Survey: Twitter data Analysis using Opinion Mining

Adarsh M J
Research Scholar
Dept. of Computer science and Engineering
Adichunchanagiri Institute of Technology

Pushpa Ravikumar, PhD
Professor & Head
Dept. of Computer science and Engineering
Adichunchanagiri Institute of Technology

ABSTRACT
Social Networking is the medium widely used for expressing emotions and opinions in public life through smart phones and other mediums on the Internet. Amongst the popular portals is the Twitter. Twitter has been the point of attraction to several people in research in important areas like prediction of democratic electoral events, consumer brands, movie collections at box office, stock market, celebrities etc. Opinion mining also called as sentiment analysis offers a fast and broader way of monitoring the public sentiments. In this paper, a study on various perspectives and approaches of Twitter data analysis being carried out in recent years using opinion mining is made by considering the words, retweets, hashtags and emoticons.

Keywords
Twitter, Opinion Mining, sentiment Analysis

1. INTRODUCTION
The term Big Data is globally used for collection of Datasets that are huge and complex, these huge Datasets makes it difficult to process using traditional data processing techniques. The challenges include analysis, pattern recognition, visualization etc. The challenges related to Big Data provide a chance to understand the data patterns and helps in prediction of events and results [1].The analysis of Big data is being carried out in many streams like text processing, Network simulation and predicting user behavior study etc. The users are the central focus with the advent of web 2.0 for any organization. Big Data analysis of user data comes very handy in predicting the correct strategies for success of any product. The study of the user data in social networks is one of the current trends of the times. The Social networkers or the authors of the messages publish their emotions and opinions on variety of topics and discuss several current issues. The Micro blogging platforms are helping this cause with restriction-less message format and also with ease of accessibility. The huge amount of messages on social networks makes it very attractive medium for data analysis [1].

A very popular social networking Micro Blogging platform is Twitter. It was launched in 2006[8]. On Twitter, any user can publish a short message referred to as a tweet with a maximum length of 140 characters, which is visible on the public display. The public timeline conveying the tweets of all the users worldwide is an extensive real-time information stream of more than one million messages per hour [1]. The tweets hence can be used to explore the social media data and find texts related to one another.Opinion Mining or Sentiment Analysis corresponds to the determination of the sentiment that the writer wanted to transmit in his/her message. The Sentiment normally represents the text polarity i.e. whether the message has a positive, negative or neutral sentiment. Hence for corporate companies, brands, celebrities and others, the sentiment acts as a measure to observe the public and market opinions. We will be looking at the various approaches involved in this process of Opinion mining.

2. LITERATURE SURVEY
Ana C.E.S Lima and Leandro N de [2] proposed three approaches for the automatic classification of sentiments, an emotion based approach, and a word based approach and a hybrid approach. In the emotion based approach they used sentiment incorporated in the emotions as criteria to automatically classify the messages. The criteria to select a tweet are the presence of at least one Emoticon. The sentiment is inferred based on the Emoticon. The word based approach uses words that express sentiment as criteria. In tweets, the presence of words such as good, bad, excellent etc will express sentiment and hence can be inferred. In the hybrid approach a combination of Emoticons and words were used to infer the sentiment. They used Naïve-bayes Algorithm for classifying tweets and concluded that the combined i.e. the hybrid approach yields better results, also, they suggested to add a label “neutral” in future classification.

Dimitrios kotsakos, panos sakkos, Ionnis katakis, Dimitrios [3] Highlighted on tagging the tweets. They did the hashtag analysis in which a hash (#) symbol used to indicate a special meaning of a word and tag content in social networks like twitter. Users used hashtags for search, annotations or viral conversations often called Memes. They revealed interesting characteristics of some expected hashtags and some not expected hashtags. They also suggested to further investigate features that characterize the behavior of popular topics and to create taxonomies of hashtags that facilitate recommendation or searches.

