Extraction and Analysis of Twitter Data using Hadoop Framework

Simranjeet Singh Randhawa[1]
Computer Science & Engineering[1], Maharaja Agrasen Institute Of Technology, Delhi, India[1]
Email:ssrbazpur@gmail.com[1]

Abstract- Analysis of data helps us to discover an unwanted pattern in the data. It also helps in the sentimental analysis. The extracted data should be mined efficiently in order discover unwanted patterns. Using Twitter API and big data technology, the Twitter data is extracted and analyzed in order to discover positive, negative or neutral sentiments. The analysis of data also helps us discover various pattern in data. Sentimental Analysis is a process of analyzing data based on the person's feelings, reviews, and thoughts. It helps to determine the sentiments of a person. Using Twitter API and big data technology, sentiments are computed and the result is depicted graphically.

Index Terms-Sentimental Analysis, Flume, and Hive, HDFS.

1. INTRODUCTION

The data extracted from twitter is analysed and mined to obtain some valuable results. . The sentimental analysis is also performed on the extracted data.The extracted data is in unstructured form,it is first converted to a structured form and then analysed to obtain various results. Sentimental analysis helps to evaluate the sentiments of a person. Sentiment analysis aims to determine the attitude of a speaker, writer, or other subject with respect to some topic or the overall contextual polarity or emotional reaction to a document, interaction, or event. Twitter contains an enormous amount of data and the data if managed effectively can get us some effective results. The main aim is to deal with Twitter data by extracting it using flume and then using sentimental analysis technique on the extracted twitter data in order to get positive, negative or neutral sentiments. The data can also be mined to get number of tweets, retweets, number of followers.

2. PROPOSED SYSTEM

Using Twitter API twitter data is extracted. With the help of consumer key, consumer secret key, access token and access token secret connection to Twitter is made and data is retrieved and stored on HDFS. The extracted data, when placed in HDFS, can be analyzed to retrieve various meaningful results like it can help in sentimental analysis, calculation of a number of tweets, retweets, etc. To get twitter tweets in HDFS we'd like associate agent therefore we will use FLUME. Flume can load twitter tweets into your HDFS. And to fireplace any question to city tweets we'll load these tweets from HDFS to PIG or HIVE. For sentiment analysis, a dictionary is formed and used that contains a worth for every word. And a specific value is assigned to every word of tweet, and finally, the full values of all the tweets are calculated. These values are negative, positive or neutral.

The proposed system is depicted in the figure:

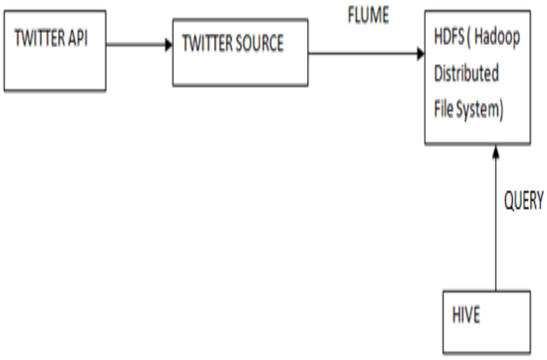

Figure 1 Proposed System flow

3. USE OF APACHE FLUME

Apache flume is responsible to deliver the data from sources to the sink with the help of a channel. Sources are broken down into set of events and the events are sent through the channel to the sink. The sink is responsible to write the event to a predefined location.[1] The Apache flume plays a vital role in Twitter's data extraction. The twitter data once extracted using flume is directed to the HDFS where the actual execution of queries takes place. Adopting the use of flume it is relatively easy to gather the data at our predefined location. Hence it becomes a vital part of the project. We will use the custom source to filter the tweets on a set of search keywords to help determine relevant tweets, instead of a pure sample of the whole Twitter firehose.

In Flume, every individual piece of data (tweets, in our case) is termed associate event; sources produce events and send the events through a channel, that connects the source to the sink. The sink then writes the events out to a predefined location. Flume supports some commonplace data sources, such as syslog or netcat. For this use case, we'll get to design a custom source that accesses the Twitter Streaming API, and sends the tweets through a channel to a sink that writes to HDFS files. in addition, we will use the custom source to filter the tweets on a set of search keywords to help determine relevant tweets, instead of a pure sample of the whole Twitter firehose.[2].

