MJRET
Open Access
Multidisciplinary Journal of Research in Engineering and Technology
Journal homepage: www.mjret.in
ISSN:2348 - 6953

SOCIAL MEDIA DATA ANALYSIS USING HADOOP

Pratik Dhotre[1],Harshal Gawali[2],Trupti Thore[3],Prasad Govardhankar[4],
Prof. Ila Savant[5],

[1234]Department of Computer Engineering,
[5]Assistant Professor in Computer Engineering,
Marathwada Mitra Mandal's College of Engineering
Pune, India

Abstract: *As of now we know present industries and some survey companies are mainly taking decisions by data obtained from web. As we see WWW is a rich collection of data that is mainly in the form of unstructured data from which we can do analysis on those data which is collected on some situation or on a particular thing. In this paper, we are going to talk how effectively sentiment analysis is done on the data which is collected from the Twitter using Flume. Twitter is an online web application which contains rich amount of data that can be a structured, semi-structured and un-structured data. We can collect the data from the twitter by using BIGDATA eco-system using online streaming tool Flume. And doing analysis on Twitter is also difficult due to language that is used for comments. And, coming to analysis there are different types of analysis that can be done on the collected data, So here we are taking sentiment analysis. Here we have categorized this sentiment analysis into 3 groups like tweets that are having positive, moderate and negative comments.*

Keywords: *Twitter, API, Hashtag, Sentiment, Hadoop, HDFS, Apache Flume, NLP*

1. INTRODUCTION

A huge number of people express themselves on the Web in a number of ways through various platforms like blogs and social networks. Of these since the launch of twitter, micro blogging has become increasingly popular. Twitter allows users to publish small updates (140 character limit) to their profiles[1]. There are a number of polls conducted to gauge public mood; a prominent example would be election polls. These polls often require a lot of manual work and resources like time, money and the like. Is it possible to come up with an automated way of analyzing public mood that is cost-effective and at the same time reliable? We would like to perform a sentiment analysis on twitter that does just that: give reliable indicators of public mood[2]. For Example Sentiment analysis is the task of identifying whether the opinion expressed in a text is positive or negative in general, or about a given

topic. For example: *"I am so happy today, good morning to everyone"*, is a general positive text, and the text: *"Django is such a good movie, highly recommends 10/10"*, expresses positive sentiment toward the movie, named Django, which is considered as the topic of this text.

2. APACHE FLUME

After creating an application in the Twitter developer site we want to use the consumer key and secret along with the access token and secret values. By which we can access the Twitter and we can get the information that what we want exactly here we will get everything in JSON format and this is stored in the HDFS that we have given the location where to save all the data that comes from the Twitter [6]. The following is the configuration file that we want to use to get the Twitter data from the Twitter. The tweets are fetched depending upon hash tag used. For example *#obama* is hashtag the tweets of obama hashtag are fetched from Twitter.

3. TEXT PREPROCESSING

3.1 Preprocessing

3.1.1.1 Filtering

URLs: People use twitter not only for expressing their opinions but also for sharing information with others. Given the short maximum length of tweets, one way of sharing is using links[2]. Tweets include various links or URLs and these do not contribute to the sentiment of the tweet. The URLs in the data used in this project are of the form http://plurk.com/p/116r50. These do not contribute to the sentiment of the tweet. Hence the URLs are removed in preprocessing phase.

Usernames: Tweets often refer to other users and such references be- gin with the @ symbol. These again do not contribute to the sentiment and hence are replaced by the generic word USERNAME.

Duplicates or repeated characters: People use a lot of casual language on twitter. For example, 'happy' is used in the form of 'haaaaaappy'. Though this implies the same word 'happy', the classifiers consider these as two different words.

3.2 Stop-words removal

In information retrieval, there exists many words that are added as conjunctions in sentences. For example, words like the, and, before, while, and so on do not contribute to the sentiment of the tweet. Also these words do not help in classifying the tweets as they appear in all classes of tweets[4]. These words are removed from the data so as to avoid using them as features. The stop words corpus was obtained from NLTK. Some modifications were required to this as the corpus also had some negative words such as nor, not, neither which are important in identifying negative sentiments and should not be removed.

3.3 Afinn Dictionary

AFINN is a list of English words rated for valence with an integer between minus five (negative) and plus five (positive). The words have been manually labeled by Finn Årup Nielsen in 2009-2011. The Afinn dictionary has the sentimental words list with positive and negative category depending upon meaning of word. Afinn dictionary useful for finding sense and adjective of word.

4. ALGORITHM

We have used **porter stemming** algorithm

Description:-

The Porter stemming algorithm (or 'Porter stemmer') is a process for removing the commoner morphological and in flexional endings from words in English. Its main use is as part of a term normalization process that is usually done when setting up Information Retrieval systems. The porter stemming algorithm used to get rigid word. For example, the word Established is converted into its rigid word in establish.