Mahanaz Roshanaei and Shivakant Mishra [4] emphasize on effect of mood and emotions on a person’s behavior. They classified users in to positive, negative and neutral users based on followers and followees. Negative users are not interested in sharing their negativity in social media. The positive users are more likely to make friendships with negative users, also, the negative users retweet more than the positive users. They use Twitter as a tool for social awareness and also to gain emotional support. Retweeting positive tweets makes the negative tweeters feel positive. Both positive and negative users avoid interacting with each other.

Malhar Anjaria, Ram Mohana Reddy Guddeti [1] used Sentiment Analysis on Twitter data and came out with some conclusions. They introduced a novel approach of exploiting the user influence factor to predict the outcome of an election result. They referred US Presidential

elections 2012 and Karnataka assembly elections 2013 and concluded that the social network based behavioral analysis parameters can increase the prediction accuracy along with the sentimental analysis. The presence of all the entities say educational background etc in unbiased and equal manner is necessary to provide accurate results. They obtained a reasonable accuracy of around 88% in case of US Presidential elections 2012 and Karnataka assembly elections 2013.Tuan-anh hoang, William w Cohen, Ee-Peng lim, Daovy Pierce, david r Redlawsk [5] examined the effects of sentiment and political affiliation on retweetability of political tweets in twitter. They performed analysis on a large dataset of tweets collected from political oriented users in US during a long politically active period and confirmed that the sentiment and political affiliation have effects on retweetability of political tweets. The effects are different with different types of users who retweet the tweets.

John p Dickerson, vadim Kagan, V S Subrahmanian [6] mentioned a SentiBot framework which addresses the classification of users as human and Bots. SentiBot relied on four classes of features say tweet syntax, tweet semantics, user behavior and network centric user properties. They concluded that the Bots flip-flop much less frequently than humans in terms of sentiment. The positive sentiment expressed by humans is always stronger than the Bots. Humans disagree more with the general sentiment of applications than the Bots.

3. METHODOLOGY

Almost 70-80% of the data available for analysis is unstructured and about 20-30% of the data is structured. The RDBMS will help to store and process structured data whereas Hadoop deals with processing both types of data [9].Fig 1 shows the general methodology used in the Twitter data analysis. The Twitter data is available for public access through streaming API. If a user has a Twitter account, then he/she can create an App in Twitter to store the related tweets. Components like Flume for Hadoop eco system can be used for data streaming. The data obtained from the Twitter can be stored in HDFS. The database has all the tweets streamed from twitter and it is necessary to filter out the tweet which needs to be analyzed.

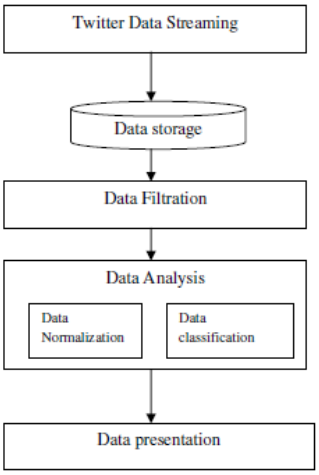

Fig 1: General Methodology of Twitter data Analysis

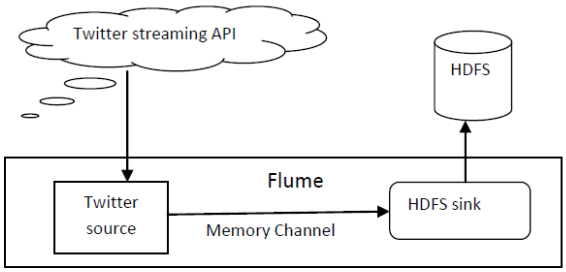

Fig 2: Streaming of Twitter data using flume

The Twitter streaming API outputs tweets in the JSON format which usually is complex. In the Hadoop eco system the Hive Project provides a query interface which can be used to retrieve the query data from the HDFS. Apache Flume is a data ingestion system that is configured by defining the end points in a data flow called sources and sinks. In flume (fig 2) each piece of data (tweets) is called an event and the events are sent through a channel, which connects the source to the sink. The sink then writes the event to some predefined location. Once the data is loaded to the HDFS, Hive can be used for querying, since Twitter data is in JSON format, one can use Hive SerDe interface. SerDe stands for Serializer and Deserializer, which are interfaces that tell Hive how it should translate the data into something that Hive can process [10]. Filtration can be done based on the keyword, Emoticons etc. The filtered data is subjected to analysis, For Example: If one is using HDFS, during analysis, the unstructured data is converted to structured data in columnar pattern. The tweets are split into words and are grouped based on the sentiments.

