Analyzing Twitter data using Big Data Analysis

BY: MEGHA SAXENA (20BIT0366)

OBJECTIVE

The main objective of this project is to focus on how data generated from Social Media can be stored and utilized by different companies to make targeted, real time and informed decisions about their product that can increase their market share. This can be done by using various algorithms. The given project will focus on how data generated from Social Media Websites can be analyzed and utilized.

Companies can use this project to understand how effective and penetrative their marketing programs are. In addition to the view counts, subscribers and shares, audience retention count, companies can also evaluate views according to date range.

ABSTRACT

Online social network platforms, with their large-scale repositories of user-generated content, can provide unique opportunities to gain insights into the emotional “pulse of the nation”, and indeed the global community. A great source of unstructured text information is included in social networks, where it is unfeasible to manually analyze such amounts of data. There is a large number of social networks websites that enable users to contribute, modify and grade the content, as well as to express their personal opinions about specific topics.

This project can help in analyzing new emerging trends and knowing about people's changing behavior with time. In addition, people in different countries have different preferences. By analyzing the comments/feedbacks/likes/view counts etc. of the videos, images uploaded, companies can understand what are the likes/dislikes of people around the world and work on their preferences accordingly.

INTRODUCTION

Social media is a web-based and mobile-based internet application that will allow the creation, access and exchange of user-generated content that is ubiquitously accessible. Besides social networking media like Twitter and Facebook, the term “social media” to encompass really simple syndication (RSS) feeds, blogs, wikis and news, all typically yielding unstructured text and accessible through the web. Social media is especially important for research into computational social science that investigates questions using quantitative techniques for example, computational statistics, machine learning and complexity and so-called big data for data mining and simulation modeling .

Social media has led to numerous data services, tools and analytics platforms. The tools available to researchers are either give superficial access to the raw data or non-superficial access. Researchers require to program analytics in a language such as Java. So the proposed work is much better than the available ones with respect to cost, efficient handling Big Data and scalability.

The analytics persons and businesses feel the need to gain new insights from social media; they require the analytics tools and expertise to transform this big data information which will have big volume and variety into the respective strategies so as to draw certain conclusions. Social media analytics is useful tool for getting details of customer sentiments that are distributed across online sources.

Literature survey

1. B.PrashanthMruthyunjayaMenduRavikumarThallapalli ,"Cloud based Machine learning with advanced predictive Analytics using Google Colaboratory" ,2021
2. L. B. and A. Cutler, "Random forests - copyright."
3. MasanoriKuroki , "Using Python and Google Colab to teach undergraduate microeconomic theory",2020
4. JPraveen Gujjar[a]H RPrasanna Kumar[b]Niranjan N.Chiplunkar[c] , "Image classification and prediction using transfer learning in colab notebook", 2021
5. L. Dey, S. Chakraborty, A. Biswas, B. Bose, and S. Tiwari, "Sentiment Analysis of Review Datasets using Naïve Bayes' and K-NN Classifier."
6. Y. Wan and Q. Gao, "An Ensemble Sentiment Classification System of Twitter Data for Airline Services Analysis," 2015

PROBLEM STATEMENT

With all the real time data collected over period of time, the system will analyze and draw meaningful inferences from the collection of tweets. Proposed system will analyze tweets data from many perspectives to make meaningful inferences. Trend analysis, sentiment analysis volume analysis are major parts of proposed system. In trend analysis, system will try to find trending discussions, parties, personalities throughout the period of time. From literature, Kmeans is more suitable algorithm for clustering of tweet data and to find trends. Volume analysis of tweets will give idea of popularity of particular topic or person over a period of time. Volume analysis with respective to geo-location and date will help to make certain conclusions. Sentiment analysis of tweets will help draw conclusion for political orientation of overall users respective to political parties, topics.