Steps:-

Step 1: Gets rid of plurals and -ed or -ing suffixes
Step 2: Turns terminal y to i when there is another vowel in the stem
Step 3: Maps double suffixes to single ones: -ization, -ational, etc.
Step 4: Deals with suffixes, -full, -ness etc.
Step 5: Takes off -ant, -ence, etc.
Step 6: Removes a final -e

5. STANFORD CORENLP

Stanford CoreNLP is a Java natural language analysis library. Stanford CoreNLP integrates all our NLP tools, including the part-of-speech (POS) tagger, the named entity recognizer (NER), the parser, the coreference resolution system, and the sentiment analysis tools, and provides model files for analysis of English.

6. EQUATIONS

6.1 Naïve Bayes/Classifier

Naive Bayes is a simple model for classification. It is simple and works well on text categorization. We adopt multinomial Naive Bayes in our project. It assumes each feature is conditional independent to other features given the class. That is,

$$\mathrm{P(c|t)} = \frac{\mathrm{P(c)|P(t|c)}}{P(t)}$$

where c is a specific class and t is text we want to classify. P(c) and P(t) is the prior probabilities of this class and this text. And P(t | c) is the probability the text appears given

this class. In our case, the value of class c might be POSITIVE or NEGATIVE, and t is just a sentence [5].

6.2 Data and Methodology

We have tested our classifier on a set of real Twitter posts hand-annotated**.** We compute accuracy of the classifier on the whole evaluation dataset, i.e.

$$\text{Accuracy} = \frac{\text{N(Correct Classification)}}{N(all\ Classification)}$$

We measure the accuracy across the classifier's decision

$$\text{Decision} = \frac{\text{N(retrived documents)}}{N(all\ documents)}$$

7. FIGURES

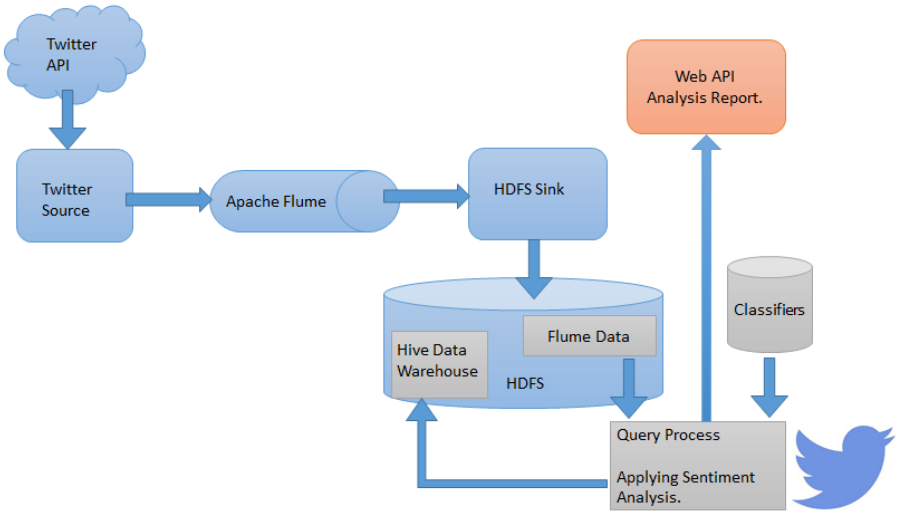
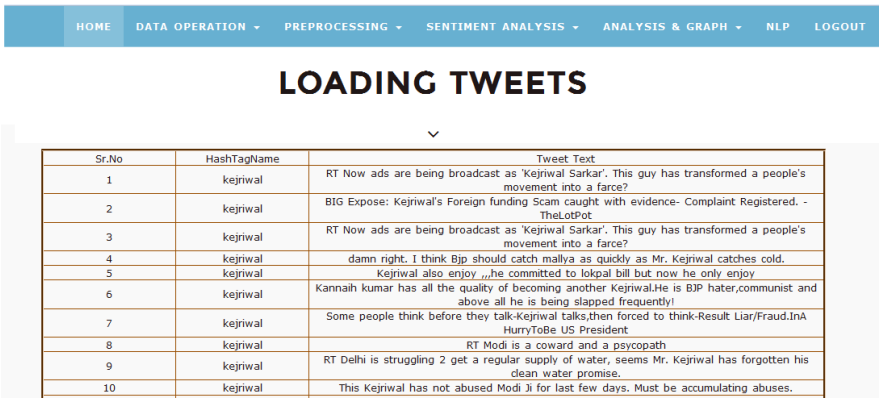


