ISSN:0975-9646

Real Time Sentiment Analysis of Twitter Data Using Hadoop

Sunil B. Mane, Yashwant Sawant, Saif Kazi, Vaibhav Shinde

College of Engineering, Pune

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

Keywords—Sentiment analysis, Streaming API's, Opennlp, Wordnet

I. Introduction

Today ,the textual data on the internet is growing at a rapid pace. Different industries are trying to use this huge textual data for extracting the people's views towards their products. Social media is a vital source of information in this case. It is impossible to manually analyze the large amount of data. This is where the need of automatic categorization becomes apparent[(4)]. Subjective data is analyzed generally in this case. There are a large number of social media websites that enable users to contribute, modify and grade the content. Users have an opportunity to express their personal opinions about specific topics. The example of such websites include blogs, forums, product reviews sites, and social networks. In this case, twitter data is used. Sites like twitter contain prevalently short comments, like status messages on social networks like twitter or article reviews on Digg. Additionally many web sites allow rating the popularity of the messages which can be related to the opinion expressed by the author.

The focus of our project is to assign the polarity to each tweet I.e. whether the author express positive or negative opinion.

II. Related Work

Sentiment analysis is most popular tend in today's world.

Lot of work has been done in this sector. Following are some approaches which are most popular in today's world.

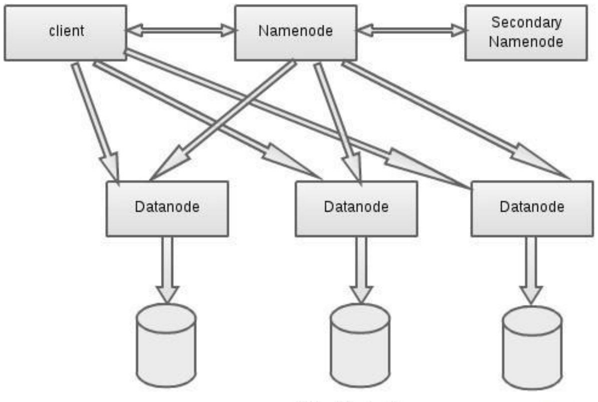
There has been a lot of research in the area of Sentiment analysis. Bo Pang and Lee were the pioneers in this field. Current works in this area includes using a mathematical approach which uses a formula for the sentiment value depending on the proximity of the words with adjectives like 'excellent', 'worse', 'bad' etc. Our project uses the Naïve-Bayes approach and an hadoop cluster for distributed processing of the textual data. Also the analysis native linguistics of a particular country along with English usage is also being worked upon.

i. Latent semantic analysis: This is natural language processing technique which works on relationship between document and words in it.

ii. Point-wise Mutual information: This is an actual mathematical term which is used to measure relationship between any two objects. This formula is used to determine the words association in a document with a pre-defined set of adjectives to determine the sentiment of the word .

HADOOP

The Hadoop platform was designed to solve problems which had lot of data for processing. It uses the divide and rule methodology for processing. It is used to handle large and complex unstructured data which doesn't fit into tables. Twitter data being relatively unstructured can be best stored using Hadoop. Hadoop also finds a lot of applications in the field of online retailing, search engines, finance domain for risk analysis etc.



Data Replications

HDFS

Hadoop Distributed File System (HDFS) is a distributed file system which runs on commodity machines. It is highly fault tolerant and is designed for low cost machines. HDFS has a high throughput access to application and is suitable for applications with large amount of data. HDFS has a **1** master server architecture which has a single namenode which regulates the filesystem access. Datanodes handle read and write requests from the filesystem's clients. They also perform block creation, deletion, and replication upon instruction from the Namenode. Replication of data in the filesystem adds to the data integrity and the robustness of the system.

Data replication is done for achieving fault tolerance. The large data cluster is stored as a sequence of blocks. Block size and the replication factor are configurable. Replication factor is set to 3 in our project which means 3 copies of the same data block will be maintained at a time in the cluster.

III. Our Approach

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.part of speech tagged using opennlp. 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 sentiment analysis are removed in this phase. Stop words are tagged as _DT in Opennlp. All the words having this tag are not considered.

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.

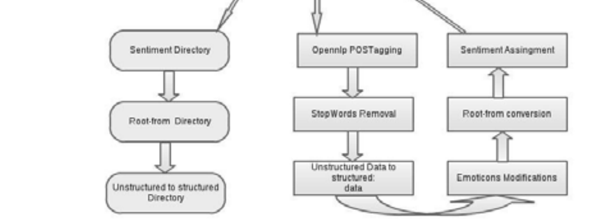
iii. Emoticons: These are most expressive method available for opinion. The emoticons symbolic representation is converted in to words at this stage i.e. □ to happy

A. *Real time data and features*

The real time that is necessary for this project is obtained from the streaming API's provided by twitter. For the development purpose twitter provides streaming API's which allows the developer an access to 1% of tweets tweeted at that time bases on the particular keyword. The object about which we want to perform sentiment analysis is submitted to the twitter API's which does further mining and provides the tweets related to only that object.

