Indian Institute of Information Technology
Allahabad
Department of Information Technology

REPORT

SENTIMENT ANALYSIS USING TWITTER

Raghvendra Kumar (IIM2016004)
Utsav Kumar Nareti (ISM2016001)
Piyush Gurnule (IIT2016035)
Avanish Chand (IIT2016120)

Supervisor: Dr. Shirshu Verma

IIIT ALLAHABAD

Indian Institute of Information Technology, Allahabad

(A UNIVERSITY ESTABLISHED UNDER SEC.3 OF UGC ACT, 1956 VIDE NOTIFICATION NO. F.9-4/99-U.3 DATED 04.08.2000 OF THE GOVT. OF INDIA) A CENTRE OF EXCELLENCE IN INFORMATION TECHNOLOGY ESTABLISHED BY GOVT. OF INDIA

August 22, 2020

Contents

Declaration

We hereby declare that the work presented in this mid semester project report of B.Tech (IT) 5th Semester entitled **"Sentiment analysis on twitter"**, submitted by us at **Indian Institute of Information Technology, Allahabad**, is an authenticated record of our original work carried out from **August 2018 to September 2018** under the guidance of **Dr. Shirshu Verma**. Due acknowledgements have been made in the text to all other material used. The project was done in full compliance with the requirements and constraints of the prescribed curriculum.

Place: Allahabad
Date: 22 November,2018
.
.

Raghvendra Kumar(IIM20160094)
Utsav Kumar Nareti (ISM2016001)
Piyush Gurnule (IIT2016035)
Avanish Chand(IIT2016120)

Certificate

This is to certify that the above statement made by the candidate is correct to the best of my knowledge.

Date:
Place: Allahabad
.

Dr. Shirshu Verma
Assistant Professor
IIIT-Allahabad

ACKNOWLEDGEMENT

We have tried our best to present this project in a complete manner without failing the deadlines, in this project. However, it would not have been possible without the kind support and help of many individuals. We would like to extend our sincere thanks to all of them. We are highly indebted to **Dr. Shirshu Verma** for his guidance and constant supervision as well as for providing necessary information regarding the project whenever we were stuck. We are largely indebted to them for their support in the project.

Abstract

This project addresses the problem of sentiment analysis in twitter; that is classifying tweets according to the sentiment expressed in them: positive, negative or neutral. Twitter is an online micro-blogging and social-networking platform which allows users to write short status updates of maximum length 140 characters. It is a rapidly expanding service with over 200 million registered users - out of which 100 million are active users and half of them log on twitter on a daily basis - generating nearly 250 million tweets per day . Due to this large amount of usage we hope to achieve a reflection of public sentiment by analyzing the sentiments expressed in the tweets. Analyzing the public sentiment is important for many applications such as firms trying to find out the response of their products in the market, predicting political elections and predicting socioeconomic phenomena like stock exchange. The aim of this project is to develop a functional classifier for accurate and automatic sentiment classification of an unknown tweet stream.

Introduction

We have chosen to work with twitter since we feel it is a better approximation of public sentiment as opposed to conventional internet articles and web blogs. The reason is that the amount of relevant data is much larger for twitter, as compared to traditional blogging sites. Moreover the response on twitter is more prompt and also more general (since the number of users who tweet is substantially more than those who write web blogs on a daily basis). Sentiment analysis of public is highly critical in macro-scale socioeconomic phenomena like predicting the stock market rate of a particular firm. This could be done by analyzing overall public sentiment towards that firm with respect to time and using economics tools for finding the correlation between public sentiment and the firm's stock market value. Firms can also estimate how well their product is responding in the market, which areas of the market is it having a favourable response and in which a negative response (since twitter allows us to download stream of geo-tagged tweets for particular locations). Predicting the results of popular political elections and polls is also an emerging application to sentiment analysis. One such study was conducted by Tumasjan et al. in Germany for predicting the outcome of federal elections in which concluded that twitter is a good reflection of offline sentiment .