Fig.1. System Architecture

8. RESULT & ANALYSIS



Sr.No	HashTagName	Tweet Text
1	kejriwal	RT Now ads are being broadcast as 'Kejriwal Sarkar'. This guy has transformed a people's movement into a farce?
2	kejriwal	BIG Expose: Kejriwal's Foreign funding Scam caught with evidence- Complaint Registered. - TheLotPot
3	kejriwal	RT Now ads are being broadcast as 'Kejriwal Sarkar'. This guy has transformed a people's movement into a farce?
4	kejriwal	damn right. I think Bjp should catch mallya as quickly as Mr. Kejriwal catches cold.
5	kejriwal	Kejriwal also enjoy ,,,he committed to lokpal bill but now he only enjoy
6	kejriwal	Kannaih kumar has all the quality of becoming another Kejriwal.He is BJP hater,communist and above all he is being slapped frequently!
7	kejriwal	Some people think before they talk-Kejriwal talks,then forced to think-Result Liar/Fraud.InA HurryToBe US President
8	kejriwal	RT Modi is a coward and a psycopath
9	kejriwal	RT Delhi is struggling 2 get a regular supply of water, seems Mr. Kejriwal has forgotten his clean water promise.
10	kejriwal	This Kejriwal has not abused Modi Ji for last few days. Must be accumulating abuses.

Fig. 2. Loading tweets from twitter

Above Screenshot shows that the loading tweets from the twitter by using four key i.e. 1.consumer key 2.consumer secret key 3.access token key 4.access token secret key.Each twets has sr.no. i.e.no of the tweet,HashtagName that topic of tweet and Tweet Text is actual tweet.

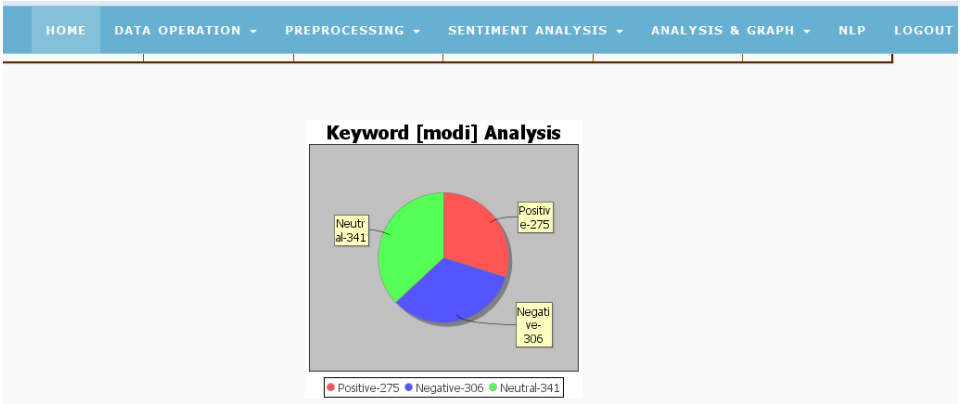


Fig.3. Analysis in Pie Chart

Above Screenshot shows that the Analysis of tweets in Negative or positive or neutral in pie-Chart.

9. CONCLUSION & FUTURE SCOPE

Sentiment analysis is an evolving field with a variety of use applications. Although sentiment analysis tasks are challenging due to their natural language processing origins, much progress has been made over the last few years due to the high demand for it. Not only do companies want to know how their products and services are perceived by consumers (and compare to competitors), but consumers want to know the opinions of others before making buying decisions. The growing need for product insights – and the technical challenges currently facing the field –will keep sentiment analysis and opinion mining relevant for the foreseeable future. Next-generation opinion mining systems need a deeper bind between complete knowledge bases with reasoning methods inspired by human thought and psychology. This will lead to a better understanding of natural language opinions and will more efficiently bridge the gap between unstructured information in the form of human thoughts and structured data that can be analyzed and processed by a machine.

ACKNOWLEDGEMENT

We would like articulate our deep gratitude to our guide Assistant Prof. Ila Savant who has been always our motivation for caring out the paper work. We special thanks to MMCOE institute for giving us such nice opportunity to work in a great environment. We thanks to our friend and colleague who have been source of inspiration and motivation that help us during dissertation time and to all other people who directly and indirectly support us to fulfill task.

REFERENCES

[1]. *Agarwal, A., Xie, B., Vovsha, I., Rambow, O., Passonneau, R.: Sentiment analysis of twitter data. In: Proc. ACL 2011 Workshop on Languages in Social Media. pp. 30–38 (2011)[2] S M Kim and E Hovy. 2004. Determining the sentimentof opinions. Coling.*

[2]. Pang and L. Lee. 2004. A sentimental education: Sentiment analysis using subjectivity analysis using subjectivity summarization based on minimum cuts. ACL.

[3]. Alec Go, Richa Bhayani and Lei Huang, Twitter Sentiment Classi_cation using Distant Supervision

[4]. Balamurali A.R., Aditya Joshi, Pushpak Bhattacharyya, Robust Sense Based Sentiment Classification, ACL WASSA 2011, Portland, USA, 2011

[5]. Liu, Bing, Sentiment Analysis and Opinion Mining, 5th Text Analytics Summit, Boston,June 1-2, 2009

[6]. Alec, G.; Lei, H.; and Richa, B. Twitter sentiment classification using distant supervision.

[7]. Technical report, Stanford University. 2009