4. MANAGING EXTRACTED DATA (PARTITION MANAGEMENT)

Once we have the Twitter information loaded into HDFS, we are able to stage it for querying by making an external table in Hive.We make use of external table because by making use of external table we can perform queries on the table without moving the data. We make use of partition technique and apply the partitioning technique on the table in order to prune the files that we tend to scan. This technique helps us manage large data sets efficiently.A table subdivided into partitions permits us to prune the files that we tend to scan once querying, which ends up in higher performance when managing giant information sets.[3]

□

5. CONVERTING DATA TO OUR PRESCRIBED FORMAT

The extracted twitter data is needed to be converted into a formatted structured so that hive can be used read the data effectively and help in analyzing the pattern.We make use of SerDe(Serializer and Deserializer)[4] in order to read the data in JSON. By processing the data we then make use of user-defined functions for performing the sentiment analysis. And acquire the results where a new table is created by partition concept such that all the comments that are having positive will go into the positive partition and all the comments that are having moderate will go into moderate partition and finally all the comments that are having negative will go into the negative partition.

6. GRAPHICAL RESULTS OBTAINED

The bar graph below depicts positive, negative and neutral sentiments. The data is collected from different sources. But these graph gives us the idea of what all information we can retrieve from twitter easily. Hence these graphical results are very beneficial to get us an idea about any page on Twitter.

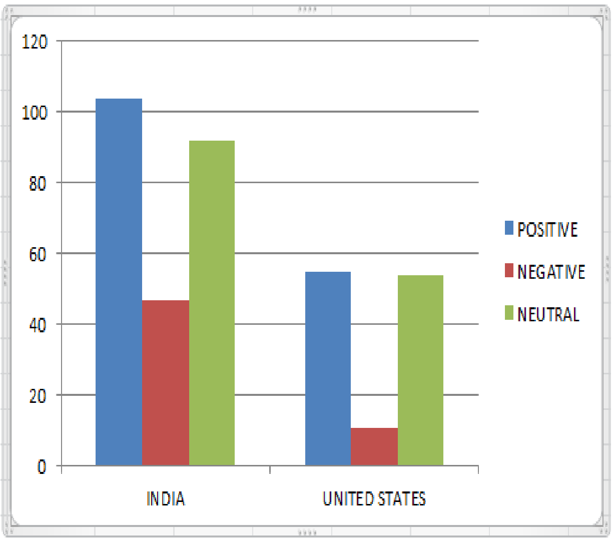

Chart 1(Positive,Negative,Neutral Count)

