Real Time Twitter Sentiment Analysis

Shivam Singh[1], Sonal Agarwal[2], Sakshi Agarwal[3]
[1,2,3]Galgotias College of Engg. & Tech. Greater Noida
[1]shivam.cs@hotmail.com, [2]ssquealer.agarwal@gmail.com, [3]sakshiagarwal95cse@gmail.com

***Abstract*—Social media is the ultimate equalizer, gives a voice & platform to anyone willing to engage, lot of people use social media to share mundane things or for self- glorification and hence expand universe. In face of this overwhelm of data, even the smallest of small fries to compete with the big guns and harness this unwidely data deluge to work for us, However with so much social media available on web, sentimental analysis is now considered as a big data task. If big data is water pouring out of your faucet, then social media is a reservoir that streams come from, therefore unending influx of content from social media is indeed what has allowed data analytics of past to balloon into "big data", which is to say : learn to understand and make productive use of views, likes, shares, follows, retweets, comments and downloads**

***Keywords*—Sentiment, hadoop, twitter, cloudera, MapReduce**

1. INTRODUCTION

Twitter is getting the lion's share of all the social media and microblogging sites from the press and blogosphere and hence is responsible for microblogging (popular form of communication on web). Twitter posts are popularly known as tweets and have limited length of 140 characters where most of users express their commotion, despondency and hence all the sensation on various issues in form of tweets, which convey a great deal about their sentiments at macroscopic level.

In present scenario sentiment analysis technique is one of the ways to predict opinion of a community ,which in collection over a given period reveal transformation in public mood at large, where empirical analysis is carried on basis of blogs, reviews , tweets and other posts on microblogging and social networking sites.

Twitter sentimental analysis is major breadth of topics that are covered nowadays, people talk about anything and everything. It gets really important to shift though chaff and narrow down to relevant keywords which is done by normalizing and validating against predefined parameters. Sliced and diced data is classified into positive, negative and neutral keywords.

2. RELATED WORK

The former advent in epoch of opinion mining includes Turney and Pang who applied different methods for detecting polarity of product reviews and movies respectively. Alternatively, task is commonly defined as classifying a given text into one of the the two classes : objective and subjective, which is far more formidable than polarity classification. However subjectivity of words and phrases may depend on their content and objective document may contain subjective sentence (Eg: This man is not a coward). However as said by Pang objective sentence from a document before classifying its polarity helped improve performance.

3. PROPOSED WORK

The focus of this research is to devise an a method which can done sentiment analysis on twitter data very fast and accuracy, though higher accuracy can be achieved by machine learning by training the system for all the words and their sentiment score value, but this is not preferred because it will decrease the speed of sentiment analysis, so a better alternative (i.e. Hadoop with natural language processing) is chosen.

Majorly sentiment analysis of twitter data can be broadly divided into four tasks that are –

1. Ingestion of tweets into HDFS
2. Post Processing
3. Query Processing using HIVE
4. Use the data for decision making.

3.1 Ingestion of tweets-

Tweets are ingested from twitter streaming using twitter 4j API. This API contains all the essential java files that are used to copy the source from twitter source into a user machine. Apache Flume provides a mechanism for real time ingestion of tweets. A .conf file is created which facilitates the transfer of tweets from twitter source into HDFS Sink using an appropriate streaming channel. This "conf" file contains the user keys and the certain keywords on the basis of which we want the tweets.

