Scientific Journal of Impact Factor (SJIF): 4.72

e-ISSN (O): 2348-4470
p-ISSN (P): 2348-6406

International Journal of Advance Engineering and Research Development

Volume 4, Issue 3, March -2017

Mining social Media Data for Understanding Different views of News

Rohit B. Pawar[1], Kishori B. Chavan [2], Shivani S. Gurav[3], Pooja G. Adake,
[4],Rohan R. Shelar[5], Prof. Kshirsagar A. P.[6]

[1, 2,3,4,5,6] *Department of Computer Science & Engineering, D.I.E.T Sajjangad Satara, Maharashtra India*

Abstract -*Recently, social media is playing a vital role in social networking and sharing of data. Social media is favored by many users as it is available to millions of people without any limitations to share their opinions, people share experience and concerns via their status.*

Twitter API, twitter4j, is processed to search for the tweets based on the geo location. People's posts on social network gives us a better concern to take decision about the particular system. Evaluating such data in social network is quite a challenging process.

In the proposed system, there will be a workflow to mine the data which integrates both qualitative analysis and large scale data mining technique. Based on the different prominent themes tweets will be categorized into different groups. It uses multi label classification technique as each label falls into different categories and all the attributes are independent to each other. Label based measures will be taken to analyze the results and comparing them with the existing sentiment analysis technique.

Keywords-*HDFS, Map-Reduce, Hive, Pig, Flume,Twitter*

I. INTRODUCTION

Now a days, various social media sites such as Twitter, Facebook, Linked In & YouTube etc. provide great platform for people to share their feelings, opinions, thinking on various social media sites. The challenge is to gather all such related data, detect and summarize the overall feelings on the topic.

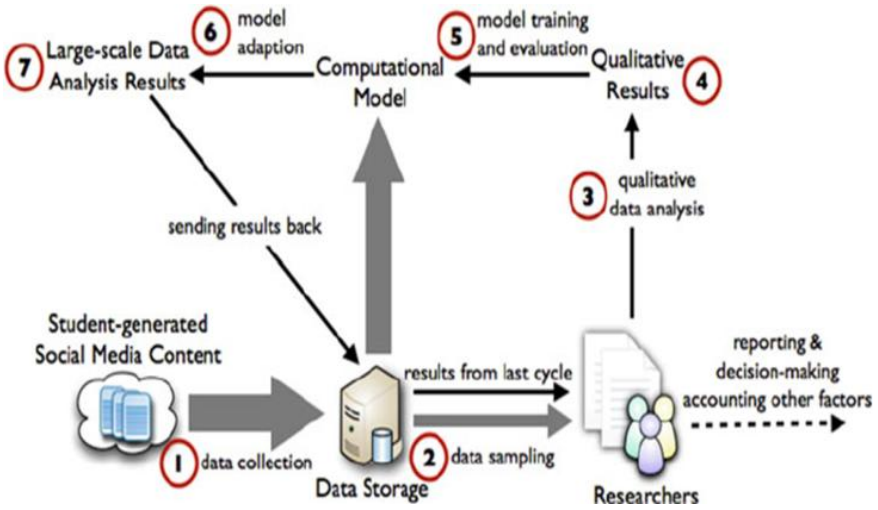

Fig.:- System Architecture

The large amount of social media data provides opportunities to understand people experiences, but also improves methodological difficulties in making sense of social media data for day to day life purposes. To explore engineering people informal conversation on twitter in order to understand issues and problems people encounter in their life Experiences.

Twitter is a popular social media site. Twitter provides API's that can be used to stream data therefore we choose to start from analyzing people posts on twitters.

II. Literature Survey

2.1 Lin, Jimmy, and Alek Kolcz. "Large-Scale Machine Learning at Twitter." In Proceedings of the 2012 ACM SIGMOD International Conference on Management of Data, pp. 793-804. ACM, 2012.[1]

This paper presents a case study of Twitter's integration of machine learning tools into its existing Hadoop-based, Pig-centric analytics platform. The core of this work lies in recent Pig extensions to provide predictive analytics

capabilities that incorporate machine learning, focused specifically on supervised classification. In particular, the authors have identified stochastic gradient descent techniques for online learning and ensemble methods as being highly amenable to scaling out to large amounts of data.

