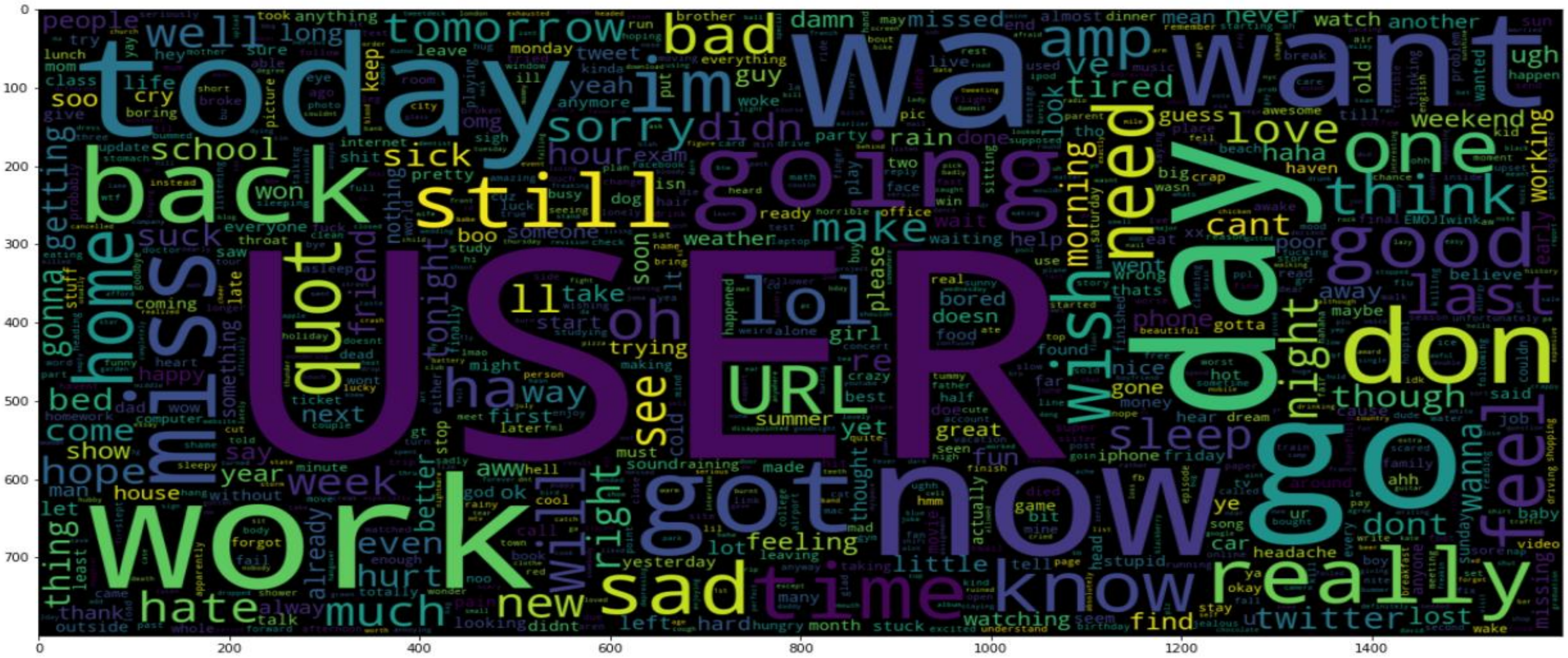
Data Processing & Analysis

- Removed unnecessary columns (All except sentiment and text)
- Process the text to lower case
- Replace emojis and non alphabets
- Removed stop words