Structured tweets are generally in sentence format, with URL's specified for images at blog articles. Data in usage format can be obtained by removing the stop words that contain general terms like 'a', etc. The tweets can be categorized based on username, URL's, repeated tweets etc.There are several classifiers available which can be used for extraction and classification like Naïve-Bayes, maximum Entropy, Support vector machines(SVM) and Artificial Neural Networks(ANN).Data presentation gives the analytical results in a form easily understandable. Visualizations can be made from the results. For Example: The Hive database in the HDFS is connected to the windows operating system using ODBC. The data is imported into MS Office excel through the ODBC.

4. CONCLUSION

In this paper, we have seen the influence of Micro blogging site twitter on the current trends and issues. The opinion mining on Twitter data helps us to analyze various brands commercially and also to analyze behaviors of people using social networks. The analysis of Twitter data is being done in various perspectives, the presence of words like good, bad and also emoticons in the tweets can be used to infer the sentiment. The Twitter users can be classified into positive, negative and neutral users based on the followers and the followees and their behaviors can be studied based on the tweeting and retweeting activity. Tweets can also be used to analyze the influence factor in elections and hence can be used to predict the results as Twitter is one of the key tools used by US presidential candidates to predict the outcomes.

Tweets can also by analyzed whether is sent by a human or a bot based on the sentiments. But the results discussed only are based on limited datasets and other logical issues. If the Opinion Mining is applied on bigger datasets taking all aspects into account will definitely lead to some convincing results which can be used for future analytics in social networking.

5. REFERENCES

[1] Malhar Anjaria and Ram Mohana Reddy Guddeti, " Influence Factor based opinion mining of Twitter data using supervised learning" sixth IEEE conference on COMSNETS, 6-10 Jan 2014, Bangalore, India.

[2] Ana c E S Lima and Leandro N de Castro, " Automatic sentiment Analysis of Twitter messages", fourth IEEE conference on CASoN, 21-23 Nov 2012, Sao Carlos.

[3] Dimitrios Kotsakos, Panos Sakkos, ioannis Katakis, Dimitrios Guanopulos, "#tag: Meme or Event?" IEEE/ACM conference on ASONAM, 17-20 Aug 2014, Beijing, China.

[4] Mahnaz Roshanaei and Shivakant Mishra, "An Analysis of positivity and Negativity attributes of users on Twitter ", IEEE/ACM conference on ASONAM, 17-20 Aug 2014, Beijing, China.

[5] Tuan Anh Hoang, William w Cohen, Ee-Peng Lim, Dovy Pierce, David R Redlawsk, "Politics, Sharing and emotion in Microblogs", IEEE/ACM conference on ASONAM, 2013, New York, USA.

[6] John P Dickerson, vadim Kagan, V S Subrahmanian, "Using sentiment to detect Bots on Twitter: Are Humans more opinionated than Bots? ", IEEE/ACM conference on ASONAM, 17-20 Aug 2014, Beijing, China.

[7] Alexander Pak and Patrick Paroubek. "Twitter as a corpus for sentiment analysis and opinion mining", Proceedings of the Seventh International Conference on Language Resources and Evaluation (LREC' 10), may 2010.

[8] https://en.wikipedia.org/wiki/Twitter

[9] http://www.thecloudavenue.com/2013/03/analyse-

[10] Tweets-using-flume-hadoop-and.htmlhttp://blog.cloudera.com/blog/2012/09/analyzing-

[11] Twitter-data-with-hadoop/