Twitter data is generally unstructured i.e use of abbreviations is very high. A tweet consists of maximum 140 characters. Also it allows the use of emoticons which are direct indicators of the author's view on the subject.

Tweet messages also consist of a timestamp and the user name. This timestamp is useful for guessing the future trend application of our project. User location if available can also help to gauge the trends in different geographical regions.



B. *Part of Speech*

The files which contained the obtained tweets are then

C. *Root form*

The given words in tweet are converted to their root form to avoid the unwanted extra storage of the derived word's sentiment. The root form dictionary is used to do that which is made local as it is heavily used is program. This lowers the access time and increases the overall efficiency of the system.

D. *Sentiment Directory*

The sentiment Directory is created using standard data from sentimwordnet and using all possible usage of a particular word i.e. "good" can be used in many different ways each way having its own sentiment value each time it is used. So overall sentiment of good is obtained from all its usage and stored in a directory which should be again local to the program (i.e. in primary memory) so that time should not be wasted in searching word in the secondary memory storage.

E. Map-reduce Algorithm

The faster real time processing can be obtained by using cluster architecture set up by hadoop. The program contains chained map-reduce structure which used to process ever tweet and assign the sentiment to each remaining words of tweet and then summing it up to decide final sentiment. Here special care should be taken for the phrasal sentences where sentiment of phrase matters rather than sentiment of each word. It can be done by dynamic directory of phrases and their sentiment values can be obtained from standard algorithm PMI-IR **2**

IV. Accuracy

The overall accuracy of project is determined by time required to access from various modules i.e. accessing from opennlp, wordnet and sentiwordnet.

As all components are in series i.e. used one after the overall, theoretically the overall accuracy of the program is the product of accuracy of all its modules .We tested our implementation on the standard dataset provided at following http://www.cs.tau.ac.il/~kfir-bar/mlproject/twitter.data

The accuracy of our project after running on this data set is as following

Sentiment	Count	Correct	%	Tolerance
Positive	729	542	74.34	-0.01
Negative	665	458	68.87	+0.09
Neutral	72	53	73.61	+- 0.005

So, overall accuracy is the mean i.e 72.27.The overall accuracy in case of the project at a Columbia university was mentioned for the Naïve-Bayes classifier used is[3] cs.columbia.edu/~julia/papersAgarwaletal11.pdf

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

VI. Future Scope

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.

VII. Features.

The previous projects in the field of sentiment analysis have used a scaling system for rating the statements like Very Positive(VP), Positive(P), Neutral(N), Negative(Neg) and Very Negative (VNeg). They achieved this by using the mutual interaction between the words in the statement with words like 'excellent', 'worse' and other extreme adjectives. Our project uses a numbering approach which can be assigned a suitable range for the different sentiments. Emoticons, 'Root form' are also included in the sentiment evaluation. Emoticons usage is a very good indicator for determining the exact emotion of the author towards the subject.

VIII. Conclusion

Sentiment analysis is a very wide branch for research. We have covered some of the important aspects. We plan ahead to improve our algorithm used for determining the sentiment value. Also the project as of now can also be expanded to other social media platform usages like movie reviews(IMDB reviews), personal blogs. The accuracy achieved is also mentioned below.(6)

Emoticons and the use of hashtags for the sentiment evaluation is a very important inference related to sentiment analysis of social media data. Our project uses emoticons but the use of hashtags to determine the context of the tweet is not done. Hence with the current limitations the accuracy is found to be 72.27 %.

References

1. Apporv Agarwal, Jasneet Singh Sabarwal, "End to End Sentiment Analysis of Twitter Data"
2. Theresa Wilson, Joanna Moore, Efthymios Kouloumpis,"Twitter Sentiment Analysis – The Good, the Bad and the OMG"
3. Apoorv Agarwal, Owen Rambow, Rebecca Passonneau, "Sentiment Analysis of Twitter Data"
4. Bo Pang, Lillian Lee, "Opinion Mining and Sentiment Analysis"
5. Andrea Esuli, Fabrizio Sebastiani, "SENTIWORDNET: A Publicly Available Lexical Resource for Opinion Mining"
6. Bing Liu, Minquing hu, "Mining and summarizing Customer Reviews"
7. Nitin Jindal, Bing Liu, "Opinion and Spam analysis"