An Efficient approach for Twitter Based Sentiment Analysis using Fuzzy C Means Clustering

Er. Amanpreet Kaur Chela, Dr. Harmaninderjeet Singh Sidhu
Research Scholor at Desh Bhagat University, Mandi Gobindgarh
Assistant Professor at Desh Bhagat University, Mandi Gobindgarh
(E-mail: preetk117@gmail.com, harmaninderjeet@gmail.com)

***Abstract* —** The objectivity and aspects or feature based sentiment analysis are two classified categories of sentiment analysis. Feature based sentiment analysis are part of movie review related to tweets then its exploration is done using objectivity based sentiment analysis. These tweets are related to love, miss, hate, etc. In general to analyze the sentiment from twitter data number of machine learning and symbolic techniques has been used. These days, sites like Twitter show the influence that surroundings have on online users. It is tough to process big data using traditional techniques. This big data can't be managed using existing analytics models and tools. In this paper, Apache Hadoop and its functionalities on Spark framework is used to analyse this large quantity of data. Computation on large datasets is performed using Hadoop as a framework and cloud computing is used to perform operations on distributed data in an efficient way. Further we have applied Fuzzy C Means clustering on tweets that helps in obtaining top terms and sentiment analysis similarity values. After that they are classified using support vector machine classifiers that will help in stripping of emotions in Twitter data. The result for different classifiers has been compared in terms to accuracy.

***Keywords*—** *Sentiment Analysis, Apache Hadoop, Fuzzy C Means, Support vector machine, Machine leanring*

I. INTRODUCTION

By increase in popularity of micro-blogging, blogging and social networking sites the amount of generated data is also getting increase. Internet is a collection of networks which has changed the way of people to express their feelings and thoughts. Online conversation frame, blog post helps people to review about the movies even before watching it in theatres that save their money and time [6]. The use of naive technique is not efficient for a person to analyze the large quantity of unreasonable information.

Each tweet emotions or opinion classification and identification is the main concern of sentiment analysis. The objectivity and aspects or feature based sentiment analysis are two classified category of sentiment analysis. Feature based sentiment analysis are part of movie review related to tweets then its exploration is done using objectivity based sentiment analysis [8]. These tweets are related to love, miss, hate, etc. In general to analyze the sentiment from twitter data number of machine learning and symbolic techniques has been used. Another way of defining sentiment analysis is model or system that takes document as input to analyze and then details document is generate. And then opinion of given input document is summarized. Firstly pre-processing is done in which stop words, emotions, white spaces, hash tags and repeating words are removed. Then training data is used to classify the tweets machine learning techniques in correct way. In machine learning techniques there is no need of database of words as required in case of knowledge based approach that makes is better and faster. The feature extraction from source text can be done using various methods. There are two phases of feature extraction the first one is twitter related data extraction by which tweet is transformed into normal text. Then more features are extracted and added into normal text in case of next phase. There is class label for each associated tweet in training data which is further pass into classifiers and trained them accordingly. Then test tweets are given to the model and classification is done with the help of various trained classifiers that results in classification of tweets into positive, negative and neutral. The classifier can be support vector machine, neural network, etc.

This complete paper is divided into several sections. Second section gives the description about sentiment analysis. Then Apache Hadoop is described in next section. In fourth section we have given a review on previous work done in this field. Next section gives problem in existing work that has been reviewed in previous section and along with problem the proposed approach used in this work is also given in brief. Then results obtained using proposed approach is given and ended with the conclusion.

II. SENTIMENT ANALYSIS

As mentioned in above section sentiment is an opinion or expression by an author about any aspect or object. It is also known as preferences, sentiment, extracting users, analyzing the opinion from subjective text. Text parsing is the main focus of sentiment analysis or in simple terms it can be defined as detection of text polarity. Further neutral, negative or positive are three types of polarity [7]. As user opinion is derived from it that's why it can also be referred as opinion mining. In understanding of user's perspective a great help is given using sentiment analysis as there is change in opinion from person to person. Sentiment can be direct opinion in

which opinion can be either negative or positive. Example of direct opinion can be poor video clarity of cell-phone.

Another is comparative statement that consists of identical objects comparison that comes under comparison opinion. Example of it can be a sentiment like picture quality of one camera is better than another [9]. There are three different levels of sentiment analysis as given below:

- Whether any given sentence is objective or subjective is given by sentiment analysis at sentence level identifiers. There is only one opinion in sentence that is assumes while analysis of sentence level.
- Opinion about particular entity is classified at sentiment analysis at document level. Complete document contains a single object opinion from single holder of opinion.
- From reviews a feature of particular object is extracted by sentiment analysis at feature level and determines whether the opinion is negative or positive. Then summarized report is produced after grouping the extracted features.