The four main important topics include data pre-processing, the machine learning algorithms used, the tools required to execute the project as well as the feature extraction techniques along with features used.

Problem Statement

Mining Twitter data is a laborious and time-consuming process due to the restrictions and difficulties in its content. The informal language, the existence of slang, abbreviations, and the short length of the message are some of the problems while analyzing this data. Tweets can have a fundamental role since the study of its content can serve to characterize topics such as adjusting marketing strategy, develop product quality, improve customers services etc.

This project aims to design and develop a text processing and semantic analysis of tweets. More specifically, to practically implement such a framework we shall accomplish the following goals:

- To retrieve the tweets.
- To classify the tweets as positive, negative or neutral.
- Performing statistical analysis on the obtained results.
- The above statistics can be further used for business purposes.

Related Work

Research work in the area of Sentiment analysis are numerous. Some of the early results on Sentiment Analysis of twitter data are by Go et al. who used distant learning to acquire sentiment data. They used tweets with positive emoticons like ":)" and ";)" as positive, and tweets with negative emoticons like ":(" as negative. Sentiment Analysis on twitter data has been done previously by Go et al. where they have built the model using Naive Bayes, MaxEnt and SVM classifiers.They also perform some pre-processing of the data that was used in modeling the pre-processing techniques used in this project. The text processing they perform includes removal of URLs, username references and repeated characters in words.

A survey report from Pang and Lee on Opinion mining and sentiment analysis gives a comprehensive study in the area with respect to sentiment analysis of blogs, reviews etc. Algorithms used in the survey include Maximum Entropy , SVM and Naive Bayes.

