SENTIMENT ANALYSIS OF TWITTER DATA IN HADOOP USING NAÏVE BAYES AND FUZZY C MEANS CLUSTERING

[1]Amanpreet Kaur Chela, [2]Harmaninder Jit Singh Sidhu
[1]Research Scholar, [2]Assistant Professor
[1,2]Department of Computer Science
[1,2]Desh Bhagat University, Mandi Gobindgarh, Punjab, India.

Abstract:

Social networking is one of the main platforms responsible for the increased amounts of data generation from the users. People across different geographical regions share their thoughts and put their opinions on the micro-blogging sites. One of the most popular micro blogging sites is Twitter where people share their reviews in the form of tweets. Due to concise and short limits of tweets, it is easier to analyze and thus extract valuable outcomes from tweets. The tweets also provide varied content of sentiments and opinions about the current technologies and affairs. Sentiment analysis is the process of analyzing various opinions and reviews given by people. Sentiment Analysis is the process which tends to understand these opinions and categorize them into positive, negative and neutral categories.

In this paper, the authors propose a concept for sentiment analysis that will help to classify various tweets on the basis of sentiment polarity. The streaming dataset from twitter will be stored in HDFS clusters which will be mined later using Naïve Bayes and Fuzzy C Means Algorithms to improve scalability and accuracy of various performance metrics and thus help any organization to formulate various strategies to promote their work process.

Keywords*: Sentiment Polarity, Accuracy, Scalability, Naïve Bayes, Sentiment Analysis.*

Introduction:

It has been found that Twitter is a source for gathering data on consumer's opinion about a product or different brands [1]. There is a reason as sentiment analysis over Twitter provides organizations or industries a rapid and effective way to monitor people's feelings about brand, product, and directors [2].The sentiment analysis on Twitter is emerging trend with researcher as it has potential applications. The challenges exclusive to this problem domain are widely attributed to the eminently informal talk of the micro blogging [8].

Sentiment analysis is concerned with the identification and classification of opinions or emotions on each post. Internet is the collection of networks and it is the biggest platform available to express one's view on a particular topic. With the increase in the popularity of social networking the way of humankind express their thoughts and feelings has changed. The mortals are connecting with each other with the help of cyberspace through the blog post, online conversation forums, micro-blogging and blogging websites, a huge quantity of data is generated. There are general social networking sites as well as sites specifically for writing reviews and opinions. These have huge collections of data, and when mined and analyzed, can be used for machine learning purposes [18]. Using sentiment analysis on data related to reviews, the product owner as well as the general public can know how well it is doing in the market. Since the proportion of such substance is enormous, it's

almost impossible for a patron to check all of them and due to that when sentiment analysis comes into place. There are various social media sites available but sentiment analysis will be done on Twitter because of its unique features.

Sentiment analysis is broadly classified in the two types first one is a feature or aspect based sentiment analysis and the other is objectivity based sentiment analysis. The tweets related to movie reviews come under the category of the feature based sentiment analysis. Objectivity based sentiment analysis does the exploration of the tweets which are related to the emotions like hate, miss, love etc. [4]

To correctly classify the tweets machine learning technique uses the training data. So, this technique does not require the database of words like used in knowledge-based approach and therefore, machine learning techniques are better and faster [4]. Accuracy is a measure of how often a sentiment rating was correct. For documents with tonality, accuracy tracks how many of those that were rated to have tonality were rated correctly, whereas scalability is the ability of a computer application or product (hardware or software) to continue to function well when it (or its context) is changed in size or volume in order to meet a user need.

Various methods are on tap for sentiment analysis and these methods provide different accuracy in the outcome. It also provides a feedback mechanism to the business organizations or industry so that they can improve product features that have negative reviews. Earlier it was easy to access the tweets but nowadays due to increased security for the unauthorized user, the accessing of reviews from tweets is not a simple task. Hence to access the tweets several steps need to follow such as generating an API (Application Program Interface). After that twitter grants an authorized key and consumer key. By possessing these keys user can access the tweets. While accessing tweets a condition needs to follow i.e. the user account should be three months old on twitter, otherwise, permission for accessing tweets won't be granted to the user.

