**Data Analysis On Twitter Data**

*Jayachitra T R -- 1MS12CS040*

*Praveen K -- 1MS12CS075*

*Manoj A -- 1MS12CS053*

*Praveen S P -- 1MS12CS076*

*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 hadoop which will process the huge amount of data on a hadoop cluster faster in real time.

The main purpose of this project is to collect large data from social networking site 'TWITTER' 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, Nature of tweets, analysis of data, hadoop, Naive Byes, Map Reduce

1. **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 eclipse, Hadoop and map reduce parallel programming concept. 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.

1. **Design**
2. ***Number of Modules***: ***4***
3. ***Modules Description***:

***1).Fetching and Extracting Data:***

The data set contains training set and test set. The training set contains 115,886 twitter users and 1,19,629 tweets from the users. The test set contains 5,136 twitter users and 1,22,891 tweets from the users. All the locations of users are uploaded from their

smart phones with the form of "UT: Latitude , Longitude".

***2). Data PreProcessing:***

Data set is pre-processed using eclipse IDE to eliminate stop word and stemming is done using porter’s algorithm.

***3). Classification:***

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

***4). Analysis:***

The Preprocessed 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 or neutral tweets and the same is represented graphically.

**C. Algorithm Design:**

1). Porter’s Algorithm (stemming) :

Step 1: Gets rid of plurals and -ed or -ing suffixes.

Step 2: Turns terminal y to i when there is another vowel in the stem.

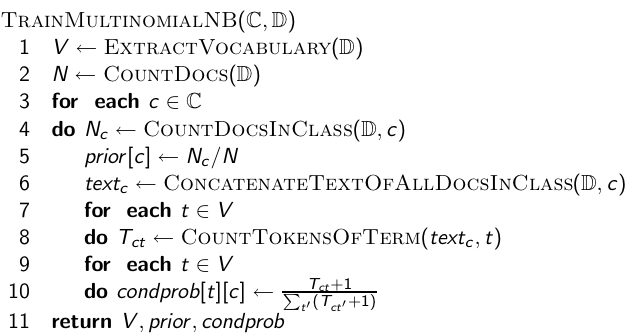
Step 3: Maps double suffixes to single ones: -ization, -ational, etc.

Step 4: Deals with suffixes, -full, -ness etc.

Step 5: Takes off -ant, -ence, etc.

Step 6: Removes a final –e.

**2). Naive bayes algorithm (classification)**



**Example:**

****

1. **Implementation**

The Project is implemented as follows:

* Twitter Data Set Identification and Extraction
* Pre-processing of data set
  + Removing unwanted attributes
  + Eliminating stop words
  + Stemming ( Porter’s algorithm)
* 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 hadoop for mapping it onto different machines.

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’, ‘happyyyyyy’ 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.

1. **Testing and Comparision**

In this project the environment in which this project has been built is python and hadoop using map reduce concept ,hence all the processing and testing that is happening will be done on the python IDLE interface and hadoop framework. 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.

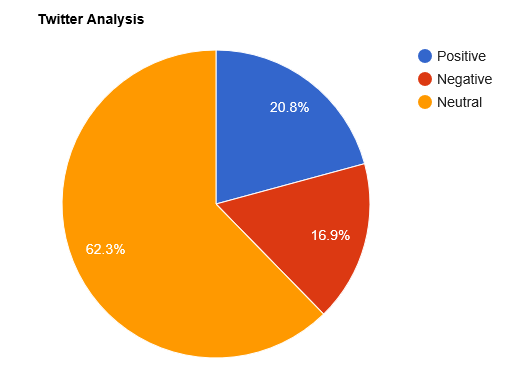
The project is implemented using hadoop, Map reduce parallel programming and python. Our project is unique compared to the earlier projects which are done on the same topic. Here we are not using any tools like NLTK or we are fetching the dataset directly from the tweeter. But we have our own dataset and we have installed hadoop and wrote our own map reduce code for classifying the datasets. Hence our project is different from all those earlier projects. Here we have made the comparision in order to show new things and features that we have done and implemented in our project.

1. **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 sample example shown in pie chart which is obtained by running the map reduce code in hadoop.

GRAPHICAL REPRESENTATION OF NATURE OF USER

TWEETS

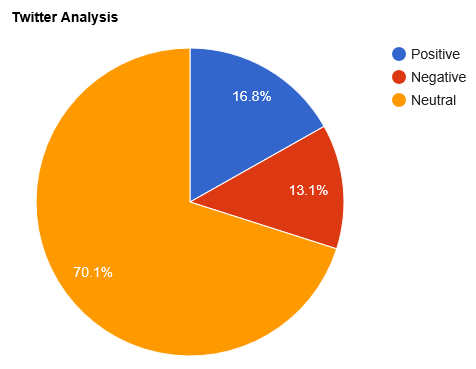


The sample count for the the given above chart is shown in the following table:

|  |  |
| --- | --- |
| Opinion | Count |
| Positive | 25540 |
| Negative | 20765 |
| Neutral | 76586 |

GRAPHICAL REPRESENTATION OF NATURE OF A SINGLE

USER TWEET



The sample count for the the given above chart is shown in the following table:

|  |  |
| --- | --- |
| Opinion | Count |
| Positive | 313 |
| Negative | 152 |
| Neutral | 555 |

1. **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 Hadoop ensures the distributed processing and it also lowers the access time. Hence overall the time efficiency increases owing to the above mentioned factors.

1. **Scope and Future work**

Twitter has lot of scope in the modern era. Twitter as a social media has many users and the number 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.

1. **Conclusion**

This project gave us hands on experience of handling and parallel processing of huge amount of data. Data collection process introduced us to java 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 parallel data processing tool: Hadoop. Apache Hadoop framework 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 installation and configuration of hadoop distributed file system but also map reduce programming model. At the end of analysis phase data visualization was performed with the help of Google Developer. 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 , our utmost Gratitude to our HOD, Dr. K G Srinivas who gave us a chance to complete this project. He has been our inspiration as we overcome all the obstacles in the completion of this project work. He had kind concern and consideration 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 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
14. http://www.bogotobogo.com/Hadoop/BigData\_hadoop\_Install\_on\_ubuntu\_single\_node\_cluster.php
15. <file:///D:/7th%20sem/PBL%20Lab/cheng_cikm10.pdf>