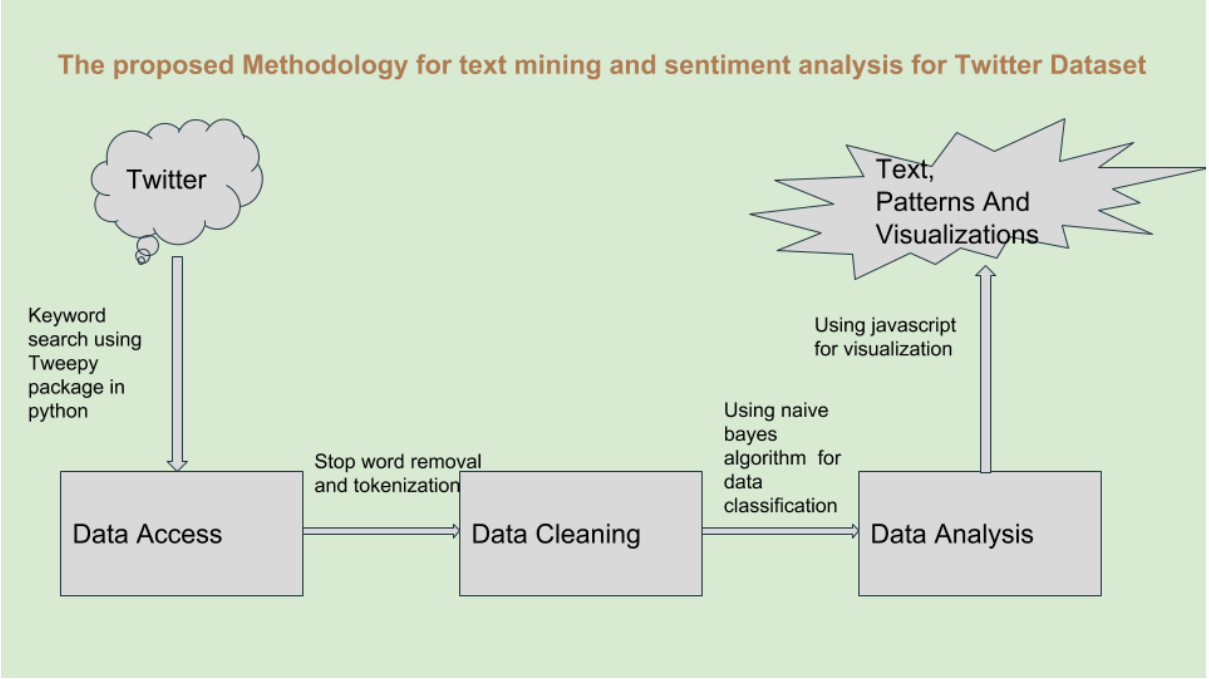
Methodology

NAIVE BAYES ALGORITHM

It is a classification technique based on Bayes Theorem with an assumption of independence among predictors. In simple terms, a Naive Bayes classifier assumes that the presence of a particular feature in a class is unrelated to the presence of any other feature. For example, a fruit may be considered to be an apple if it is red, round, and about 3 inches in diameter. Even if these features depend on each other or upon the existence of the other features, all of these properties independently contribute to the probability that this fruit is an apple and that is why it is known as 'Naive'.

Application of Naive Bayes Algorithm

Text classification/ Spam Filtering/ Sentiment Analysis: Naive Bayes classifiers mostly used in text classification (due to better result in multi class problems and independence rule) have higher success rate as compared to other algorithms. As a result, it is widely used in Spam filtering (identify spam e-mail) and Sentiment Analysis (in social media analysis, to identify positive and negative customer sentiments)


The proposed Methodology for text mining and sentiment analysis for Twitter Dataset
Twitter
Keyword search using Tweepy package in python
Data Access
Stop word removal and tokenization
Data Cleaning
Using naive bayes algorithm for data classification
Data Analysis
Using javascript for visualization
Text, Patterns And Visualizations

Data

Tweets are short length messages and have a maximum length of 140 characters. This limits the amount of information that the user can share with every message. Due to this reason, users use a lot of acronyms, hashtags, emoticons, slang and special characters. Acronyms and slang such as 2moro for tomorrow and so on are used to keep sentences within the word limit. People also refer to other users using the @ operator. Users also post URLs of webpages to share information. Emoticons are a great way to express emotions without having to say much.

Along with the twitter data, the project also required other datasets like stopwords , a dictionary of negative and positive words , an emoticon dictionary and an acronym dictionary for twitter slang words . The use of these are described in the next section.

Feature Set of negative and positive words.
The Feature Set of negative and positive words is a dataset containing around 6800 negative and positive words. This dataset is used to determine the numeric features of number of negative and positive words in the tweets, based on which sentiment classification is done. The process of stemming is also performed on this dataset, so that it maps to the training and test dataset.

Dataset Description

Twitter Data Set Description :
Tweets for Companies are downloaded using tweepy API provided by python. Each company has a hashtag associated on Twitter, for ex. #AAPL for Apple, #MSFT for Microsoft, #GOOG for Google. Using Tweepy library in python we are downloading the tweets.

Downloaded Twitter Dataset contains following attributes :

- **Date:** time stamp of the tweet.
- **Text:** contains the text of the tweet.

Hardware and Software requirements

Hardware requirements

Hardware	Operating System	Processor	RAM
System	Ubuntu 16.04 LTS	Intel Core i5	8GB

