Advances in Computer Science and Information Technology (ACSIT)
Print ISSN: 2393-9907; Online ISSN: 2393-9915; Volume 2, Number 1; January-March, 2015 pp. 27-29

http://www.krishisanskriti.org/acsit.html

Sentimental Analysis Using Social Media and Big data

Arpita Gupta[1] and Anand Singh Rajawat[2]

[1]PG Scholar, Department of CSE, SVITS, Indore, India
[2]Department of CSE, SVITS, Indore, India
E-mail: [1]arpitagupta0505@gmail, [2]comanandsrajawat@gmail. com

Abstract—*Social media is bestest way of communication. Millions of people communicate through this everyday. There are two major problems encounter when we talk about processing of the data related to social media. One is ambiguity of data and the other is data is completely unstructured. To overcome this major area of problems in social media here we can use some techniques of Big data in which analysis and collection of data plays major role or key role. It allows the data collection and analysis from a big data without hindrance, obstruction and time delay. The focus of our project is to analysis the unstructured data and overcomes the problem of ambiguity using hadoop and this will increase performance and security will also be improved.*

Keywords:-*Sentiment analysis, Text mining, Machine learning, Big data, Wordnet.*

1. INTRODUCTION

Big Data is trending research area in computer Science and sentiment analysis is one of the most important part of this research area. Big data is considered as very large amount of data which can be found easily on web, Social media, remote sensing data and medical records etc. in form of structured, semi-structured or unstructured data and we can use these data for sentiment analysis.

Sentimental Analysis is all about to get the real voice of people towards specific product, services, organization, movies, news, events, issues and their attributes[1]. Sentiment Analysis includes branches of computer science like Natural Language Processing, Machine Learning, Text Mining and Information Theory and Coding. By using approaches, methods, techniques and models of defined branches, we can categorized our data which is unstructured data may be in form of news articles, blogs, tweets, movie reviews, product reviews etc. into positive, negative or neutral sentiment according to the sentiment is expressed in them.

Sentiment analysis is done on three levels [1]

- Document Level
- Sentence Level
- Entity or Aspect Level.

Document Level Sentiment analysis is performed for the whole document and then decide whether the document express positive or negative sentiment. [1]

Entity or Aspect Level sentiment analysis performs finer-grained analysis. The goal of entity or aspect level sentiment analysis is to find sentiment on entities and/or aspect of those entities.

Sentence level sentiment analysis is related to find sentiment form sentences whether each sentence expressed a positive, negative or neutral sentiment. Sentence level sentiment analysis is closely related to subjectivity classification. Many of the statements about entities are factual in nature and yet they still carry sentiment. Current sentiment analysis approaches express the sentiment of subjective statements and neglect such objective statements that carry sentiment [1]. For Example, “I bought a Motorola phone two weeks ago. Everything was good initially. The voice was clear and the battery life was long, although it is a bit bulky. Then, it stopped working yesterday. [1]” The first sentence expresses no opinion as it simply states a fact. All other sentences express either explicit or implicit sentiments. The last sentence “Then, it stopped working yesterday” is objective sentences but current techniques cannot express sentiment for the above specified sentence even though it carry negative sentiment or undesirable sentiment. Context-aware sentiment analysis tackles the problem of ambiguity by attempting to determine the superordinate concept of the sentiment term in a given context. Straightforward for humans with ample domain experience, this can be a difficult task for automatedsystems. [2]. The focus of our project is to analysis the unstructured data and overcomes the problem of ambiguity.

2. LITERATURE SURVEY

Research is carried out in two basic ways: qualitative and quantitative. In a qualitative approach, the researcher makes knowledge claims based primarily on a constructivist perspectives (i. e. the multiple meanings of individual

experiences, meanings socially and historically constructed, with an intent of developing a theory or pattern) or advocacy/participatory perspectives (i. e. political, issue-oriented, collaborative or change oriented) or both [3]. Qualitative research involves finding out what people think, and how they feel - or at any rate, what they say they think and how they say they feel. This kind of information is subjective. It involves feelings and impressions, rather than numbers. On the other hand, quantitative research focuses on measuring an objective fact. Key to conducting quantitative research is definition of variables of interest and to a large extent a sense of detachment in the data collection by the researcher. Quantitative research analyses data using statistics and relies on large samples to make generalized statements.

A new trend has emerged in research today - the mixed method research design or plural research designs. Plural research design combines both qualitative and quantitative research methods in market studies and is becoming quite a fashion in social science research. Triangulation also can be impractical to some research situations given the high research cost of multiple data collection and the time delays in data collection and data analysis[4]. Sentiment analysis provides a faster, simpler and less expensive alternative to traditional qualitative market research techniques like observations, interviews and even ethnography as well as provides information in real time

3. 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[5], support vector machine[6] , maximum entropy and an hadoop cluster for distributed processing of the textual data. Also the analysis native linguistics of a particular country along with Englishusage is also being worked upon.

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.

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 name node which regulates the file system access. Data nodes handle read and write requests from the file system's clients. They also perform block creation, deletion, and replication upon instruction from the Name node. Replication of data in the file system adds to the data integrity and the robustness of the system.

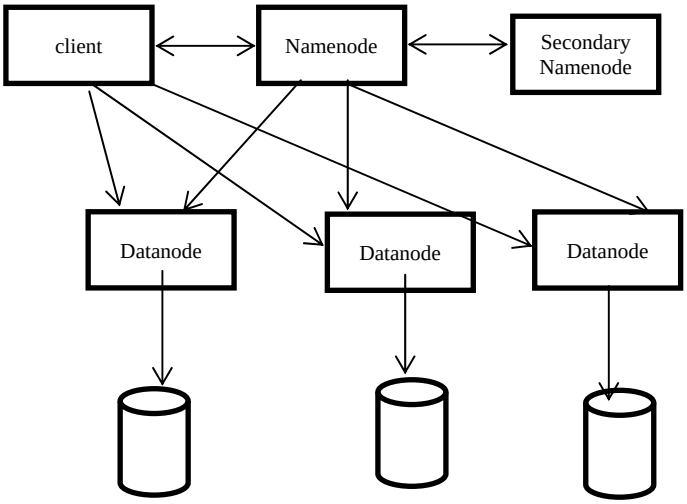

Fig. 1: Data Replication

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 time in the cluster.

4. 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 and facebook 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.
iv. To overcome the problem of ambiguity.

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 or facebook. 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 sentiment wordnet 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 secondarymemory 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 **.**

5. 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, Posts and commentscontaining 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.

6. 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. [7]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 or post is not done. Hence with the current limitations the accuracy is found to be 74%

REFERENCES

[1] Bing Liu, Sentiment Analysis and Opinion Mining, Morgan and Claypool Publishers, May 2012. p. 18-19, 27-28, 44-45, 47, 90-101.

[2] http://systems-sciences. uni-graz. at/etextbook/bigdata/sentiment_analysis. html.

[3] Creswell, John. (2007). Qualitative Inquiry and Research Design. Choosing Among FivemApproaches. 2nded. Sage Publications Inc: California.

[4] Kelle, Udo. (2006). "Combining Qualitative and Quantitative Methods in Research Practice: Purposes and Advantages. " Qualitative Research in Psychology, 3 (4): 293 311.

[5] ApoorvAgarwal, Owen Rambow, Rebecca Passonneau, "Sentiment Analysis of Twitter Data"

[6] Vapnik, Vladimir, N. (1995). *The Nature of Statistical Learning Theory*, Springer-Verlag: New York.

[7] Bing Liu, Minquinghu, "Mining and summarizing Customer Reviews"