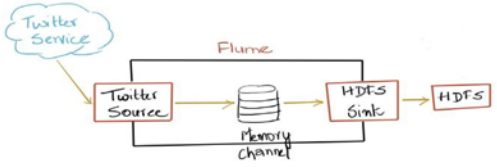


Fig.1. Tweets Ingestion Process

3.2 *Post Processing*

Most of the tweets (~70%) requires post processing to increase the accuracy of sentiment analysis, the eight step post processing is the unique feature of our research because we are filtering more and more irrelevant data to achieve only meaningful information. This eight step filtering is done in a sequential order,which is as follows :-

Removal of links, special symbols like @
Removal of stop words like a,an,the.
Construction of n-grams: Attaching the not keyword in word preceding it.
Translation of acronyms: Replacing lol with lough out loud. There are 84 such acronyms.
Spelling correction.
Booster Words- Suppose "very" is used two times, then double the sentiment score.
Emoticon List- Replace □ with positivesentiment.
Repeated punctuation- keywords likegood!!! Increases the strength.

3.3 *Query processing using HIVE*

Once the tweets are ingested into HDFS, we need to convert twitter data stored in HDFS from ".JSON" to a readable form so use the command:-

ADD jar /home/cloudera/hive-serdes-1.0-SNAPSHOT.jar

TABLE I. SAMPLE DICTIONARY

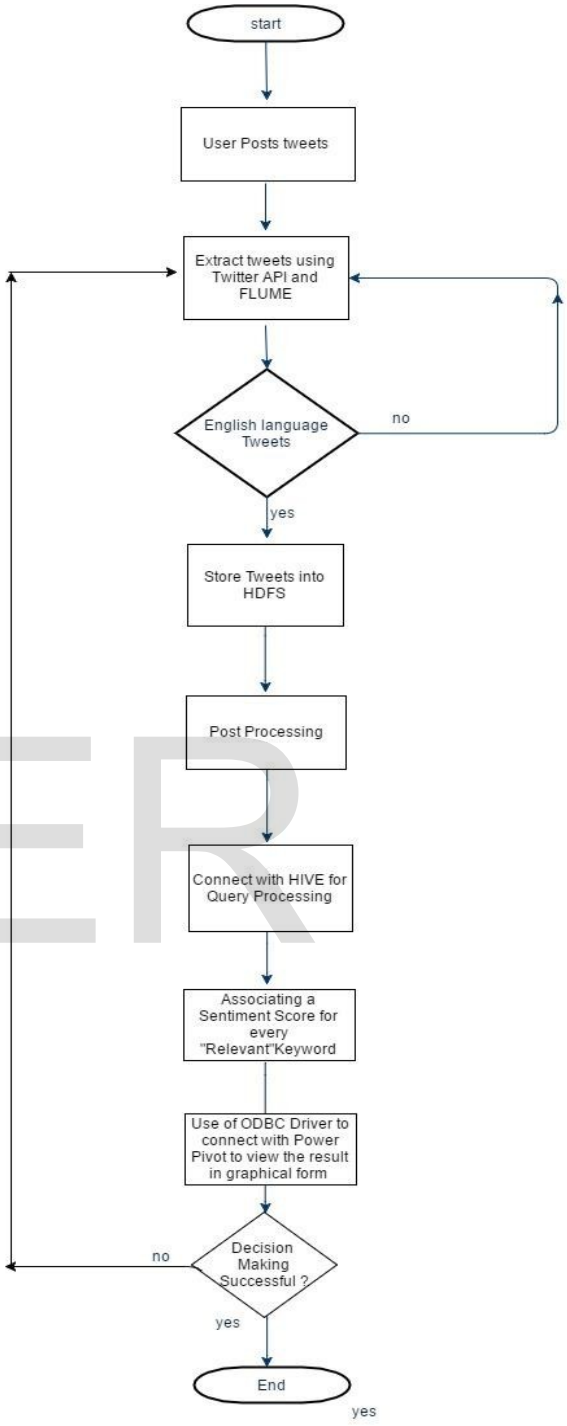
Strength	Word	Polarity
weaksubj	Terminate	negative
weaksubj	Terminated	negative
weaksubj	termination	negative
strongsubj	needed	Blind negation

1. After this create a new table having enteries as much as required for analyzing relevant information, most of data is not relevant to us such as followers, follow count; so we need only relevant data, therefore create table only with few attributes.