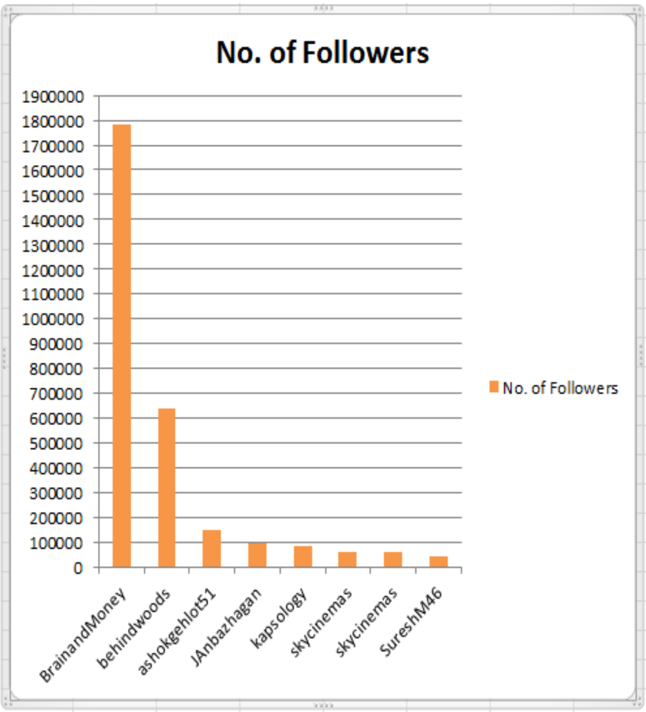

Chart 2: Number of Followers

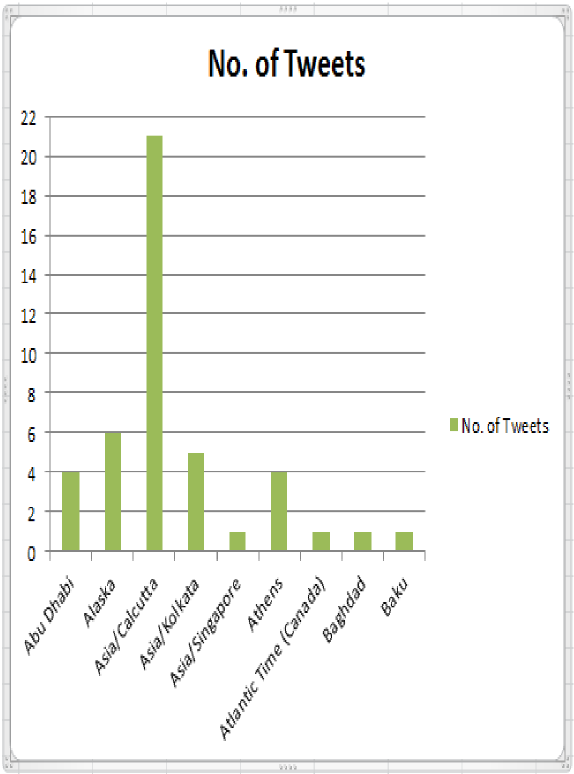

Chart 3: Number of tweets

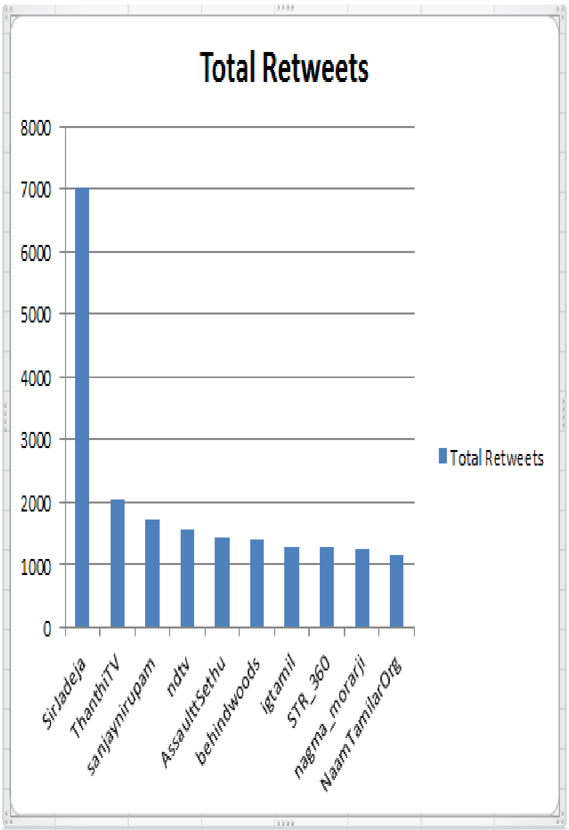

Chart 4: Number of Retweets

7. CONCLUSIONS

The data extracted from twitter is analysed such that important results such as positive, negative, neutral count is calculated. Moreover extracted data can be used to calculate the total tweets or retweets. Such analyses help to rate a particular page. This analysis also helps in calculating the number of bad tweets on a page. Using this strategy anyone can know the whether there are more positive, negative or neutral tweets on a page. This research is really helpful in analyzing data of trending topics on Twitter. It can help us to find some valuable asset at the page.

REFERENCES

[1] Maletic, J. I., Collard, M. L., and Marcus, A., "Source Code Files as Structured Documents", in Proceedings 10th IEEE, International Workshop on Program Comprehension (IWPC'02), Paris, France, June 27-29 2002, pp. 289-292.

[2] Judith Sherin Tilsha S, Shobha M S, “A Survey on Twitter Data Analysis Techniques to Extract Public Opinion”, International Journal of Advanced Research in Computer Science and Software Engineering, Volume 5, Issue 11, November 2015, pp 536-540.

[3] Marcus, A. and Maletic, J. I., "Recovering Documentation-to-Source-Code Traceability Links using Latent Semantic Indexing", in Proceedings 25th IEEE/ACM International Conference on Software Engineering (ICSE'03), Portland, OR, May 3-10 2003, pp. 125-137.

[4] Salton, G., and Automatic Text Processing: The Transformation, Analysis and Retrieval of Information by Computer,Addison-Wesley, 1989.