The existing systems and tools for analytics are not efficient to deal with the complexity of big data. The present scenario of existing systems reveals following limitations. Firstly, the available systems like Real Time Twitter Trend Mining System and Twitter-Monitoring require thorough data cleaning, scraping and integration planning that will consecutively increase the overhead [3]. Secondly, the existing system is inefficient for real-time analytics. Also, it is a tedious process to analyze the bulk amount of data in a short span of time. The proposed system helps to reduce almost the above-mentioned drawbacks. The eight generally used Twitter sentiment analysis data file are demonstrated in [7].

Literature Survey:

A structure is implemented in Hadoop for analysis of Twitter data while forming a cluster of nodes [19]. Here the data is gathered by using Twitter API. Author has implemented a system where data is categorized in the form of positive, negative and neutral tweets. In this context, preprocessing is the technique for cleaning and preparation of text which are going to be classified. [5] In this paper author has compared 15 normally used preprocessing techniques on Twitter data. They applied different machine learning algorithms like Linear SVC,

Bernoulli Naïve Bayes, and Logistic Regression. In [6] the authors explored the Pre-processing techniques for two languages on news and emails. They used techniques like stemming, stop word removal, expansion of abbreviations, stop word removal, removal of non-alphabetic signs and negation handling with the addition of the prefix 'NOT'. They have implemented the system using SVM classifier and also correlated the number of features to its accuracy. In [9], the authors observed the effects of pre-processing on twitter dataset for the sentiment classification. They worked on the tweets which were full of folksonomy, symbols, unidentified words, and abbreviations. They recognized the significance of slang words and spelling correction. They took down the hash tags, URLs, stop words, user mentions and punctuation. They use an SVM classifier for their experiment. Cui [10] presented that most relevant classification for sentiment analysis is Support vector machine(SVM) since it can give better results for both opinions i.e. negative and positive terms, however in case of very small training data set the Naive Bayes classifier provides accurate results. The SVM classifier is known for high quality and to develop this quality of classifier, Support vector machine require a bulky training dataset. Zhen Niu [11] proposed a model to increase the efficacy of Naive Bayes. In this model, dynamite methods for computing the weight, classification and feature extraction are used. Authors used a Bayesian algorithm for the proposed model. In this work, to tune, the weights of the classifiers unique feature and representative features are provided. Representative featureis the information which represents a class and the information which can differentiate between the classes is called Unique feature. The probability of each classification is determined on the basis of weights and this feature refines the Bayesian algorithm.

The tweets related to movie reviews are given in [12]. Twitter sentiment analysis is difficult because it is very tough to identify emotional words from tweets and also due to the presence of the repeated characters, slang words, white spaces, misspellings etc. The feature vector aids in better sentiment analysis despite of classifier selected. Naïve Bayesian and Support vector machine performs well and also provide higher accuracy. The results show that they got 75 % accuracy from SVM and 65% accuracy from Naïve Bayesian classifier, come under the category of the feature based sentiment analysis. Apache Mahout is a machine learning library for clustering, classification and filtering, implemented on top of Hadoop, the open source version of Map Reduce described in [13]. Although there are some machine learning algorithms implemented in Mahout, it is still helpful to study how to convert a machine learning algorithm to a Hadoop program and to optimize the algorithm scalability in large datasets. The scalable and real-time sentiment analysis of Twitter data for system is emphasized in [14]. They demonstrate the merits of the proposed system, both in terms of classification accuracy as well as scalability and performance. They proposed a methodology for extracting useful features from posts in order to represent them in sentiment analysis process. Scalable systems for sentiment analysis can be categorized in real-time systems and systems for batch processing. A system is presented for real-time sentiment analysis on Twitter streaming data towards presidential candidates.

Twitter, the most popular micro blogging platform, for the task of sentiment analysis has focused in [14]. The tweets are important for analysis because data arrive at a high frequency and algorithms that process them must do so under very strict constraints of storage and time. 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. All public tweets posted on twitter are freely available through a set of APIs provided by Twitter. Using the corpus, a sentiment classifier, is constructed that is able to determine positive, negative and neutral sentiments. A novel