In contrast to other linguistic approaches the authors adopt a knowledge-poor, data-driven approach. It provides a base-line for classification accuracy from content, given only large amounts of data. The data set involves a test set consisting of one million English tweets with emoticons from Sept. 1, 2015, at least 20 characters in length. The test set was selected to contain an equal number of positive and negative examples. For training, they have prepared three separate datasets containing 1 million, 10 million, and 100 million English training examples from tweets before Sept. 1, 2015 (also containing an equal number of positive and negative examples). In preparing both the training and test sets, emoticons are removed.

2.2 Bian, Jiang, Umit Topaloglu, and Fan Yu. "Towards Large-Scale Twitter Mining for Drug-Related Adverse Events" In Proceedings of the 2012 international workshop on Smart health and wellbeing, pp. 25-32. ACM, 2012. [2]

In this paper, the authors describe an approach to find drug users and potential adverse events by analyzing the content of twitter messages utilizing Natural Language Processing (NLP) and to build Support Vector Machine (SVM) classifiers. Due to the size nature of the dataset (*i.e*., 2 billion Tweets), the experiments were conducted on a High Performance Computing (HPC) platform using Map Reduce, which exhibits the trend of big data analytics. The results suggest that daily-life social networking data could help early detection of important patient safety issues The data set used is a collection of over 2 billion Tweets collected from May 2009 to October 2010, from which they try to identify potential adverse events caused by drugs of interest. The collected stream of Tweets was organized by a timeline. The raw Twitter messages were crawled using the Twitter's user timeline API that contains information about the specific Tweet and the user. The work is indexed only with the following four fields for each Tweet:
1) Tweet id that uniquely identifies each Tweet;
2) User identifier associated with each Tweet;
3) Timestamp of the Tweet; and 4) the Tweet text.
They utilized the Amazon Elastic Compute Cloud (EC2) to run the Twitter indexers on 15 separate EC2 instances, 34.2 GB of memory, and 13 EC2 Compute Units) in parallel, which were able to parse and index all 2 billion Tweets within two days. The size of the Lucene indexes is 896 GB.
To mine Twitter messages for AEs, the process can be separated into two parts:
1) Identifying potential users of the drug;
2) Finding possible side effects mentioned in the users' Twitter timeline that might be caused by the use of the drug concerned. Both processes involve building and training classification models based on features extracted from the users' Twitter messages. Two-sets of features (*i.e*., textual and semantic features) are extracted from Twitter users' timeline for both classification models. Textual features such as the bag-of-words (BoWs) model are derived based our analysis of the actual Twitter messages. Semantic features are derived from the Unified Medical Language System (UMLS) Metathesaurus concept codes extracted from the Tweets using Metamap developed at the National Library of Medicine (NLM). Two-class Support Vector Machine (SVM) was used for the purpose of classification. Evaluation of the SVM was done using parameters such as, the Area under the Curve (AUC) value, and the Receiver operating characteristic (ROC) curve. The ROC curve using the mean values of the 1000 iterations was drawn. The prediction accuracy on average over the 1000 iterations was evaluated to 0.74 and the mean AUC value is 0.82.

2.3 Liu, Bingwei, Erik Blasch, Yu Chen, Dan Shen, and Genshe Chen. "Scalable Sentiment Classification for Big Data Analysis Using Naive Bayes Classifier" In Big Data, 2013 IEEE International Conference on, pp. 99-104. IEEE, 2013. [3]

