

Title

Twitter Sentiment Analysis

Abhay Sastha S

Gunjan Pancholi

14th September, 2021



Problem Definition

Twitter sentiment analysis system is an easily customisable software that can be used to understand the user's(target group) sentiments on a particular topic through the tweets on twitter. The software should rank the tweets under 3 categories, negative, neutral and positive. Make use of python and its libraries such as re, stopwords, pandas, PorterStemmer, CountVectorizer for the backend and use Comma Separated Values (CSV) for storing data; make use of pandas library to connect the front end and the back end of the system. The output system should be made through Graphical Representation Interface for better User Interface/User Experience (UI/UX). To create the representation for the system make use of the matplotlib or plotly library of python. The system should also provide a graphical representation of the sentiments on the particular topic.



Objective

The main objective of the twitter sentiment analysis system is to create an efficient and effective system, with an easily accessible and understandable mechanism so as to provide better UI/UX for the user of the system. The system shall ease the problems faced in identifying the public opinion of a particular topic on the twitter forum. The system shall be designed in a format which makes it easy to customize the outputs for analysis. The usage of graphs to display the output will help the user/analysts to compare and comprehend the public opinions with ease.

Why Twitter?

Twitter has been chosen for this project as it provides us with reliable and varied opinions as it boasts a total of 206 million users on its platform. The platform is filled with personal, statistical, expert and governmental opinions which makes this the ideal choice to understand the overall sentiment of a topic through varied backgrounds, not limited to only but including differences like gender, age, culture, race etc. Twitter is the most successful microblogging service with 150 million daily users. 6,000 tweets are written every second. People tweet about everything that comes to mind and use hashtags to associate the tweet with a topic. Moreover the provision of the twitter api from Twitter itself makes it easy for the collection of reliable and real life data sets to work on. The usage of real life data sets will provide us with an unbiased dataset to work on which is important while creating a system such as sentiment analysis systems as these are based on opinions and opinions can be biased.



Scope of the Project

Why use Sentiment Analysis?

Business:

In marketing, companies use it to develop their strategies, to understand customers' feelings towards products or brands, how people respond to their campaigns or product launches and why consumers don't buy some products.

Politics:

In the political field, it is used to keep track of political views, to detect consistency and inconsistency between statements and actions at the government level. It can be used to predict election results as well!

Public Actions:

Sentiment analysis also is used to monitor and analyse social phenomena, for the spotting of potentially dangerous situations and determining the general mood of the blogosphere.



Structure of the Project

1. Extract tweets from twitter api and save as csv file
2. Remove unwanted value columns from the csv file
3. Create a list of positive words
4. Create a list of negative words
5. Create empty csv file that would store the in-process results temporarily
6. Data pre processing
 - a. Remove punctuations and symbols
 - b. Remove URLS and username tags
 - c. Remove numbers, hashtags, retweet symbol(RT),
 - d. Stopwords removal
 - e. Convert words to base words (stemming) and to lowercase
7. Create a nested list for each tweet in tweets with each inner list containing a tweet and the over all list containing all tweets
8. Compare each word in the tweet with list of all positive words and negative words and make count of them individually
9. Save the output onto the empty csv file
10. Read the csv file that stores the result and plot the values on to a graph.

Feasibility Study

This is an important component of this project as this study provides us with the information on how this given solution would be able to solve the given problem. The main objective is not to solve the problem but to acquire its scope. It shall focus on the following points -

1. Meet user requirements (Value Proposition feasibility)
2. Effective utilization of resources provided (Operational feasibility)
3. Cost and time efficient system (Economical feasibility)
4. Technically sound to avoid errors (Technical feasibility)

The study of the above mentioned points are as follows -

1. Value Proposition Feasibility-

The twitter sentiment analysis system provides the users with a range of features such as a graphical representation of the data, ease of usage and modification of the software, fast and robust mechanism that provides precise data and can handle huge datasets with ease.

2. Operational Feasibility-

Is there sufficient support for users? Is the current method acceptable to users? Will the proposed system cause any problem?