Data Split : Train data - 95%
Test data - 5%

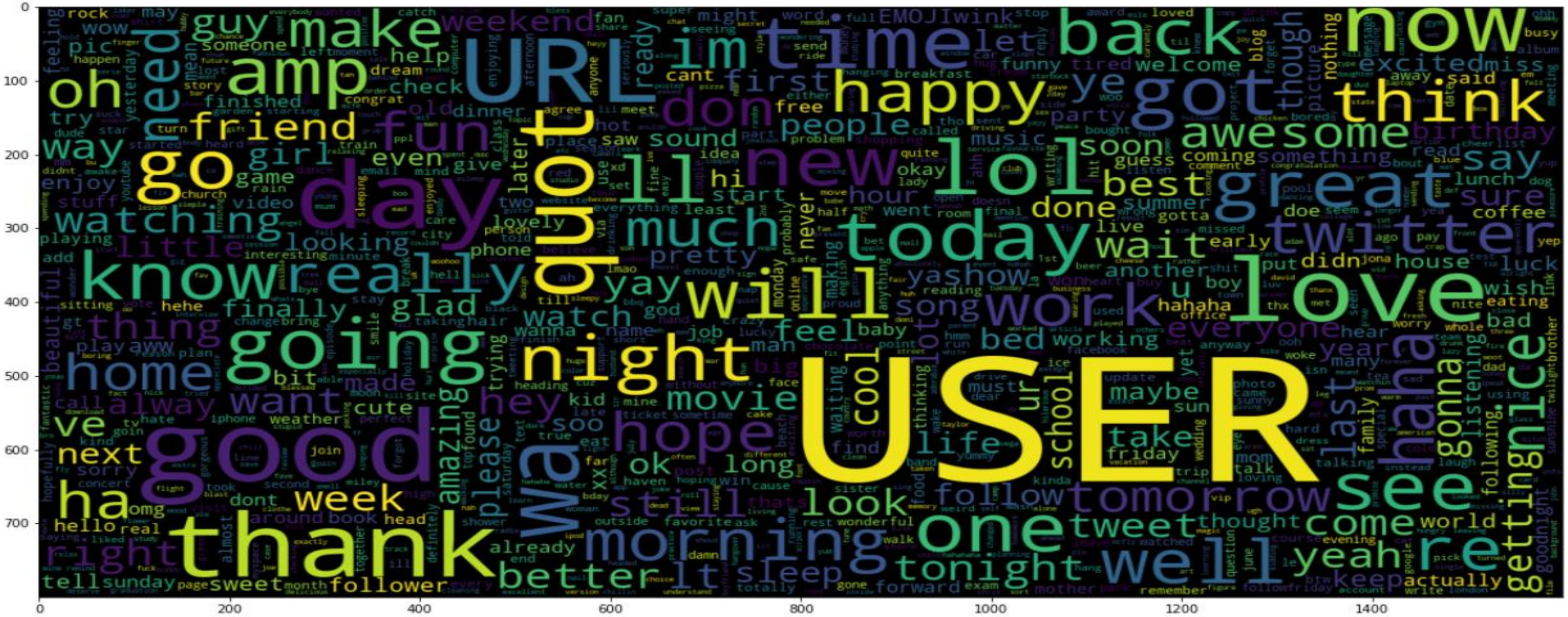
Data Processing & Analysis

Negative wordcloud



Data Processing & Analysis

Positive wordcloud



Libraries & Packages

- Numpy - python library to work with arrays
- pandas - To analyze and manipulate the data
- seaborn - Data visulatization library
- matplotlib - visualization library to plot array data
- nltk - natural language toolkit (provodes text processing libraries)
- sklearn - library with lot of machine learning tools for statistical modelling
- wordcloud - To visualize the text data (size ~ frequency or importance)
- pickle - To store the python objects into byte strams

Methodology

1. TF-IDF Vectoriser
2. Bernoulli Naive Bayes
3. Linear Support Vector Classification
4. Logistic Regression

TF-IDF Vectoriser

- $TF(t) = \frac{\text{(Number of Reputation of word in sentence)}}{\text{(number of words in the sentence)}}$

- $IDF(t) = \log \frac{\text{(number of sentences)}}{\text{(Number of sentences containing words)}}$

- $TF-IDF=TF \ast IDF$

Bernoulli Naive Bayes

- Bernoulli Naive Bayes is one of Naive Bayes algorithms. It only takes binary values. The most general example is where we check if each value will be whether or not a word that appears in a document.
- This is a very simplified model.
- The key benefit of this approach is that only binary values for characteristics, like:
  - True or False
  - Spam or Ham
  - Yes or No
  - 0 or 1

Linear Support Vector Classification

- Method that is useful for sorting linear problems. A LinearSVC's (Support Vector Classifier) purpose is to classify the data and return the best hyperplane for data classification.
- LinearSVC uses a linear kernel to execute SVC in a flexible manner.
- Linear SVM is used for linearly separable data, which means if a dataset can be classified into two classes by using a single straight line.