Machine learning technologies are widely used in sentiment classification because of their ability to "learn" from the training dataset to predict or support decision making with relatively high accuracy. However, when the dataset is large, some algorithms might not scale up well. In this paper, the authors evaluate the scalability of Naive Bayes classifier (NBC) in large-scale datasets. They have presented a simple and complete system for sentiment mining on large datasets using a Naive Bayes classifier with the Hadoop framework. Instead of using Mahout Library, they implemented NBC to achieve fine-grain control of the analysis procedure for a Hadoop implementation. They haveInternational Journal of Database Theory and Application Vol.9, No.1 (2016) 80 Copyright ⓒ 2016 SERSC
Demonstrated that NBC is able to scale up to analyze the sentiment of millions movie reviews with increasing throughput.
The raw data comes from large sets of movie reviews collected by research communities. In their experiments, they use two datasets: the Cornell University movie review dataset3 and Stanford SNAP Amazon movie review dataset4. The Cornell dataset has 1000 positive and 1000 negative reviews. The Amazon movie review dataset is organized into eight lines for each review, with additional information such product identification (ID), user ID, profile Name, score, summary *etc.* They have used only unigrams for the Naive Bayes classifier. The classification task is divided into three sequential jobs as follows.
1) Training job - All training reviews are fed into this job to produce a model for all unique words with their frequency in positive and negative review documents respectively.

2) Combining job - In this job, the model and the test reviews are combined to a intermediate table with all necessary information for the final classification.
3) Classify job - This job classifies all reviews simultaneously and writes the classification results to HDFS. The experimental setup consists of a Virtual Hadoop cluster of seven nodes. It is a fast and easy way to test a Hadoop program in the Cloud, although the performance might be weaker compared to a physical Hadoop cluster. The cloud infrastructure is built on a Dell server with 12 Intel Xeon E5- 2630 2.3GHz cores and 32G memories. They tested their code on Cornell dataset and resulted in a 80.85% average accuracy. Without changing the Hadoop code, the program was able to classify different subsets of Amazon movie review dataset with comparable accuracy. To test the scalability of Naive Bayes classifier, the size of dataset in their experiment varies from one thousand to one million reviews in each class.

3.4 ÁlvaroCuesta, David F., and María D. R-Moreno. "A Framework for Massive Twitter Data Extraction and Analysis", In Malaysian Journal of Computer Science, pp 50-67 (2014):1. [4]

The authors propose an open framework to automatically collect and analyze data from Twitter's public streams. This is a customizable and extensible framework, so researchers can use it to test new techniques. The framework is complemented with a language-agnostic sentiment analysis module, which provides a set of tools to perform sentiment analysis of the collected tweets. The capabilities of this platform are illustrated with two study cases in Spanish, one related to a high impact event (the Boston Terror Attack), and another one related to regular political activity on Twitter. The first case study involves the activity on Twitter around a high impact event, the Boston Terror Attacks. In this case, they tracked a hash tag. The second case study was focused on regular Twitter usage, tracking the activity around well-known Spanish political actors, *i.e.* politicians, political parties, journalists and activist organizations as well. The authors have selected controversial accounts to have a good foundation for sentiment analysis.

There are several layers of processing and these modules need to interchange data among them, using open data formats such as JSON. Most tools in the framework are implemented in Python, but the Classifier and Tester web interfaces run on NodeJS and are programmed in CoffeeScript (a language which can be pre-processed into JavaScript). The chosen backend database is MongoDB, which is a good fit for our purposes since its atomic representation is JSON, just like tweets The implementation was based on the Natural Language Toolkit (NLTK) framework

A complete procedure of data extraction and sentiment analysis is divided into three separate steps: data acquisition, training for sentiment analysis and report generation. The first step is, gathering data from Twitter with the Miner. Then the classifier is trained and the sentiment analysis carried out. Finally, the platform generates a set of reports, including the sentiment analysis if it is enabled. Classification was done according to three classes, "positive", "negative" and "neutral". Several Naïve Bayes classifiers using a set of ngrams in order to select the one with the best performance. In particular, they have tried {1}, {2}, {3}, {1, 2}, {1, 3} and {2, 3} ngrams and minimum score of 0, 1, 2, 3, 4, 5, 6 and 10. All these different options were tried using ten-fold cross-validation to avoid biases induced by the partition of the training set. The parameters such as accuracy mean and variance, precision, recall and fmeasure mean and variance were used for evaluation. The conclusion is that the best trainers had 1-grams included and a minimum score between 2 and 4.

