Data Analysis On Twitter Data

Krunal P Bhatt -- 1MS12CS048
Nishant Washisth -- 1MS12CS069
Raghav Chawla -- 1MS12CS079
Rajat Jain -- 1MS12CS085

Computer Science & Engineering
M S Ramaiah Institute Of Technology

Abstract: Twitter, one of the largest social media site receives tweets in millions every day. This huge amount of raw data can be used for industrial or business purpose by organizing according to our requirement and processing. This paper provides a way of analysing tweeter data using python programming.

The main purpose of this project is to collect large data from the social networking site and analyse this data which can depict the public sentiments and emotions that tells the nature of public .The aim of this project is to come up with a good result that helps the public from the analysis of user tweets. This includes identifying the locations where the tweets are made more and differentiating between positive, negative and neutral tweets.

Keywords: Twitter, Sentiment Analysis, Analysis of data, Machine Learning, Naive Byes Algorithm.

I. Introduction

The rise of micro blogging services like Twitter has spawned great interest in these systems as human-powered sensing networks. Since its creation in 2006, Twitter has experienced an exponential explosion in its user base, reaching a lot of people across the globe. So analysis on the twitter data will help the people in knowing the mindset of the people around us.

Hence in this project we are doing an overall analysis of the twitter dataset which will classify the tweets as positive, negative or neutral. The main technologies used in this are NLTK, and machine learning algorithms to identify the sentiment The detailed design and implementation is explained below. Here we concentrate on the analysing the dataset and visualising the output using pie chart which is done using python.

Design

a. ***Number of Modules***: *4*

b. ***Modules Description***:

1).Fetching and Extracting Data:

The data set contains training set and test set. The training set contains 125432 twitter users and 229,677 tweets from the users. The test set contains 5050twitter user and 145098 tweets from the users. All the locations of users are uploaded from their smart phones smart phones with the form of "UT: Latitude , Longitude".

2). Data PreProcessing:

The data is in JSON encoded format which makes it easier to extract the important data.

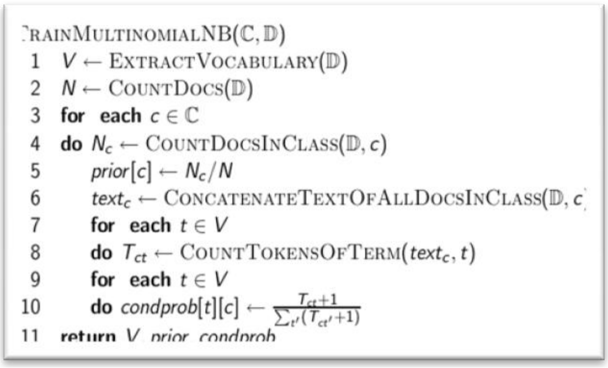
3). Classification:

Tweet classification will be performed on the user Tweet to determine the nature of the Tweet relative to the geo location. The data analysis will provide a negative, or positive value or numeric values.

4). Analysis:

The Pre-processed data is considered for analysis. Based on the number of positive and negative words in the user tweets, the tweet is classified as positive, negative tweets and the same is displayed to the user.
In the naive-bayes algorithm a training set is given to the program after which it classifies all words into 2 categories and the results are calculated and displayed for the user.

2). Naive bayes algorithm (classification)



```
TrainMultinomialNB(ℂ, 𝔻)
V ← ExtractVocabulary(𝔻)
N ← CountDocs(𝔻)
for each $c \in ℂ$
do $N_c$ ← CountDocsInClass(𝔻, c)
    $prior[c] \leftarrow N_c / N$
    $text_c$ ← ConcatenateTextOfAllDocsInClass(𝔻, c
    for each $t \in V$
    do $T_{ct}$ ← CountTokensOfTerm($text_c$, t)
    for each $t \in V$
    do $condprob[t][c] \leftarrow \frac{T_{ct}+1}{\sum_{t'}(T_{ct'}+1)}$
return V, prior, condprob
```

II. Implementation

The Project is implemented as follows:

- Twitter Data Set Identification and Extraction
- Pre-processing of data set
  - Removing unwanted attributes
  - Eliminating stop words and URL's
- Tweet classification
  - Positive
  - Negative
  - Neutral

Our Approach In Implementation:

In our approach we focused more on the speed of performing analysis than its accuracy i.e. performing sentiment analysis on big data which is achieved by splitting the various modules of data in following steps and collaborating with each other. This tagging is used for following various purposes.

i. **Stop words removal:**

The stop words like a, an, this which are not useful in performing the analysis are removed in this phase. Stop words are removed using java in eclipse platform. All the words are not considered are not considered for analysis.

ii. **Unstructured to structured**:

Twitter comments are mostly unstructured i.e. 'aswm' is written 'awesome', 'happyyyyyyy' to actually 'happy'. Conversion to structured is done by dynamic data records of unstructured to structured and vowels adding.

iii. **Stemming**:

In stemming we use porter's algorithum to remove suffix and prefixes from the tweeter data set file and normalise it.

