Twitter Sentiment Analysis

Achyuth Pilly

Table of Contents

What is Sentiment Analysis?

Sentiment Analysis is the process of ‘computationally’ determining whether a piece of writing is positive, negative or neutral. It’s also known as opinion mining, deriving the opinion or attitude of a speaker.

Why Twitter?

Twitter is a microblogging website. Every second, on average, around 6,000 tweets are tweeted on Twitter.

It is utilized by individuals to express their feelings and show estimations on various events.

How's it different from other approaches?

- We are analyzing the streaming data on the go, not storing the tweets in any database or file. By this you can get the sentimental trends on the current feelings of the users.
- Dynamically user can give the keyword/hashtag and number of tweets on which topic sentiment of tweets has to analyzed and on how many tweets.
- Visualizing the trends of the sentiments, which could make easier to make decision on the trends.

How Sentiment Analysis Works?

1. Using a vocabulary: looking at important keywords (usually verbs and adjectives) along with modifiers like negation words

2. Using rules: look at presence of vocabulary words in sentences, use rules to categorise them by sentiment

3. Applying ML techniques: treat this as a classification problem, amass a dataset and set up features (which could be the keywords in the vocabulary), and train to identify sentiment

Approaches in Sentiment Analysis

There are mainly three approaches in sentiment analysis

1. Lexicon based - considers lexicon dictionary for identifying polarity of the text
2. Machine learning based approach - Needs to develop classification model, which is trained using prelabeled dataset of positive, negative, neutral content.
3. Combined approach - Which uses lexicon dictionary along with pre-labelled data set for developing classification model.

Our Approach using Naive Bayes ML Method

Sentimental analysis will be performed after the data i.e., tweets are collected from the source (here twitter).

The collected tweets need some preprocessing like cleaning, removing emoticons and unnecessary words. So that our tweets can be categorized into positive, negative and neutral.

Tweet Collection

Tweet collection includes gathering applicable tweets about the specific zone of interest.

We are using Twitter's API for collecting large number of tweets.

For the analysis of tweets, we make a python script which will be utilized to bring tweets from twitter.

Preprocessing Tweets

The preprocessing of the tweets is a critical step as it chooses the effectiveness of alternate steps.

In pre-handling we first concentrate our principle message from the tweet, at that point we evacuate every unfilled space, stops words, hash labels, rehashing words, URL's and so on.

Cleaning twitter information is vital, since tweets contain a few syntactic highlights that may not be valuable for examination.

Feature Extraction

1. Unigram : includes gathering all words in the record and utilizing them as highlights
2. Bigrams : are highlights comprising of sets of two neighboring words in a sentence.
3. Part-of-speech tagging in semantics and data recovery is the way toward labeling each word in a sentence to a specific grammatical feature.

Sentiment Classifiers

- Bayes' theorem is formally expressed by the following equation.

$$P(A|B) = \frac{P(B|A) \times P(A)}{P(B)}$$

- where P(A) and P(B) ar nd B without regarding each other. P(A | B) is the probability of A conditional on B and P(B | A) is the probability of B conditional on A. In naïve Bayes classification, A is categorical outcome events and B is a series of predictors. The word "naïve" indicates that the predictors are independent on each other conditional on the same outcome value.

Sentiment Analysis using Python

Python

Python is a high-level programming language extensively used in the field of data science.It is interpreted, interactive and object-oriented scripting language.

Python provides a vast library which can be used for various applications for example natural language processing, machine learning, data analysis etc.

Natural Language Toolkit

Natural Language Toolkit is a Python library for Natural Language Processing.

Tokenization, tagging, filtering,text manipulation operations can be performed using NLTK.

It embodies various trainable classifiers into itself like Naïve Bayes Classifier.

Tweepy

Tweepy is a python client for the official Twitter API.We can install it through pip.

pip install tweepy

Tweepy supports oauth authentication.

Textblob

TextBlob is a python library built over top of NLTK and offers a simple API to perform NLP tasks.

It is used for processing textual data.

It is easy to learn and offers a lot of features like sentiment analysis, pos-tagging, noun phrase extraction, etc.

Output

- **Result 1:** When the code is executed, it asks for keyword/tag to search in Tweets for categorizing the tweets into positive, negative and neutral.

```
Python 3.6.4 |Anaconda, Inc.| (default, Jan 16 2018, 10:22:32)
[MSC v.1900 64 bit (AMD64)]
Type "copyright", "credits" or "license" for more information.

IPython 6.2.1 -- An enhanced Interactive Python.

In [1]: runfile('E:/college/Term_02/Raed/2033/
Project_Sentiment_Analysis/Main_dev.py', wdir='E:/college/
Term_02/Raed/2033/Project_Sentiment_Analysis')
Connection is Successful!!!

Enter Keyword/Tag to search about:
```

- **Result 2:** We need to enter the keyword. For example: Data Analytics

```
Python 3.6.4 |Anaconda, Inc.| (default, Jan 16 2018, 10:22:32)
[MSC v.1900 64 bit (AMD64)]
Type "copyright", "credits" or "license" for more information.

IPython 6.2.1 -- An enhanced Interactive Python.

In [1]: runfile('E:/college/Term_02/Raed/2033/
Project_Sentiment_Analysis/Main_dev.py', wdir='E:/college/
Term_02/Raed/2033/Project_Sentiment_Analysis')
Connection is Successful!!!

Enter Keyword/Tag to search about: Data Analytics
```