3.5 Skuza, Michal, and Andrzej Romanowski. "Sentiment analysis of Twitter Data within Big Data Distributed Environment for Stock Prediction" In Computer Science and Information Systems (FedCSIS), 2015 Federated Conference on, pp. 1349-1354. IEEE, 2015. [5]

This paper discusses a possibility of making prediction of stock market basing on classification of data coming from Twitter micro blogging platform.

Twitter messages are retrieved in real time using Twitter Streaming API. Tweets were collected over 3 month's period from 2nd January 2013 to 31st March 2013. It was specified in the query that tweets have to contain name of the company or hashtag of that name. Predictions were made for Apple Inc. in order to ensure that sufficiently large datasets would be retrieved Only tweets in English are used in this research work. Reposted messages are redundant for classification and were deleted. After pre-processing each message was saved as bag of words model – a standard technique of simplified information representation used in information retrieval. System design consists of four components: Retrieving Twitter data, pre-processing and saving to database (1), stock data retrieval (2), model building (3) and predicting future stock prices (4). Polarity mining is a part of sentiment in which input is classified either as positive or negative. Automatic sentiment detection of messages was achieved by employing SentiWordNet. Prediction of future stock prices is performed in this work by combining results of sentiment classification of tweets and stock prices from a past interval. Taking into consideration large volumes of data to be classified and the fact they are textual, Naïve Bayes method was chosen due to its fast training process even with large volumes of training data and the fact that is it is incremental. Considered large volumes of data resulted also in decision to apply a map reduce version of Naïve Bayes algorithm.

III. Classification of existing System

As we outlined in the previous section, there have been only one approach to the construction of natural language processing system. The only approach of those system is that they only support the static database or the database provide them by programmer. They don't support any change into that database. It was the serious limitation of those system.

In this article we will explore the system which supports the dynamic database of user to access it for different person of different organization. As others paper doesn't include in their system.

In the next section, we will explore the each of these approaches. It is important to emphasize that we do not claim one of these approaches to be better than others, as each of the approaches has its advantage and disadvantages. However, we will evaluate the convenience of each of these approaches in regards to the main goal of our research work: Knowledge Extraction from Database using NLP.

IV.Contribution of our approach compared to previous related work

The main improvement in our proposal compared to the other existing system is to support the Dynamic databases of user. This system supports the partial change or all changes in database of the end user. The previous system WASP(Word Alignment-based Semantic Parsing) is a system developed at the University of Texas by Yuk Wah Wong. This system was designed to get "complete,formal,symbolic,meaningful, representation of natural language sentence".The main drawback in this system is that it only parses given natural language input string and generates query, but could not execute query and take appropriate actions. This makes database untouched.

Another improvement in our system is that it not only generates query but also provides user expected result.

V.System modules

5.1 Text-Preprocessing

Text preprocessing is eliminating unwanted, meaningless contents from user tweets. Preprocessing module removes the symbols from tweets. Preprocessing means converting user tweets into proper readable form. Grammatical mistakes, spelling mistakes and unwanted words are handle also match with our dictionary. For ex. "Goooooood, soooo much, niceeee, woooow, okkkk" like this words replace with Good, so much, nice, wow, ok.

Use following dictionary:-

http://getintopc.com/softwares/education/oxford-dictionary-free-download/

5.2 Data Sampling

Before preprocessing data is available in an only unstructured format. That unstructured data was in unreadable format. After performing preprocessing we got data in readable format. Data (tweets) matched with the words available in dictionary, then rating passed to the matching tweets. According to rating input tweets was classifying into three ways such as Positive, Negative, Neutral.

