JOURNAL OF CRITICAL REVIEWS

ISSN- 2394-5125 VOL 7, ISSUE 15, 2020

SENTIMENT ANALYSIS OVER TWITTER BIGDATA USING MODIFIED MEANSHIFT CLUSTERING ALGORITHM

Sharon Susan Jacob[1], Dr. R. Vijayakumar [2]

[1]Research Scholar, School of Computer Sciences, Mahatma Gandhi University, Kottayam, Kerala, India.
[2]Former Professor & Dean, School of Computer Sciences, Mahatma Gandhi University, Kottayam, Kerala, India
Emails:[1]sharonsusanjacob@gmail.com, [2]vijayakumar@mgu.ac.in

Received: 14 March 2020 Revised and Accepted: 8 July 2020

ABSTRACT: The amount of digital data that exists is growing at a rapid rate, doubling e very two years. A huge repository of terabytes of data is generated each day and analysis of these massive data requires a lot of efforts to extract knowledge for decision making. This means we need to have the technical tools, algorithms, and models to clean, process, and understand the available data in its different forms for decision-making purposes. The aim of this work is to develop a clustering technique to extract relevant information from unstructured Bigdata. Twitter is one of the important and popular social media where people may express their views/opinions or emotions freely. Thus, twitter is some of the central source of Bigdata and taken as the dataset. Cluster based techniques on sentiment analysis is a novel approach for analyzing sentiments expressed in social media sites. This paper presents the analysis of twitter feeds by implementing the algorithms Bisecting K Means, basic Meanshift algorithm and its modified form Gaussian mixture model based Meanshift and a comparison is made with these methods based on metrics namely accuracy, precision, recall and F1 score. For analyzing the
data, Python programming and PySpark DataFrame was used. The experiments perform on big data of tweets from a dataset of one lakh and show that the work provides an accuracy of 87% on detecting whether a tweet is "positive", "negative", or "neutral".

KEYWORDS: Bigdata, Meanshift Clustering, Gaussian mixture model, Bisecting K means, Sentiment analysis.

I. INTRODUCTION

Big Data is used to describe the massive volume of both structured and unstructured data that is so large and is difficult to process it using traditional techniques. Bigdata essentially describes data that falls under three categories: Fast, Large and Complex or 3V's viz., Velocity, Volume and Variety. Velocity point out the speed at which data is getting generated.Volume refers to the amount of Data that is getting generated. Variety indicates the different types of data that is getting generated. The best use case of big data is the data that keeps flowing on social media networks like, Facebook, Twitter, etc. The data is collected and observed in the form of comments, images, social statuses etc. Twitter is one of the important and popular social media where anyone can post tweets about any event and thus contains huge amount of data. Sentiment analysis or opinion mining is nothing but analysis of opinions or emotions from text data. Sentiment analysis identifies opinion or sentiment of each person with respect to specific event.

Clustering is the process of reducing a set of data by grouping the data objects such that objects within the same cluster are similar to each other. Clustering is the most popular unsupervised and exploratory data analysis [17]. Cluster based techniques on sentiment analysis is a novel approach for analyzing sentiments expressed in social media sites. This paper mainly focuses on the extracting the emotions from Bigdata of tweets using GMM based Meanshift clustering technique. **Meanshift** is falling under the category of a clustering algorithm in contrast of Unsupervised learning that assigns the data points to the clusters iteratively by shifting points towards the mode (mode is the highest density of data points in the region, in the context of the Meanshift). As such, it is also known as the **Mode-seeking algorithm**. The problem with the algorithm is that, it does not work well in case of high dimension and we do not have any direct control on the number of clusters but in some applications, we need a specific number of clusters. To overcome this problem, Gaussian Mixture model can be used.**Gaussian mixture models** are a probabilistic model for representing normally distributed subpopulations within an overall population. Mixture models in general don't require knowing which subpopulation a data

ISSN- 2394-5125 VOL 7, ISSUE 15, 2020

point belongs to, allowing the model to learn the subpopulations automatically. Since subpopulation assignment is not known, this constitutes a form of unsupervised learning[22].

1.1 Motivation and need of the study

Day by day, with the advent of technology, data has grown rapidly not only in size but also in variety and theanalysis of all this data is required. Data mining technique is used to extract relevant information among data but it faces difficulties to analyze Bigdata. Clustering is one of the key techniques in which mining is performed by finding out clusters having similar group of data. Companies use big data techniques to understand the customers' requirements and check what they say on social media. This helps companies to analyse and come up strategies that will be beneficial for the company's growth. Traditional and existing clustering algorithms are best suited with small datasets. Scalability is an important aspect that should be considered for an efficient clustering method. So, we had a motivation to develop improved clustering methods for analysis of Bigdata. In this paper, we have discussed an improved version of Meanshift clustering algorithm on Bigadata of twitter feeds.

The remainder of this paper is organized as follows: In Section II. Related works are discussed. Section III presents Methodology. Experimental analysis and results are described in Section IV and conclusion is given in Section V.

II. RELATED WORK

Damir Demirović [1] explained the detailed implementation of meanshift algorithm. An analysis of the algorithm, a well-commented implementation, a short overview of kernel density estimation and kernel function and insight into the working of the menashift algorithm are provided.

In [2] Comaniciu et al. present two solutions for the scale-space problem. The first is completely non-parametric and based on the adaptive estimation of the normalized density gradient. They define variable bandwidth mean shift and show superiority over the fixed bandwidth procedure.

In [3] Meng et al. propose a novel adaptive bandwidth strategy that combines different adaptive bandwidth strategies and bidirectional adaptive bandwidth mean shift which have the ability to escape from the local maximum density.