III. APACHE HADOOP

By use of Hadoop open source framework an Apache is used for parallel processing of large datasets across nodes cluster. MapReduce programming model and Hadoop distributed file system (HDFS) are major components of Hadoop. Commodity machine across clusters of nodes or Cloud computing services is used to run Hadoop that makes it accessible. Hadoop is intended to run on commodity hardware still it can handle failures in an efficient manner that makes it robust. To parallel handle a large data any number of nodes can be added in Hadoop cluster. It can keep multiple copies of data that help in overcoming the failures of hardware.

Different modules of Hadoop [11]:

- **Hadoop common utilities:** There is need of file system level abstractions and operating system level by Hadoop module that is provided by utilities and libraries of java. Hadoop common utilities facilitated java scripts and files are used to carry out the execution of Hadoop.
- **Hadoop Yam framework:** It carry out the managing of clusters resources and jobs scheduling.
- **Hadoop Distributed File System (HDFS):** A large amount of data is stored by Hadoop file system that gives easier access to stored data and results in high throughput.
- **MapReduce paradigm:** It enable the parallel processing of data.

IV. SPARK AND APACHE FLUME

Spark streaming in apache spark uses the same abstraction for real time event streaming. Spark is used to offer data streaming that groups incoming data into micro-batches for processing by the apache spark platform. This can be completed through a discretized stream that is used for representation of set of RDD (Resilient distributed datasets). This made the batch processing as available for streaming as well it allows MLib and GraphX. MLib and GraphX are used to operate with steaming data. [12] New Data Frame API is used with query optimizer that has equal performance for Scala, Java, Python, and R.

Apache Flume is a scattered, consistent, and available service for efficiently collecting, aggregating, and shifting large amounts of log data. Apache Flume is a tool/service/data ingestion mechanism for collecting aggregating and transporting large amounts of streaming data such as log files, events, etc… from various sources to a centralized data store. Flume is used to move the log data generated by application servers into HDFS at a higher speed.

Following objective function is minimized by formulating conventional FCM and the sum of the errors between intensity at every pixel and the centroids of each class with respect to membership and centroids.

$$J = \sum_{i,j} \sum_{k=1}^{c} \mu_{k(i,j)^q} \ \|I(i,j) - v_k\|^2$$

When high values are assigned to pixels then objective function is minimized. Pixels whose intensities are close to centroids and low values are assigned to pixels whose intensities are distant from the centroids. The degree of fuzziness of clustering is controlled by constant parameter and satisfies the condition of constant parameter greater than 1. By increase in value of q membership function become increasingly fuzzy.

V. PREVIOUS WORK

Various researchers have worked on movie review taken from tweets [1]. The emotional words identification from tweets by twitter sentiment analysis is tough. This is due to use of repeated characters, white spaces, slang words and misspellings. Despite of selected classifier improved sentiment analysis is achieved by feature vector. Results show that an improved accuracy has been achieved by the use of support vector machine and Naive Bayesian. In case of Naive Bayesian classifier accuracy achieved is 65% and 75% of accuracy is achieved using SVM. In [5] open source version of MapReduce is used to described and filtering, classification and clustering of machine learning library is apache Mahout is implemented on Hadoop. Mahout is used to implement machine learning algorithms that help in studying the machine learning algorithm to a Hadoop program. Even in case of large datasets an optimize algorithm scalability is achieved.

In [4] author have emphasized on Twitter data real time sentiment analysis and scalable. In terms of performance and scalability classification accuracy terms are used to describe the merits of proposed system. Useful features are extracted from posts using proposed methodology to represent a process of sentiment analysis. The systems for batch processing and real time system are used for sentiment analysis scalable systems. A system is presented for real-time sentiment analysis on Twitter streaming data towards presidential candidates. In [12] authors have focused analyzing the sentiment analysis task for Twitter that is considered as most popular micro blogging platform. Data is arrived at high frequency that makes tweets as important part to be analyse and used algorithms to process them under strict constraints of time and storage. It will be shown how to automatically collect a corpus for sentiment analysis and opinion mining purposes and then perform linguistic analysis of the collected corpus. Twitter provided set of APIs make it available of public tweets posted on twitter. The neutral, negative and positive sentiments are determined by constructing a sentiment classifier using corpus.