multilayer data clustering framework based on feature selection and modified K-Means algorithm are proposed in [15]. To facilitate the clustering, the proposed algorithm selects a representative feature subset to reduce the dimension of the raw data set. Besides, the selected feature subset has fewer missing values than the raw data set, which may improve the cluster accuracy. Another unique property of the proposed algorithm is the use of partial distance strategy. The concepts of big data and sentiment analysis are explored in [16]. The concepts are allowed to know more issues and challenges in the area of sentiment analysis on big data for the further research. Generally, the machine learning approaches are very important to classify the text or reviews from an ambiguous data. However, the notions are helpful to the researcher and whoever new to the area of applications, approaches and techniques in Sentiment Analysis. The overall literature contains analytics of big data, measurement of big data, issues, approaches, methods for predict the accuracy with a help of evaluation metrics/statistical analysis and ecosystem of the sentiment analysis on big data. The focus on the analysis of Hadoop framework for sentiment analysis of social media data are described in [17]. The major challenge with this extensive growth in the usage of social media is processing and analyzing the huge sets of data produced as a result. With the implementation of Map Reduce paradigm, Hadoop framework proves to be a reliable framework as it processes the huge sets of data in a fault- tolerant manner. The authors can implement the technologies such as Apache Pig, Hive, Sqoop, HBase, Zookeeper, and Flume on top of Hadoop in-order to improve the efficiency and performance of Hadoop.

Problem Formulation:

Big Data volumes bring new challenges in social media. It is relevant for developers to store, analyze and process massive datasets to offer and handle opportunities for new innovative solutions. Twitter data streaming to client presents opportunities for applying machine learning techniques and finding valuable insights, which can exponentially be useful for companies and marketing departments to learn about their current customers or attract new ones with interesting product offers.

In this paper, the authors propose a concept for sentiment analysis that will help to classify various tweets on the basis of sentiment polarity. The streaming dataset from twitter will be stored in HDFS clusters which will be mined later using Naïve Bayes and Fuzzy C Means Algorithms to improve scalability and accuracy of various performance metrics and thus help any organization to formulate various strategies to promote their work process. The proposed work flow is shown in Figure 1.