The Twitter Sentiment Analysis System is Operationally feasible. This application will provide necessary information to the users as to how the data regarding different operations to be performed on databases will be shown. The application will be planned in a way that no prior knowledge would be required to use it. The user will just need to have the basic knowledge of operating a computer.

3. Economical Feasibility-

Whether the new system will be cost effective or not? Will it be beneficial in term of cost reduction?

The project is economically feasible. The cost on the hardware is fairly low and the software being used i.e. python is a free open source software, where we are not spending any money. Moreover, the technical equipments are already available, so no further expenditure will be made to buy software packages.

4. Technical feasibility-



It refers to if the work for the project will be done with the current available equipments, i.e. existing H/W and S/W technology and available personnel. If the new technology is required, can it be developed?

For our project, we require a front end and back end s/w system, for which we are using python and CSV files respectively. Python is a free Open Source Software and is available to us. We also have an adequate amount of hardware for running python on our system. So it is technically feasible to develop this project.



Methodology

Methodology will include the steps to be followed to achieve the objectives of the project during the project development.

- Scope of the project
- Reporting on project status
- Financial planning
- Resource Planning
- Quality Control Plan
- Risk Management Plan



Dataset



Tools and Platforms Used

Recommended System Requirements

-->Processors:-Intel® Core™ i3 processor 4300M at 2.60 GHz

-->Disk Space:- 2 to 4 GB

-->Operating System:- Windows10, MACOS, UBUNTU

-->Python Versions:- 3.X.X or higher

Bibliography

CODE SAMPLE DRAFT

```
projectTwitterDataFile = open("C:/Users/MSI/Desktop/project_twitter_data.csv", "r")
resultingDataFile = open("C:/Users/MSI/Desktop/resulting_data.csv", "w")
```

```
import pandas as pd
csv1 = pd.read_csv("C:/Users/MSI/Desktop/project_twitter_data.csv")
print(csv1.head())
```

```
punctuation_chars = ["'", '"', ",", ".", "!", ":", ";", '#', '@']
# lists of words to use
positive_words = []
with open("C:/Users/MSI/Desktop/positive_words.txt") as pos_f:
    for lin in pos_f:
        if lin[0] != ';' and lin[0] != '\n':
            positive_words.append(lin.strip())
```

```
def get_pos(strSentences):
    strSentences = strip_punctuation(strSentences)
    listStrSentences = strSentences.split()

    count = 0
    for word in listStrSentences:
        for positiveWord in positive_words:
            if word == positiveWord:
                count += 1
    return count
```

```
negative_words = []
with open("C:/Users/MSI/Desktop/negative_words.txt") as pos_f:
    for lin in pos_f:
        if lin[0] != ';' and lin[0] != '\n':
            negative_words.append(lin.strip())
```

```
def get_neg(strSentences):
    strSentences = strip_punctuation(strSentences)
    listStrSentences = strSentences.split()

    count = 0
    for word in listStrSentences:
```

```

        for negativeWord in negative_words:
            if word == negativeWord:
                count += 1
    return count

def strip_punctuation(strWord):
    for charPunct in punctuation_chars:
        strWord = strWord.replace(charPunct, "")
    return strWord

def writeInDataFile(resultingDataFile):
    resultingDataFile.write("Number of Retweets, Number of Replies, Positive Score,
Negative Score, Net Score")
    resultingDataFile.write("\n")

    linesPTDF = projectTwitterDataFile.readlines()
    headerDontUsed = linesPTDF.pop(0)
    for linesTD in linesPTDF:
        listTD = linesTD.strip().split(',')
        resultingDataFile.write(
            "{}, {}, {}, {}, {}".format( listTD[1], listTD[2], get_pos(listTD[0]),
get_neg(listTD[0]), (get_pos(listTD[0]) - get_neg(listTD[0]))))
        resultingDataFile.write("\n")

writeInDataFile(resultingDataFile)
projectTwitterDataFile.close()
resultingDataFile.close()

csv2 = pd.read_csv("C:/Users/MSI/Desktop/resulting_data.csv")
print(csv1.head())

```