In [2] author has proposed a modified K-Means algorithm that has added the feature selection based on multilayer data clustering framework. Then a representative feature subset is selected using proposed algorithm to facilitate the clustering that results in reducing raw data set dimension. Besides, the selected feature subset has fewer missing values than the raw data set, which may improve the cluster accuracy. Partial distance strategy is used in proposed algorithm as a unique property. In [3] authors have proposed a concept of sentiment analysis and big data. The concepts are allowed to know more issues and challenges in the area of sentiment analysis on big data for the further research. In order to classify review or text from an ambiguous data use of machine learning approaches is prove to be important. Big data analytics contain in their review along with measurement of big data, approaches, issues and methods to predict the accuracy. For doing it they have used evaluation metrics/statistical analysis and ecosystem of the sentiment analysis on big data.

VI. PROBLEM AND PROPOSED APPROACH

From last few years a lot of fascination has been gain by social media and Twitter is one of the social media site. It is a significant platform for people where they express their opinions and views about anything. By the increase in its popularity its quality is also getting increased. Person moods can be analyze by the use of sentiment analysis that helps in deciding neutral, negative and positive views of person. Earlier use of sentiment analysis has been done for syntax or lexical feature extraction and assigning polarity label to every document. These days, sites like Twitter show the influence that surroundings have on online users. It is tough to process big data using traditional techniques. This big data can't be managed using existing analytics models and tools that create a need to use a cloud storage. So, in this work we have introduced a Hadoop that is used to store this large big data.

In this work, Apache Hadoop and its functionalities on Spark framework is used to analyse this large quantity of data. Computation on large datasets is performed using Hadoop as a framework and cloud computing is used to perform operations on distributed data in an efficient way. Further we have applied fuzzy c means clustering on tweets that helps in obtaining top terms and sentiment analysis similarity values. After that they are classified using SVM classifiers that will help in stripping of emotions in Twitter data. We have introduced it to improve the accuracy.

VII. RESULTS AND DISCUSSION

MapReduce and HDFS are two main components of Hadoop. The MapReduce make it possible to write applications in effective manner and also helps in processing a large amount of data in an efficient way [11]. Map Task and Reduce Task are two tasks of MapReduce paradigm. Inputs are captured by Map Task that is divided into pair of data. Then value/key pair are formed by dividing the data into tuples. Output from Map task is used as a input for Reduce task. Then smaller set of tuples are formed after dividing the tuples in the Map task.

The MapReduce component of the Hadoop framework schedules monitors the tasks and also re-executes the failed task. There is one Task tracker and single Job tracker in MapReduce paradigm that act as slave and master respectively. Then task is executed after directing the Task tracker by Job tracker and it also manages the resource, consumption, resource distribution tracks and availability. After that status information is given to Job tracker by Task tracker.

To implement this work we have used Python. Twitter data is taken as dataset and input that is also used to train various machine learning classifiers. The code will be integrated with the Twitter API using consumer keys/secrets and access token key/secrets. Finally, a training set will be prepared by creating three folders- positive, negative and neutral. Tweets will store as values in files and names of documents as tweet ids.

Twitter provides API functions to facilitate third-party users to access the data. There are two main types of Twitter APIs: the REST API and the stream API. The REST API supports queries to Twitter user accounts and tweets, and it usually has very strict limits on the query rate (e.g., 150 requests per hour). Although the REST API provides flexible access to Twitter data from almost every angle, the rate limits make it not suitable for collecting large amounts of Twitter data and monitoring updates. On the other hand, the stream API provides almost real-time access to Twitter's global stream of public tweets. The stream API produces near real-time samples of Twitter's public tweets in large amounts. The

streaming API is used to provide the push deliveries of tweets and other events to be happening on twitter.

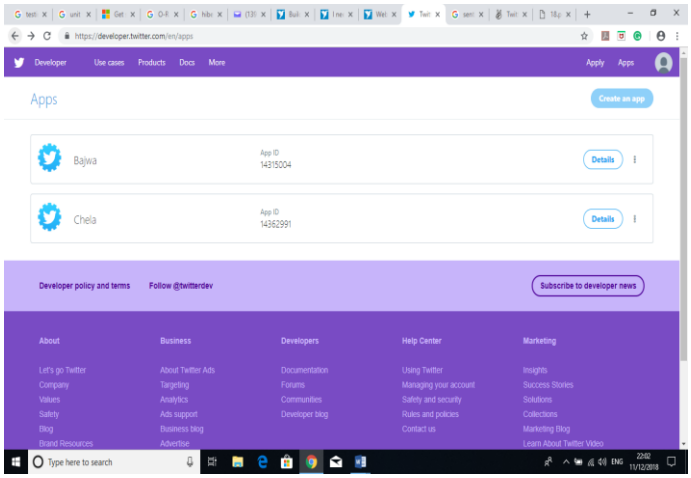

Fig. 1: Creation of two API's Bajwa & Chela

In Twitter API streaming we have created two API's name as Bajwa and Chela. Then keys and token are considered for both Bajwa and Chela as shown in below figures respectively.