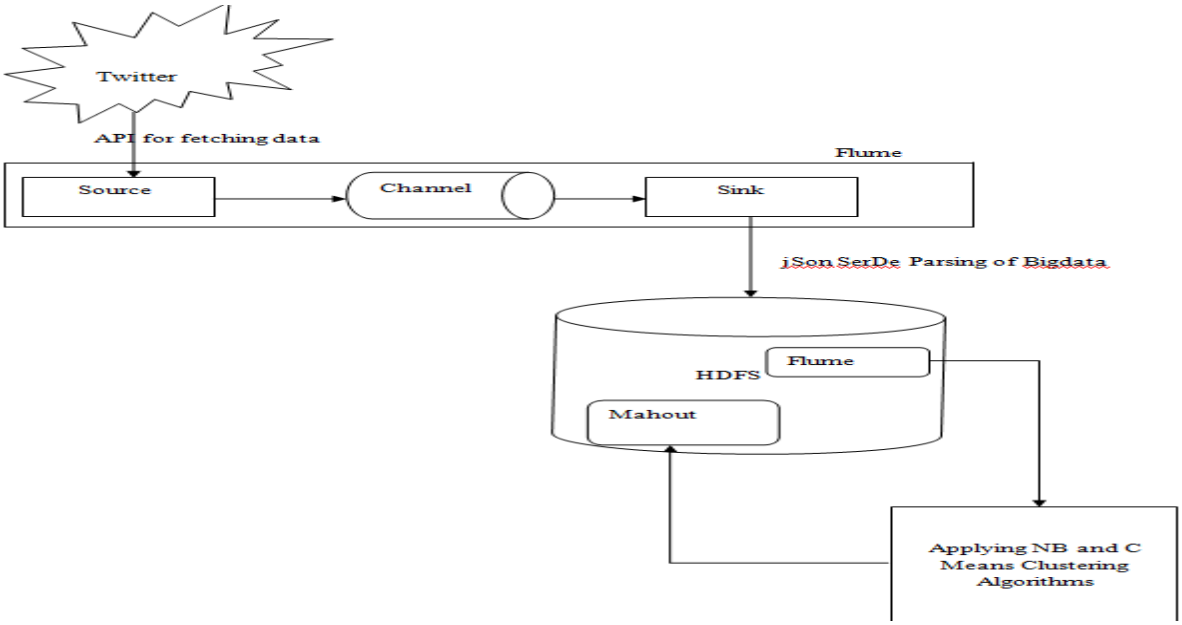

Fig. 1 Flow Diagram of Sentiment analysis of Twitter data with Hadoop

The authors propose that the streaming twitter data of Indian Currency Demonetization will be fetched from twitter services using Apache Flume and then transported to the centralized store of HDFS in Hadoop as shown in Figure 2. Machine learning algorithms will later be applied to this data on HDFS to perform sentiment analysis and classify emotions of people regarding this dataset. The authors propose two techniques to be applied: Naïve Bayes classification and Fuzzy C Means Clustering. The idea is to extract sentiments of people regarding their satisfaction or dissatisfaction with regard to their initiative of currency denomination. The accuracy of proposed work is expected to enhance by applying machine learning classifiers to plot in terms of precision, recall, g-mean and f-measure. Maximum Entropy and SVM Classifiers will be used for stripping of emotions in Twitter data to improve accuracy. These two classifiers will be used to categorize each emotion separately in terms of positive, negative or neutral. This will help in improving accuracy due to the individual definition of each sentiment for data set. Fuzzy c means clustering will be applied on tweets to obtain the top terms and similarity values for sentiment analysis. In this fuzzy c clustering will be applied to input dataset. Fuzzy c will help us to form clusters based on positive and negative reviews. Then it will help authors to calculate the percentage of people's opinion on positive, negative or neutral sides and also will help them to evaluate the similar values for each opinion.

Work Process using Fuzzy C Means Clustering Algorithm:

1. Fetch data from Twitter regarding "Indian Currency Demonetization" using twitter API.

2. Removal of unwanted data for review analysis.

3. Identification of word synonyms of different words and design the matrix.

4. Conversion of input data into <key,value> pairs.

5. Find importance of terms by using TF-IDF method. Weight can be assigned of each term Wi as:
Wi = tfi * log (D/dfi)

tfi is the term frequency of the term, D is the number of documents, dfi is the number of documents in which term is present. Sparse vectors are created from sequence files.

6. On the basis of calculated weight, fuzzy c-means clustering is applied.

a) Random selection of k points as the initial centroid.

b) Find the distance of each point from each of the centroid.

c) Create the membership matrix.

d) Total membership for a point in all the clusters must add to 1.

e) Generate new centroid for each cluster with iteration all these steps. Iteration will stop if centroids will be same as previous.

7. New clusters will be created after applying fuzzy clustering. Hence the top terms and similarity values between them are obtained for sentiment analysis.

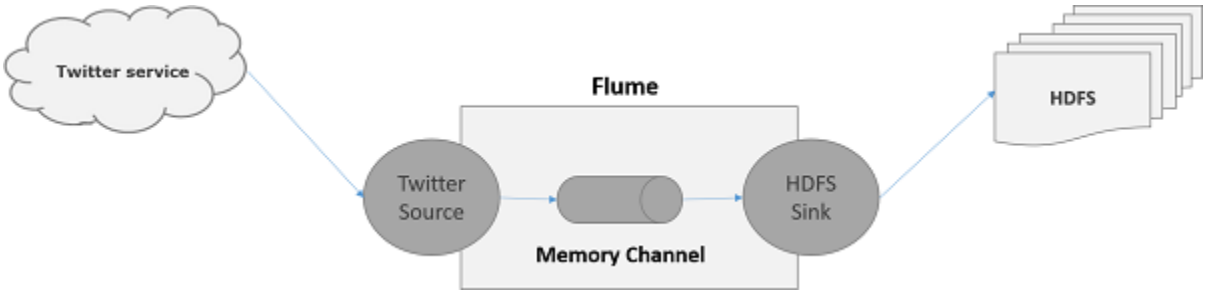

Fig. 2: Fetching Twitter data using Apache Flume in HDFS.

Conclusion and Future Work:

In context of this work, the authors have proposed work process for sentiment analysis that will help to classify various tweets on the basis of sentiment polarity using machine learning and flume. The streaming dataset from twitter will be stored in HDFS clusters which will be mined later using Naïve Bayes and Fuzzy C Means Algorithms to improve scalability and accuracy of various performance metrics and thus help any organization to formulate various strategies to promote their work process.

In the near future, the authors are planning to implement the proposed work and try to extend and improve this work by exploring more features that may be added in the feature vector and will increase the classification performance. One other future work will be the experimentation with different clusters so as to better evaluate the performance of Hadoop in regards to time and scalability. Moreover, the authors plan to compare the

classification performance of proposed work with classification methods, such as Naive Bayes or Support Vector Machines.