Logistic Regression

- The Logistic Regression Model is used to simulate the likelihood of a particular class of event circumstances, such as passing or falling, winning or losing, alive or dead, healthy or unhealthy.
- Logistic Regression is used when the dependent variable is categorical.

Evaluation metric

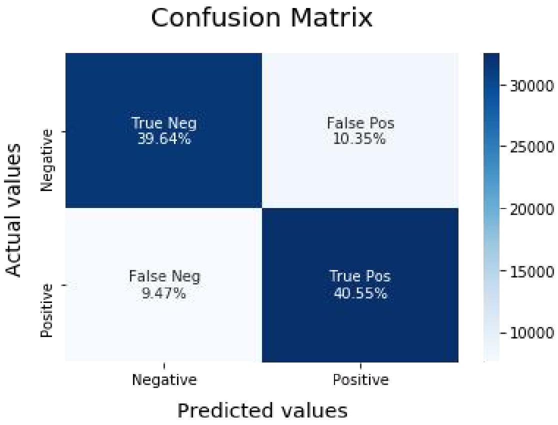
Confusion matrix - Summarizes and visualize the performance in table format

TP - True positive
TN - True negative
FP - False positive
FN - False negative

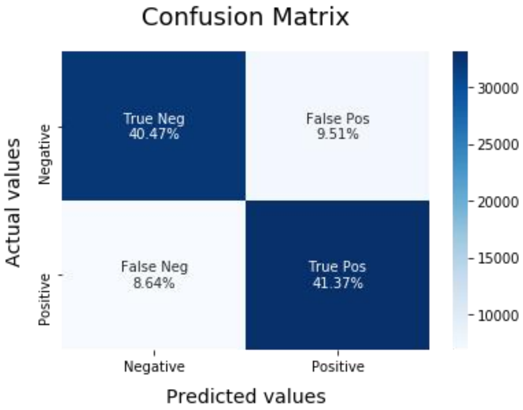
Accuracy = (TP +TN) / (TP+TN+FP+FN)

Results:

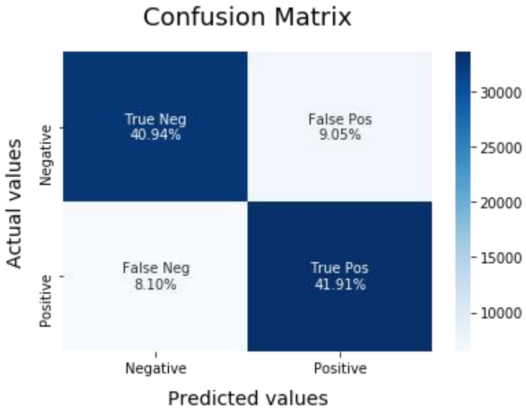
BernoulliNB Model:



LinearSVC Model:



Logistic Regression Model:



Conclusion:

- Out of all the models we tested, the Logistic Regression Model performs the best. Which is giving accuracy of about 82%.
- The BernoulliNB Model, although, is the quickest to train and predict on, it should also be mentioned. Additionally, it has 80% accuracy while calculating.

References:

[1] F. Neri, C. Aliprandi, F. Capeci, M. Cuadros and T. By, "Sentiment Analysis on Social Media," 2012 IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining, 2012, pp. 919-926, doi: 10.1109/ASONAM.2012.164. https://ieeexplore.ieee.org/stamp/stamp.jsp?tp=&arnumber=6425642

[2] Y. HongDa and K. Takano, "A Recommendation Method for Social Media Users based on a Sentiment Analysis Model," 2022 IEEE 4th Global Conference on Life Sciences and Technologies (LifeTech), 2022, pp. 485-488, doi: 10.1109/LifeTech53646.2022.9754863. https://ieeexplore.ieee.org/stamp/stamp.jsp?tp=&arnumber=9754863

KENT STATE UNIVERSITY®

Thank You!