5.3 Data Visualization

In this, We can show output by u graphical representation, such as

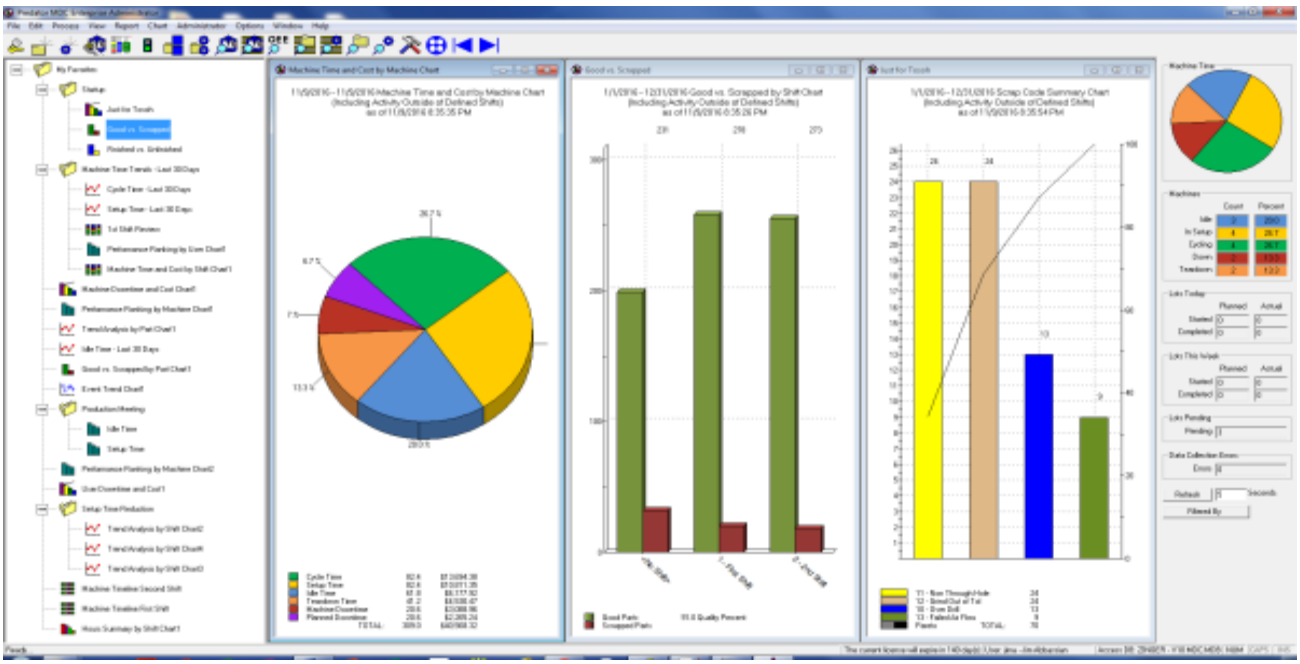

Fig. Data Visualization

VI. Conclusion

In our project, we try to understand opinions of people related to particular news and according to thatnews we analysis at three categories. Then we can represent this by graphical format.

It is use to analyzing social media data for different news.

It is proposed to stream real time live tweets from twitter using Twitter API, and the large volume of data makes the application suitable for Big Data Analytics. A method to predict or deduct the location of a tweet based on the tweet's information and the user's information should be found in the future.

VII. References

[1] Lin, Jimmy, and Alek Kolcz. "Large-Scale Machine Learning at Twitter." In Proceedings of the 2012 ACM SIGMOD International Conference on Management of Data, pp. 793-804. ACM, 2012.

[2] Bian, Jiang, Umit Topaloglu, and Fan Yu. "Towards Large-Scale Twitter Mining for Drug-Related Adverse Events" In Proceedings of the 2012 international workshop on Smart health and wellbeing, pp. 25-32. ACM, 2012.

[3] Liu, Bingwei, Erik Blasch, Yu Chen, Dan Shen, and Genshe Chen. "Scalable Sentiment Classification for Big Data Analysis Using Naive Bayes Classifier" In Big Data, 2013 IEEE International Conference on, pp. 99-104. IEEE, 2013.

[4] ÁlvaroCuesta, David F., and María D. R-Moreno. "A Framework for Massive Twitter Data Extraction and Analysis", In Malaysian Journal of Computer Science, pp 50-67 (2014):1.

[5] Skuza, Michal, and Andrzej Romanowski. "Sentiment analysis of Twitter Data within Big Data Distributed Environment for Stock Prediction" In Computer Science and Information Systems (FedCSIS), 2015 Federated Conference on, pp. 1349-1354. IEEE, 2015.