References:

[1] Jansen, B.J., M. Zhang, K. Sobel, and A. Chowdury, “Twitter power: Tweets as electronic word of mouth”, Journal of the American Society for Information Science and Technology 60:2169–2188,2009.

[2] Saif, H., Y. He., H. Alani, “Semantic sentiment analysis of twitter”, In The Semantic Web–ISWC, Springer, 508–524, 2012.

[3] Gaurav D Rajurkar, Rajeshwari M Goudar, "A speedy data uploading approach for Twitter Trend And Sentiment Analysis using HADOOP", HADOOP, 2015 International Conference on Computing Communication Control and Automation

[4] Amolik, A., Jivane, N., Bhandari, M., & Venkatesan, Dr. M, “Twitter sentiment Analysis of Movie Reviews using Machine Learning Techniques”, International Journal of Engineering and Technology (IJET),*7*(6), 2038-2044, 2016.

[5]Effrosynidis D., Symeonidis S., Arampatzis A, “ A Comparison of Pre-processing Techniques for Twitter Sentiment Analysis”, In: Kamps J., Tsakonas G., Manolopoulos Y., Iliadis L., Karydis I. (eds) Research and Advanced Technology for Digital Libraries. Lecture Notes in Computer Science, Springer, Cham, vol 10450, 2017.

[6] Uysal, A.K., G¨unal, S, “The impact of preprocessing on text classification. Inf. Process. Manage”, 50(1), doi:10.1016/j.ipm.2013.08.006, 104–112, 2014.

[7] Saif, H., Fern´andez, M., He, Y., Alani, H, “Evaluation datasets for twitter sentiment analysis: a survey and a new dataset, the STS-gold”, In Proceedings of the First International Workshop on Emotion and Sentiment in Social and Expressive Media: Approaches and Perspectives from AI, Turin, Italy, pp. 9–21, 2013.

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

[9] Singh, T., Kumari, M, “Role of text pre-processing in twitter sentiment analysis”, Proc. Comput.Sci.89, pp 549–554, 2016.

[10] H. Cui, V. Mittal, and M. Datar, “Comparative Experiments on Sentiment Classification for Online Product Reviews.”, In Proceedings of AAAI-06, pp.1265-1270,2006.

[11] Z. Niu, Z. Yin, and X. Kong, “Sentiment classification for micro blog by machine learning,” in Computational and Information Sciences (ICCIS), 2012,Fourth International Conference on, pp. 286–289, IEEE, 2012.

[12] Karanasou, M., Ampla, A., Doulkeridis, C., & Halkidi, M, “Scalable and Real-Time Sentiment Analysis of Twitter Data”, IEEE 16th International Conference on Data Mining Workshops (ICDMW). doi:10.1109/icdmw.2016.0138, 2016.

[13] Liu, B., Blasch, E., Chen, Y., Shen, D., & Chen, G, “Scalable sentiment classification for Big Data analysis using Naïve Bayes Classifier”, IEEE International Conference on Big Data. doi:10.1109/bigdata.2013.6691740, 2013.

[14] Sheela, L. J, "A Review of Sentiment Analysis in Twitter Data Using Hadoop", International Journal of Database Theory and Application,*9*(1), 77-86. doi:10.14257, 2016.

[15] Duan, G., Hu, W., & Zhang, Z, "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,9(4), 81-90. doi:10.14257/ijsip.2016.9.4.08, 2016.

[16] Edison, M., & Aloysius, A, "Concepts and Methods of Sentiment Analysis on Big Data", International Journal of Innovative Research in Science, Engineering and Technology,*5*(9), 16288-16296. doi:10.15680/IJIRSET.2016.0509102, 2016.

[17] P., A., N., N., & Rao, A, "Sentiment Analysis of Social Media Data using Hadoop Framework: A Survey", International Journal of Computer Applications,151(6), 7-10. doi:10.5120/ijca2016911833, 2016.

[18] Pak, A., & Paroubek, P, "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.

[19] Ajinkya Ingle, Anjali Kante, Shriya Samak, Anita Kumari, "Sentiment Analysis of Twitter Data Using Hadoop", International Journal of Engineering Research and General Science Volume 3, Issue 6, November-December, 2015.