RELATED PROJECT

Trend analysis and based on that predicting public opinions. It plays important role, many researcher working on automatic technique of extraction and analysis of huge amount of twitter data. In Sensing Trending Topics in Twitter author compare six trend detection method and find that standard natural language processing technique perform well for social streams on particular topic. They conclude that n-gram give best performance other than state-of-art techniques. In Comparison of Machine Learning Approaches on Arabic Twitter Sentiment Analysis , the authors have used three different machine learning algorithms Naïve Bayes, Decision Trees and Support Vector Machine for sentiment classification of Arabic dataset which was obtained from twitter. This research has followed a framework for Arabic tweets classification in which two special sub-tasks were performed in pre-processing, Term Frequency-Inverse Document Frequency (TF-IDF) and Arabic stemming. They have used one dataset with three algorithms and performance has been evaluated on the basis three different information retrieval metrics precision, recall, and f-measure.

PROPOSED SYSTEM

- Cleaning of data

- ➢ Tweets contain many slang words and punctuation marks. We need to clean our tweets before they can be used for training the machine learning model. Cleansing data will reduce noise in the tweet data. Unimportant words will be removed such as URL, hashtag (#), username (@username), email, emoticons (: @,: *,: D), (,), dot (.) and also other punctuation

- *Case folding*

- ➢ This stage serves to change letters character in the comments into all lowercase letters characters. In social media, especially Twitter, writing tweets, there must be differences in the shape of letters, case-folding stages is a changing process the shape to lowercase letters (lower case) or can also be called uniformity of letters

- *Tokenizing*

- ➢ Tokenizing or parsing stage is the cutting stage of the input string based on each word arrange. In principle, this process is to separate every word that composes a document.

- *Stemming*

- ➢ *Stemming* is the stage to make the word affixes into basic words. In stemming, conversion of morphological forms of a word to its stem is done assuming each one is semantically related

Tools used: Google colab

Colab is a free notebook environment that runs entirely in the cloud. It lets you and your team members edit documents, the way you work with Google Docs. Colab supports many popular machine learning libraries which can be easily loaded in your notebook.

This tutorial gives an exhaustive coverage of all the features of Colab and makes you comfortable working on it with confidence

Language used: python

Python is a high-level, general-purpose programming language. Its design philosophy emphasizes code readability with the use of significant indentation. Its language constructs and object-oriented approach aim to help programmers write clear, logical code for small- and large-scale projects.

Python is dynamically-typed and garbage-collected. It supports multiple programming paradigms, including structured (particularly procedural), object-oriented and functional programming. It is often described as a "batteries included" language due to its comprehensive standard library.

Guido van Rossum began working on Python in the late 1980s as a successor to the ABC programming language and first released it in 1991 as Python 0.9.0. Python 2.0 was released in 2000 and introduced new features such as list comprehensions, cycle-detecting garbage collection, reference counting, and Unicode support. Python 3.0, released in 2008, was a major revision that is not completely backward-compatible with earlier versions. Python 2 was discontinued with version 2.7.18 in 2020.

Python consistently ranks as one of the most popular programming languages.