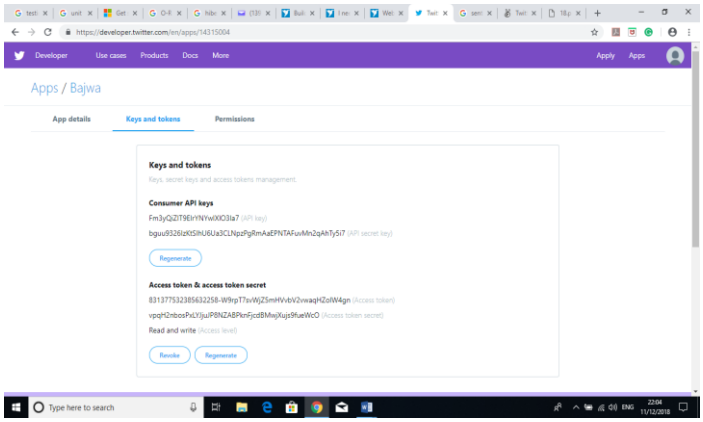

Fig. 2: Keys and Tokens for Bajwa

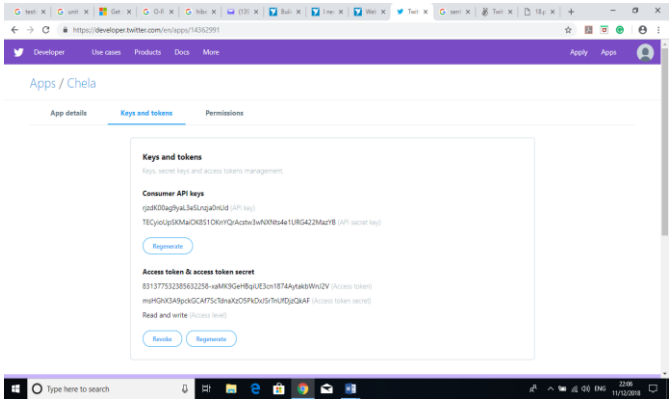

Fig. 2: Keys and Tokens for Chela

Once we get keys and tokens for both Bajwa and Chela then apply Fuzzy C Mean classiifer. Fuzzy C Means (FCM) is a method of clustering which allows one piece of data to belong to two or more clusters. This method is used frequently found in pattern recognition.

Algorithm of Fuzzy c-mean:

1. *Initialize* $U=[u_{ij}]$ *matrix,* $U^{(0)}$
2. *At k-step: calculate the centers vectors* $C^{(k)}=[c_j]$ *with* $U^{(k)}$

$$c_j = \frac{\sum_{i=1}^{N} u_{ij}^{m} \cdot x_i}{\sum_{i=1}^{N} u_{ij}^{m}}$$

3. *Update* $U^{(k)}$, $U^{(k+1)}$

$$u_{ij} = \frac{1}{\sum_{k=1}^{C} \left(\frac{\|x_i - c_j\|}{\|x_i - c_k\|} \right)^{\frac{2}{m-1}}}$$

4. *If* $\| U^{(k+1)} - U^{(k)} \| < \varepsilon$ *then STOP; otherwise return to step 2.*

In this work we have used MultinomialNB, LinearSVC, ComplimentNB, SGDC Classifier, Logisticregression, Nearest centroid, Passie aggressive classifer and Ridge classifier.

In this work Fuzzy c mean is applied on Twitter data and cluster will be made using training data and do predictions and compared with other used classifiers in terms of accuracy. Cluster is calculated in the first phase of fuzzy c means clustering and then points are assigned to centre using distance formula in the next phase. This will be performed till that cluster centers are not get stabilized. The algorithm similar to k-means clustering in many ways but it will give data items a membership value between 0 and 1. Then SVM is applied for classification purpose. Support Vector Machine is used for sentiment analysis that analyze the data, define decision boundaries and then computation is performed using kernels. Both regression and classification is supported by SVM that is prove to be useful in statistical learning theory and helps in recognizing various factors.