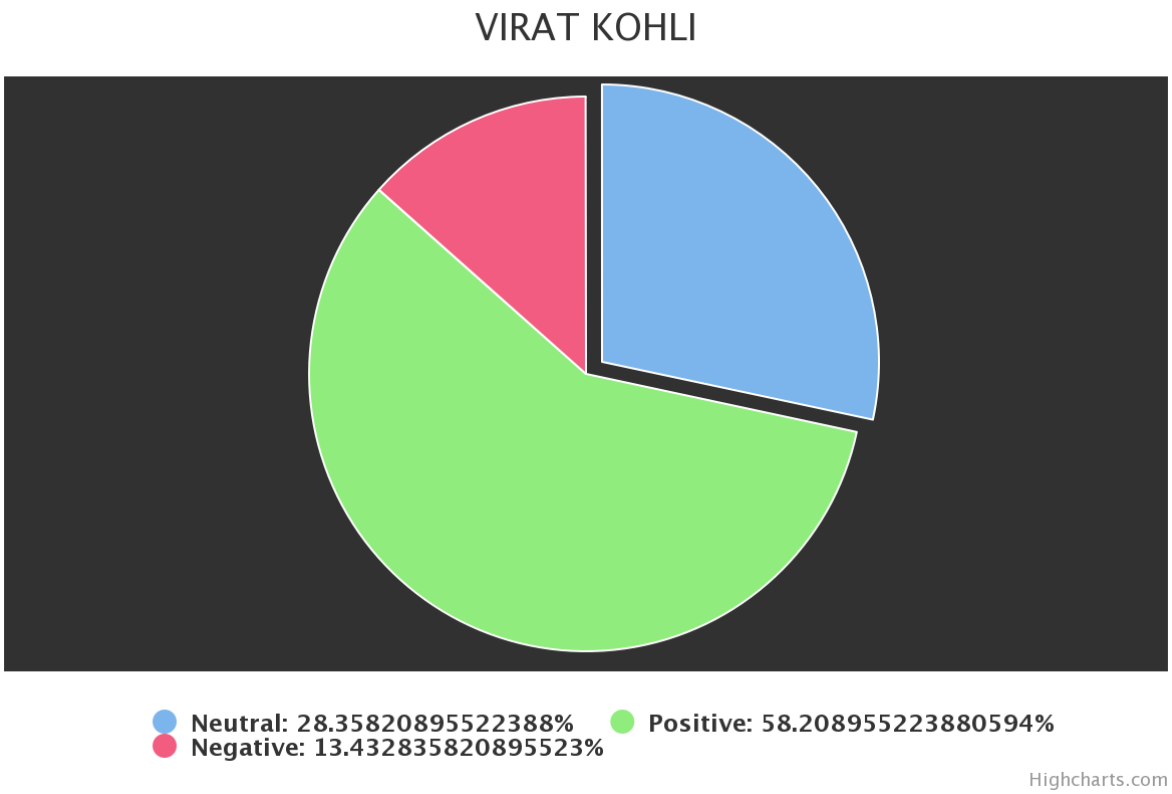
Software requirements

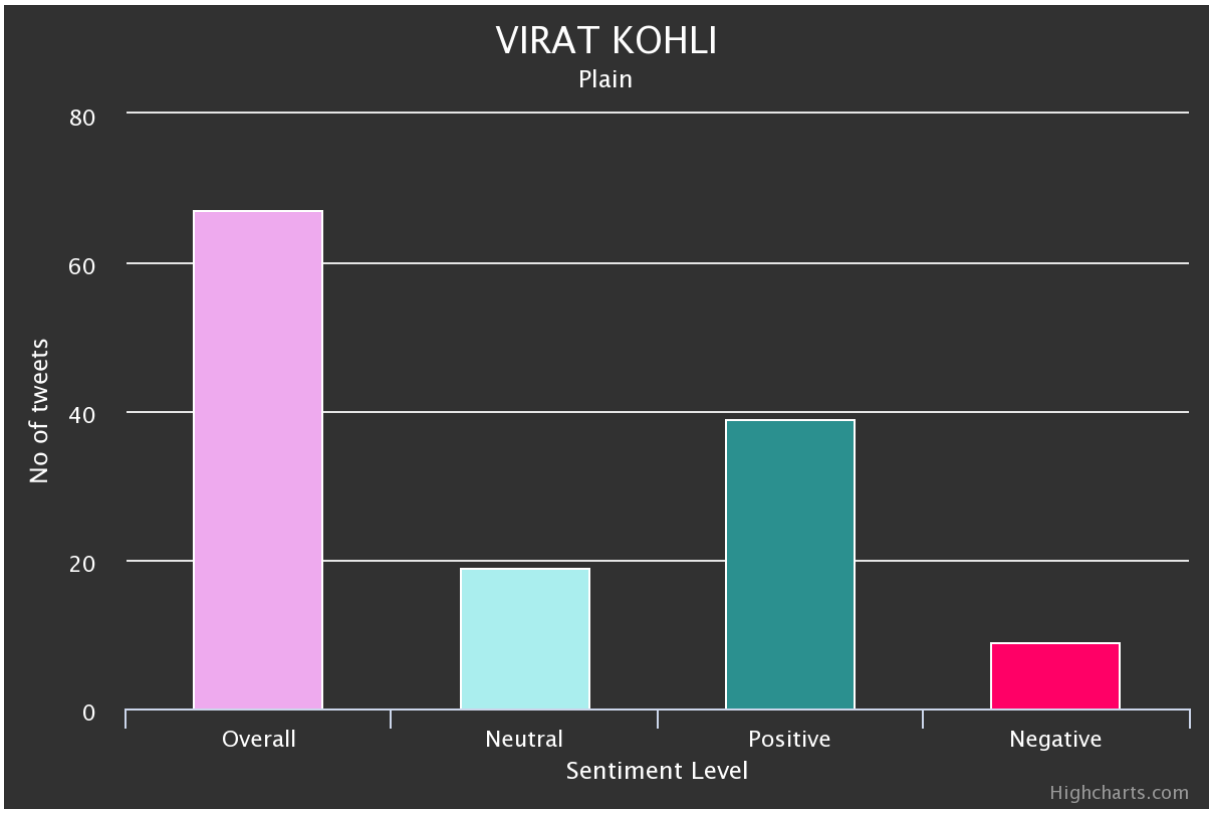
- Languages Used : Python , HTML, CSS .
- Librabies : Tweepy, NLTK(Natural Language ToolKit),NumPy,Scikit-Learn of python.
- Framework : Django
- Additional Requirements : Twitter API