2. Split each tweet into words using the split() UDF available in Hive. To split the text as words use the split() function, an array of values will be returned.

3.4 *Use the data for decision making:-*

Fig.2. Overall Sentiment Analysis Process



3. Split each word inside the array as a new row. A UDTF (User Defined Table Generating Function) will be needed for this. We have built-in UDTF called 'explode' which will extract each element from an array and create a new row for each element.

Connect the HIVE to Micrososft Power Pivot view in

Excel using the ODBC driver to get the processed data in the form of graphs, charts and geographical location based data because the culture and diversity of a location matters very much. Following is an example of data representation

Fig.3. Geolocation based sentiments

4. CONCLUSION

As we know that the twitter post are very important source of opinion on different issues and topics, It can give a keen insight about a topic and can be a good source of analysis. These Analysis can facilitate the process of decision making in various areas such as health care analysis, market anaylsis, weather forecasting, advertising analysis, fraud detection, traffic flow optimization etc.

5. FUTURE WORK

Analysing Sentiments from Sarcastic tweets.

Sarcasm is the use of words that mean the opposite of what the speaker wants to say with the "hidden" or rather apparent intention of insulting someone, showing irritation or to be funny. Recognition of sarcasm can ease many Sentiment Analysis NLP applications, such as review summarization, dialogue systems, review ranking systems, etc.

This work was implemented on a single data node machine and can also be implemented on a machine with multiple data node and name nodes which is expected to increase the speed of processing.

6. ACKNOWLEDGMENT

This research would not have been completed without the guidance of Mr. R.P. Singh (Project Guide), Mr. Manish Kumar Sharma (Project Co-ordinator), Ms. Swati Sethi (Webtek Labs) and the HOD CSE Dr. Bhawna Mallick. I would like to thank all the professors who guide us in this project.

7. REFERENCES

[1] Matthew Koehler, Spencer Greenhalgh, Andrea Zellner, Michigan State University, United States , "Potential Applications of Sentiment Analysis in Educational Research and Practice ", 2015 in Las Vegas, NV, US Publisher: Association for the Advancement of Computing in Education (AACE).

[2] Sunil B. Mane , Y.t Sawant, S. Kazi, V. Shinde , "Real Time Twitter Data Sentiment Analysis ", International Journal of Computer Science and Information Technologies, Vol. 5 (3) , 2014, 3098 – 3100 , ISSN:0975-9646.

[3] M. Wiegand, Alexandra Balahur, B. Roth, D. Klakow, A. Montoyo. 2010. A survey on the role of negation in Sentiment Analysis. Proceedings of the workshop on negation speculation in natural language processing 60–68, Association for Computational Linguistics.

[4] Chihil Hung and Hao-kai Lin, "Using Objective Word in SentiWordNet to Improve Word-of-Mouth Sentiment Classification", IEEE Computer Society, P.48- 53, March-April 2013.

[5] Building Machine Learning Algorithms on Hadoop for Bigdata Asha T, Shravanthi U.M, Nagashree N, Monika M International Journal of Engineering and Technology Volume 3 No. 2, February, 2013.

[6] Jintao Mao and Jian Zhu, "Sentiment Classification based on Random Process", IEEE Computer Society, International Conference on Computer Science and Electronics Engineering, p.473-476, 2012.

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

[8] M. Taboada, J. Brooke, M. Tofiloski, K. Voll and M. Stede. 2011. Lexicon based methods for Sentiment Analysis. Computational linguistics, volume 37, number2, 267–307, MIT Press.

[9] Cui, M. Zhang, Y. Liu, S. Ma, Emotion Tokens: Bridging the interval among Multilingual Twitter Sentiment Analysis, Springer-Verlag, Berlin, Heidelberg, 2011, pp. 238–249.

[10] https://blog.cloudera.com/blog/2012/11/analyzing-twitter-data-with- hadoop-part-3-queryingsemi-structured-data-with-hiv