CODE

```python
#Description : This is a sentiment analysis program that parses the tweets fetched from Twitter using Python


#import the libraries
import tweepy
from textblob import TextBlob
import pandas as pd
import numpy as np
import re
import matplotlib.pyplot as plt
plt.style.use('fivethirtyeight')


# Load the data
from google.colab import files
uploaded = files.upload()


#Get the data
log = pd.read_csv('Login.csv')


# Twitter API credentials
consumerKey = log['CLIENT ID'][0]
consumerSecret = log['CLIENT ID SECRET'][0]
accessTokenKey = log['API TOKEN'][0]
```

```
accessTokenSecret = log['API SECRET KEY'][0]



#Create the authentication object
authenticate = tweepy.OAuthHandler(consumerKey,consumerSecret)

#Set the access token and access token secret
authenticate.set_access_token(accessTokenKey, accessTokenSecret)

#Create API object while passing i the auth info
api = tweepy.API(authenticate, wait_on_rate_limit = True)



#Extract 2320 tweets from the twitter API
postVA = api.user_timeline(screen_name ="VirginAtlantic", count=2320, lang="en", tweet_node="extended")
postU = api.user_timeline(screen_name="united", count=2320, lang="en", tweet_node="extended")
postSA = api.user_timeline(screen_name="southwestAir", count=2320, lang="en", tweet_node="extended")
postD = api.user_timeline(screen_name="Delta", count=2320, lang="en", tweet_node="extended")
postA = api.user_timeline(screen_name="LATAMAirlinesUS", count=2320, lang="en", tweet_node="extended")
postAA = api.user_timeline(screen_name="AmericanAir", count=2320, lang="en", tweet_node="extended")



#Create a dataframe with a column called tweets
dfVA = pd.DataFrame([tweet.full_text for tweet in postVA], columns=['TweetsVA'])
dfVA.head()
dfU = pd.DataFrame([tweet.full_text for tweet in postU], columns=['TweetsU'])
dfU.head()
dfSA = pd.DataFrame([tweet.full_text for tweet in postSA], columns=['TweetsSA'])
dfSA.head()
dfD = pd.DataFrame([tweet.full_text for tweet in postD], columns=['TweetsD'])
dfD.head()
dfA = pd.DataFrame([tweet.full_text for tweet in postA], columns=['TweetsA'])
dfA.head()
```

```
dfAA = pd.DataFrame([tweet.full_text for tweet in postAA], columns=['TweetsAA'])
dfAA.head()
df=dfVA.dfU.dfSA.dfD.dfA.dfAA
df['Tweets'] = df['TweetsVA']df['TweetsU']df['TweetsSA']df['TweetsD']df['TweetsA']df['TweetsAA']
```

```
#Clear the text
#Create a function to clean the tweets
def cleanTxt(text):
  text =re.sub(r'@[A-Za-z0-9]+', '', text) #removes @
  text =re.sub(r'#', '', text) #remove #
  text =re.sub(r'*', '', text) #remove *
  text =re.sub(r'.', '', text) #remove .
  text =re.sub(r'https?:\/\/\S+', '', text) #removes hyperlinks

  return text

  #Cleaning the txt
  dfVA['TweetVA']=df['TweetsVA'].apply(cleanTxt)
  dfU['TweetU']=df['TweetsU'].apply(cleanTxt)
  dfSA['TweetSA']=df['TweetsSA'].apply(cleanTxt)
  dfD['TweetD']=df['TweetsD'].apply(cleanTxt)
  dfA['TweetA']=df['TweetsA'].apply(cleanTxt)
  dfAA['TweetAA']=df['TweetsAA'].apply(cleanTxt)
```

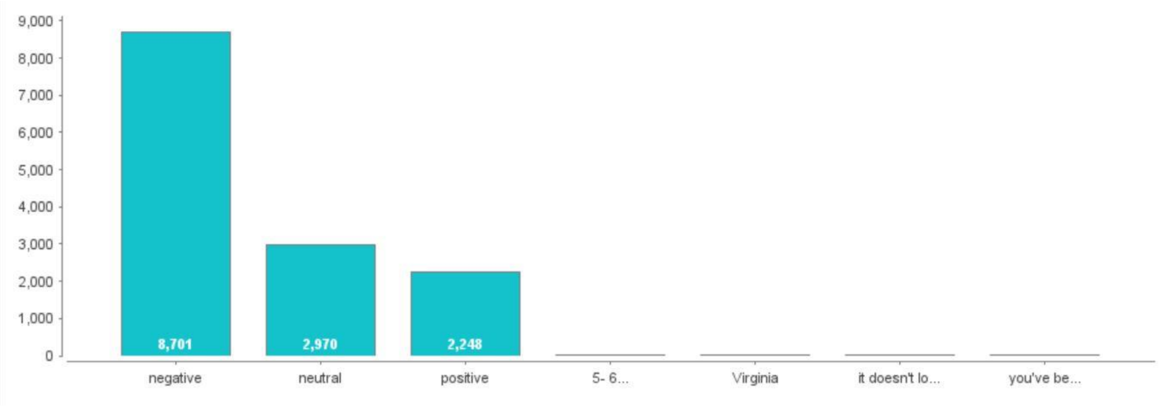
```
#Create a function to get the subjectivity
def getSubjectivity(text):
  return TextBlob(text).sentiment.subjectivity

  def getPolarity(text):
  return TextBlob(text).sentiment.polarity

  #Create two new columns
  df['Subjectivity'] = df['Tweets'].apply(getSubjectivity)
  df['Polarity'] = df['Tweets'].apply(getPolarity)
```