III. Testing and Comparision

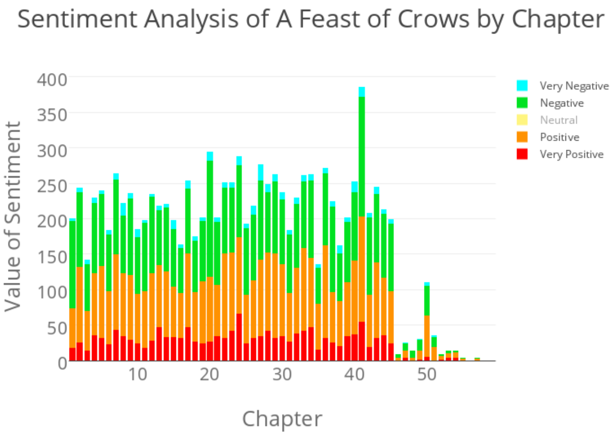
In this project the environment in which this project has been built is python, hence all the processing and testing will be done on the python IDLE interface. Since Machine Learning algorithm like Naïve bayes is used in this project, the maximum amount of testing will be done to test how well does the system diagnose to classify the whole data set into different nature of tweets.

The project is implemented using python. Our project is unique compared to the earlier projects which are done on the same topic.

IV. Results

After streaming the tweets into HDFS in real time, Naïve bayes is used in analyzing the tweets. Tweets are tagged as documents where categories are the hash tags defined in the Flume configuration file. Later the tweets are grouped as positive, negative and neutral based on subjectivity corpus forming a dictionary of words and its polarity. The graph shown before is only as an example and not done by our application.

IV.GRAPHICAL REPRESENTATION OF NATURE OF A SINGLE USER TWEET



This is a sample count for the US 2016 elections based on twitter sentiment

Bernie Sanders:

Opinion	Count
Positive	313
Negative	152
Neutral	555

Donald Trump:

Opinion	Count
Positive	54
Negative	500
Neutral	245

V. Time Efficiency

Time efficiency is an important aspect where our project scores well. Lower response time has achieved by use of data structures as local variables. This reduces the access time from a hard-disk. Also the use of Python and twitter streaming API lowers the access time. Hence overall the time efficiency increases owing to the above mentioned factors.

VI. Scope and Future work

Twitter has lot of scope in the modern era. Twitter as a social media has many users and the numbers of users are increasing day by day. At this moment, the code can handle the analysis part with a very good accuracy. But there are a few areas which have a lot of scope in this aspect. Sarcastic comments are the ones which are very difficult to identify. Tweets containing sarcastic comments give exactly opposite results owing to the mindset of the author. These are almost impossible to track. Also depending on the context in which a word is used, the interpretation changes. For ex: the word ‘unpredictable’ in ‘unpredictable plot’ in context of a land plot is negative whereas ‘unpredictable plot’ in context of a movie’s plot is positive. So it’s important to relate the interpretation with the context of the tweets. Also the use of native language combined with English usage is difficult to interpret.

Nowadays big data has become the buzzword in IT industry organizations. The need of analysing and processing of information has grown a lot. This paper implemented the analysing of big data (tweets) only for text. Further analysis can be done to images and all types of multimedia files based on index support. The result of Text mining and data analysis would help in suggesting related pages based on different types of data. So that industries make the data easily available to people who is using and trying accessing such type of data.

VII. Conclusion

This project gave us hands on experience of handling and parallel processing of huge amount of data. Data collection process introduced us to twitter streaming API. It was very interesting to gather and then aggregate the social networking data so as to extract interesting patterns and recent trends from it. We got exposure to work with prominent analytics tool: twitter streaming API is gaining significant momentum from both industry and academia as the volume of data to analyze growth rapidly.

This project helped us not only to gain knowledge about Python and Twitter streaming API but also Naïve Bayes programming model.. Amongst the many fields of analysis, there is one field where humans have dominated the machines more than any – the ability to analyze sentiment, or sentiment analysis.

ACKNOWLEDGMENT

First and foremost, my utmost Gratitude to our HOD, Dr. K G Srinivas who gave us this opportunity to complete this project. He had kind concern and considerations regarding project work which helped us to successfully complete the project.

Secondly we would like to thank Sowmya B J, Assistant Professor, MSRIT and Chandrika Prasad, Assistant Professor, MSRIT who reviewed our project and gave us a valuable guidance which helped us in completing this project.

This work would not have been possible without the guidance and help of several individuals who in one way or another contributed their valuable assistance in preparation and completion of this study.

We would like to express sincere thanks to all the teaching and non teaching faculty of CSE department and my dear friends who helped in all the ways while preparing the Report.

References

[1] https://archive.org/download/twitter_cikm_2010

[2] http://www.cloudera.com/content/www/enus/documentation/other/tutorial/CDH5/Hadoop-Tutorial/ht_example_4_sentiment_analysis.html

[3] https://github.com/omarshammas/sentiment_analysis

[4] https://github.com/timvandermeij/sentiment-analysis/tree/master/words

[5] https://github.com/madhusudancs/sentiment-analyzer/blob/master/analyzer/train.py

[6] http://academictorrents.com/details/d8b3a315172c8d804528762f37fa67db14577cdb

[7] https://archive.org/details/twitter_cikm_2010

[8] http://www.mrgeek.me/technology/datascience/data-mining-1-5-million-tweets-for-twitter-sentiment-analysis/

[9] http://www.slideshare.net/sumit786raj/sentiment-analysis-of-twitter-data?related=1

[10] http://www.slideshare.net/niteshsinghns/twitter-sentiment-analysis-project-report

[11] http://www.alex-hanna.com/tworkshops/lesson-6-basic-sentiment-analysis/

[12] http://alexdavies.net/twitter-sentiment-analysis/

[13] http://wwwnlp.stanford.edu/courses/cs224n/2009/fp/3.pdf