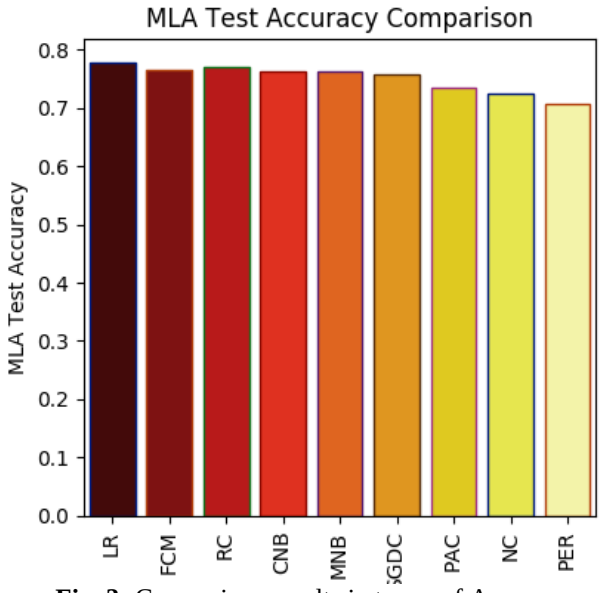

Fig. 3: Comparison results in terms of Accuracy

The results have been obtained by applying Fuzzy c mean on Twitter data and compared for other different classifier in terms of accuracy. The results show that use of Fuzzy c means gives better results in terms of accuracy.

VIII. CONCLUSION

Sentiment is an opinion or expression by an author about any aspect or object. It is also known as preferences, sentiment, extracting users, analyzing, and opinion from subjective text. Text parsing is the main focus of sentiment analysis or in simple terms it can be defined as detection of text polarity. Further neutral, negative or positive are three types of polarity. As user opinion is derived from it that's why it can also be referred as opinion mining. In understanding of user's perspective a great help is given using sentiment analysis as there is change in opinion from person to person. The existing approaches used for sentiment analysis are not proves to be efficient in large datasets. In this paper, we have introduced the cloud or Hadoop to handle the large dataset of twitter in which clusters are made using fuzzy c means and classified by SVM. The proposed approach has been compared with other existing approaches of classification. The results shows that use of SVM prove to be efficient in terms of accuracy.

IX. REFERENCES

[1] A. Amolik, N. Jivane, M. Bhandari, Dr. M. Venkatesan, "Twitter sentiment Analysis of Movie Reviews using Machine Learning Techniques", International Journal of Engineering and Technology (IJET),Vol. 7, pp. 2038-2044, 2016.

[2] G. Duan, W. Hu, Z. Zhang, "A Novel Multilayer Data Clustering Framework based on Feature Selection and Modified K-Means Algorithm", International Journal of Signal Processing, Image Processing and Pattern Recognition, Vol. 9, pp. 81-90, 2016.

[3] M. Edison, A. Aloysius, "Concepts and Methods of Sentiment Analysis on Big Data", International Journal of Innovative Research in Science, Engineering and Technology, Vol. 5, pp. 16288-16296, 2016.

[4] M. Karanasou, A. Ampla, C. Doulkeridis, M. Halkidi, "Scalable and Real-Time Sentiment Analysis of Twitter Data", 2016 IEEE 16th International Conference on Data Mining Workshops (ICDMW), 2016.

[5] B. Liu, E. Blasch, Y. Chen, D. Shen, G. Chen, "Scalable sentiment classification for Big Data analysis using Naïve Bayes Classifier", 2013 IEEE International Conference on Big Data, 2013.

[6] S. Mandal, "Comparison Of Results – Sentiment Analysis On Movie Reviews From Twitter Using Different Classifiers", International Journal of Computer Engineering and Applications, Vol. 10, pp. 56-62, 2016.

[7] N. M. "Characteristics of Twitter", Retrieved March 9, 2017, from http://impactoftwitter.weebly.com

[8] Anisha P. Rodrigues, Niranjan N. Chiplunkar, Anujna Rao, "Sentiment Analysis of Social Media Data using Hadoop Framework: A Survey", International Journal of Computer Applications, Vol. 151, pp. 7-10, 2016.

[9] A. Pak, P. Paroubek, "Twitter as a Corpus for Sentiment Analysis and Opinion Mining", In Proceedings of the International Conference on Language Resources and Evaluation, pp. 1320-1326, 2010.

[10] "Perform sentiment analysis in a big data environment", Retrieved February 18, 2017, from http://www.ibm.com/developerworks/library/ba-sentiment-analysis-big-data/index.html.

[11] S. Shang, M. Shi, W. Shang, Z. Hong, "Research on public opinion based on Big Data", 2015 IEEE/ACIS 14th International Conference on Computer and Information Science (ICIS), pp. 1-4, 2015.

[12] L. J. Sheela, "A Review of Sentiment Analysis in Twitter Data Using Hadoop", International Journal of Database Theory and Application, Vol. 9, Vol. 77-86, 2016.