- **Result 3:** After entering the keyword, we need to enter the number of tweets on which keyword will be searched.

```
Python 3.6.4 |Anaconda, Inc.| (default, Jan 16 2018, 10:22:32)
[MSC v.1900 64 bit (AMD64)]
Type "copyright", "credits" or "license" for more information.

IPython 6.2.1 -- An enhanced Interactive Python.

In [1]: runfile('E:/college/Term_02/Raed/2033/
Project_Sentiment_Analysis/Main_dev.py', wdir='E:/college/
Term_02/Raed/2033/Project_Sentiment_Analysis')
Connection is Successful!!!

Enter Keyword/Tag to search about: Data Analytics

Enter how many tweets to search: 1500
```

- **Result 4-1:** Percentage of Positive, Negative and Neutral tweets are displayed.

```
Python 3.6.4 |Anaconda, Inc.| (default, Jan 16 2018, 10:22:32)
[MSC v.1900 64 bit (AMD64)]
Type "copyright", "credits" or "license" for more information.

IPython 6.2.1 -- An enhanced Interactive Python.

In [1]: runfile('E:/college/Term_02/Raed/2033/
Project_Sentiment_Analysis/Main_dev.py', wdir='E:/college/
Term_02/Raed/2033/Project_Sentiment_Analysis')
Connection is Successful!!!

Enter Keyword/Tag to search about: Data Analytics

Enter how many tweets to search: 1500
Positive tweets percentage is: 75.00 %
Negetive tweets percentege is: 8.33 %
Neutral tweets percentage is: 16.67 %
```

- **Result 4-2:** Five positive tweets are displayed on the output console as follows:

```
Positive tweets:

RT @Harry_Robots: #AI is “available anytime, polite and faster”.
#ArtificialIntelligence &gt;&gt; #Infographics &gt;&gt; via
@adrian_drones @ravikikan…
RT @INM7_ISN: Looking for two PhD students for exciting new
projects combining multi-modal imaging and machine-learning.

Join is to combi…
RT @Ronald_vanLoon: 12 #Python Resources for #DataScience
by @granvilleDSC @DataScienceCtrl |

Read more at https://t.co/MxaXSUXaBo

#BigDa…
RT @JCullenNow: There are some great new features in #ArcGIS
Business Analyst like the new population by generation #data
update. I did a q…
RT @Zac_Urback: For the 3rd straight year the @OHLSteelheads are
looking to expand our analytics department! If you are
interested in data…
RT @Sophia_May_00: #AI is “available anytime, polite and
faster”. #ArtificialIntelligence &gt;&gt; #Infographics &gt;&gt;
via @adrian_drones @ravikikan…
RT @DataAnalytics_1: #AI is “available anytime, polite and
faster”. #ArtificialIntelligence &gt;&gt; #Infographics &gt;&gt;
via @adrian_drones @ravikik…
RT @Ronald_vanLoon: #BI Success Beyond #DataVisualization
by @CIOwhitepapersreview @Sisense |

Read more at https://t.co/UozW0o8rBx

#DataV…
Get Smart with Pillow’s Data, Visibility and Control Tools
https://t.co/gRSvPTWsuQ #Analytics #data https://t.co/6zr4HuiLq3
```

- **Result 4-3:** Neutral tweets are displayed. If tweets are less than 5 than all of them will be displayed on output console.

```
Neutral tweets:

RT @BerniceRogowitz: Needed: teaching "associate"  for a Vis
course in the Columbia University Applied Analytics Program this
fall. I wrote…
Business Analytics News: Data Analytics Trends - Blog - Stay
Informed | MicroStrategy https://t.co/vH8y4lQBNo, see… https://
t.co/yMUagLRfP9
```

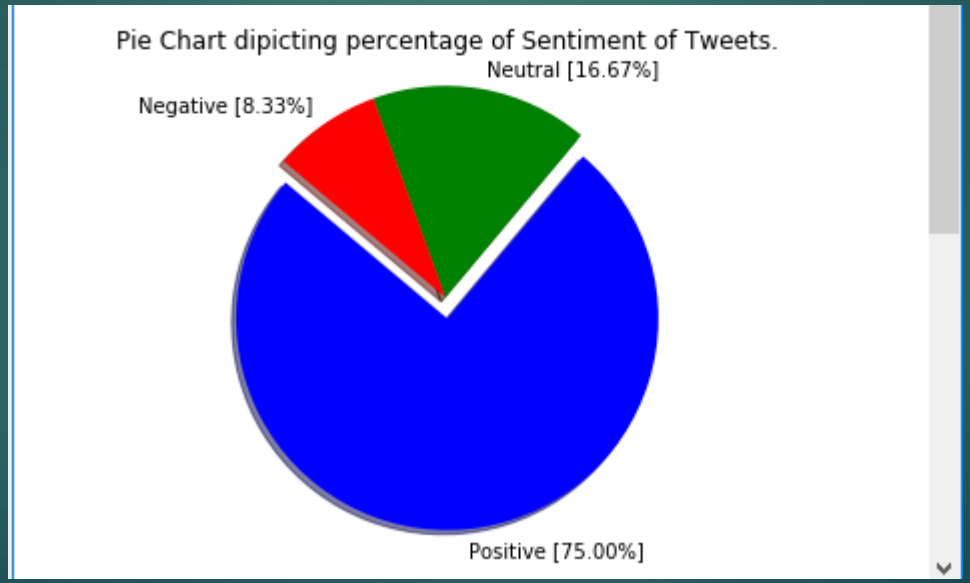
- **Result 4-4:** Negative tweets are displayed on the output console.

```
Negative tweets:

RT @ashwinmegha: Big Data: Numbers mean nothing without context:
https://t.co/bQPrBytsWo

#SoftwareTesting #news #SoftwareDevelopment #outs…
```

- **Result 4-5:** Pie chart is also displayed on the output console. Blue part depicts Positive tweets, Red depicts Negative tweets and Green part depicts Neutral tweets.



THANK YOU!