Results

Overall , our result contains the pie-charts and bar diagram which shows the percentage of positive tweets ,negative tweets and neutral tweets in graphical form.
Also,we are displaying the 5 tweets from the positive tweets as well as 5 tweets from negative tweets as our result.

Illustration:
Suppose we have searched "virat kohli" in our query in the webpage.We obtained following graphs and analysis as the output in which the pie chart shows the positive,negative and neutral tweets percentage.




VIRAT KOHLI
Plain
80
60
40
20
0
No of tweets
Overall
Neutral
Positive
Negative
Sentiment Level
Highcharts.com

Conclusion

Nowadays, sentiment analysis or opinion mining is a hot topic in machine learning. We are still far to detect the sentiments of corpus of texts very accurately because of the complexity in the English language.

In this project we tried to show the basic way of classifying tweets into positive or negative category using Naive Bayes as baseline and how language models are related to the Naive Bayes and can produce better results. We could further improve our classifier by trying to extract more features from the tweets, trying different kinds of features, tuning the parameters of the naïve Bayes classifier, or trying another classifier all together.

Future Work

- We can use the current sentiment analysis along with the datasets so that we can predict the outcome of stock market,election,etc.
- Potential improvement can be made to our data collection and analysis method.
- Future work can be done with possible improvement such as more refined data and more accurate algorithm.

References

- Alec Go, Richa Bhayani and Lei Huang. Twitter Sentiment Classification using Distant Supervision. Project Technical Report, Stanford University, 2009
- Albert Biffet and Eibe Frank. Sentiment Knowledge Discovery in Twitter Streaming Data. Discovery Science, Lecture Notes in Computer Science, 2010, Volume 6332/2010, 1-15, DOI: 10.1007/978-3-642-16184-1
- Sentiment Analysis : http://nbviewer.jupyter.org/github/marclamberti/TwitterEmotionAnalysis/blob/master/TwitterSentimentAnalysis.ipynb
- Steven Bird, Even Klein Edward Loper. Natural Language Processing with Python.
- Twitter API : https://developer.twitter.com/en/apps
- Tweepy: http://docs.tweepy.org/en/v3.5.0/
- How To Build Twitter Sentiment Analyzer : https://www.ravikiranj.net/posts/2012/code/how-build-twitter-sentiment-analyzer/
- Tom M. Mitchell, generative and discriminative classifiers: Naive Bayes and Logistic Regression