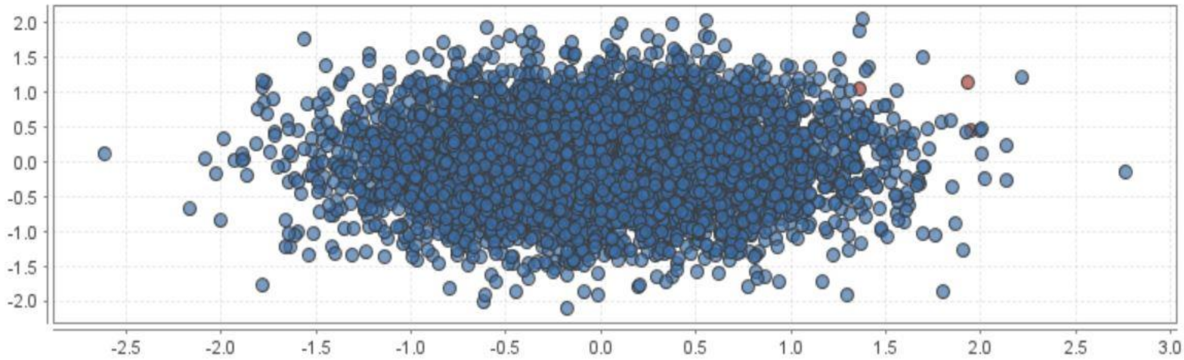
```
#Create a function to comput the negative, neutral and positive analysis
```

```
def getAnalysis(score):
  if score<0:
    return 'negative'
    elif score==0:
    return 'nneutral'
    if score>0:
    return 'positive'

    df['Analysis'] = df['Polarity'].apply(getAnalysis)



#Positive Tweets
j=1
sortedDF =df.sort_values(by=['Polarity'])
for i in range(0,sortedDF.shape[0]):
  if(sortedDF['Analysis'][i]=='Positive'):
  print(str(j)+')'+sortedDF['Tweets'][i])
  print()
  j=j+1


#Negative Tweets
j=1
sortedDF =df.sort_values(by=['Polarity'],ascending='False')
for i in range(0,sortedDF.shape[0]):
  if(sortedDF['Analysis'][i]=='Negative'):
  print(str(j)+')'+sortedDF['Tweets'][i])
  print()
  j=j+1


#Plot the polarity and subjectivity
plt.figure(figsize=(8,6))
for i in range(0,df.shape[0]):
  plt.scatter(df['Polarity'][i],df['Subjectivity'][i], color='blue' , c
olor='red')
  plt.xlabel['Polarity']
  plt.ylabel['Subjectivity']
  plt.show()



#Show value count
df['Analysis'].value_counts()

#plot and visualize the count
plt.xlabel('Sentiment')
plt.ylabel('Counts')
```

```
df['Analysis'].value_counts().plot(kind='bar')
plt.show()
```

OUTPUT





DATASET

- The data set is taken from Twitter API.
- It is taken about 6 US airlines namely Virgin American, United, Southwest, American, Delta and US Airlines.
- The data is taken using Twitter developer app.

CONCLUSION

Sentiment analysis or opinion mining is a field of study that analyzes people's sentiments, attitudes, or emotions towards certain entities. This paper tackles a fundamental problem of sentiment analysis, sentiment polarity categorization. Tweets about six airline data from kaggle.com are selected as data used for this study. We performed sentiment analysis using python code and google colab.