M. Farhadloo et al. [4] proposed multiclass sentiment analysis for English language using clustering and score representation. The model used aspect level sentiment analysis. Bag of nouns was preferred instead of bag of words to enhance clustering results, score representation and more accurate sentiment identification.

Chunxu Wu [5] proposed a method for synthesizing the semantic orientations of context-dependent opinions that cannot be determined using WordNet. This method is utilized to decide the sentiment of opinions by utilizing semantic closeness measures. This approach relies on such measures to determine the orientation of reviews when there is insufficient relevant information.

Sharma et al. [6] proposed an unsupervised document based sentiment analysis system able to determine the sentiment orientation of text documents based on their polarities. This system categorizes documents as positive and negative [6,7] and extracts sentiment words from document collections, classifying them according to their polarities.

Musto et al. [8] proposed a lexicon-based approach to identify the sentiment of any given tweet T, which began by breaking down the tweet into a number of small-scale phrases, such as as indicated by the part signs occurring in the content. Punctuations, adverbs and conjunctions constituted the part signal and, at whatever point a part signal occurred in the text, another micro-phrase is constructed.

X.Hu et al. [9] exploited emotional signals to detect sentiments appearing in social media data. These emotional signals were defined as any information that correlated or was associated with sentiment polarities.

Paltoglou and Thelwall [10] employed a lexicon-based approach to estimate the level of emotional intensity to make predictions. This approach was appropriate for detection of subjective texts expressing opinion and for sentiment polarity classification to decide whether the given text was positive or negative.

Turney [11] used bag-of-words method in which the relationships between words was not considered at all for sentiment analysis and a sentence is simply considered as a collection of words. To determine the sentiment for the whole sentence, sentiment of every individual word was determined separately and those values are aggregated using some aggregation functions.

Georgescu et al. [12] proposed an accelerated MS by performing a fast nearest neighbour search with spatially coherent hash tables. For the same scope, another acceleration technique is performed via locality sensitive hashing, which approximates the adjacent feature space elements around the mean.

C Xiao et al. [13] proposed a MS procedure using a reduced feature space. This reduced feature space is used for the adaptive clustering of the original data set, and is generated by applying adaptive KD-tree in a high dimensional affinity space.

III. METHODOLOGY

ISSN- 2394-5125 VOL 7, ISSUE 15, 2020

The work performs clustering of twitter feeds into three clusters positive, negative, and neutral with a given set of Bigdata of tweets. Twitter sentiment analysis was performed using the language Python and PySpark. DataFrame in Apache Spark has the ability to handle massive amount of data and has support for wide range of data format and sources. While Apache Spark is an open-source, cluster-computing framework, Python is a general-purpose, high-level programming language with an array of useful libraries. Thus, PySpark is a Python API for Spark that allows to leverage the simplicity of Python and speed and power of Apache Spark for BigData applications. The following conceptual diagram illustrates the steps involved in analyzing Bigdata for our sentiment analysis.

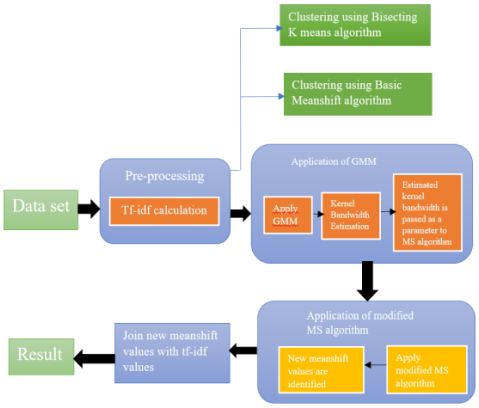

Fig.1Diagram showing Existing and Proposed Framework

A. Dataset Description:

The data set that we have collected is Twitter Sentiment Analysis [21], which is available online with free access. It comprises of around 1,00,000 tweets.

B. Data Pre-processing

Unstructured data does not lend itself to the programming tasks. It has to be processed in various different ways as applicable, to be able to serve as an input to any machine learning algorithm. Data pre-processing is the most vital step in unstructured data analysis and Spark offers these techniques out of the box through the ml.features package. Most of the techniques aim to convert text data to concise numerical vectors that can be easily consumed by machine learning algorithms.

1. Create RDD

In the first step, create RDD(Resilient Distributed Datasets). There are two ways to create RDDs: parallelizing an existing collection in your driver program, or referencing a dataset in an external storage system. RDD offers various transformations to parse and process the unstructured data like map, flatMap, filter, union, reduceByKey, etc.

2. TF-IDF Calculation

TF-IDF is a common pre-processing step for other machine learning algorithms. It is essentially used to vectorize text (convert text into vector) for further processing. [22]. TF-IDF stands for **"Term Frequency — Inverse Document Frequency"**. This is a technique to quantify a word in documents, we generally compute a weight to each word which signifies the importance of the word in the document and corpus.

TF-IDF = Term Frequency (TF) * Inverse Document Frequency (IDF) (1)

$$tf\text{-}idf(t,d) = tf(t, d) * \log(N/(df + 1)) \quad (2)$$

Where,

- **t** — term (word)
- **d** — document (set of words)
- **N** — count of corpus

ISSN- 2394-5125 VOL 7, ISSUE 15, 2020

- **corpus** — the total document set

C. SENTIMENT CLUSTERING USING MACHINE LEARNING ALGORITHMS

Here we used unsupervised machine learning approach for clustering based on sentiments. In the taken dataset Clustering have been carried out with two existing techniques viz., Bisecting K Means and basic Meanshift algorithm.Then a modified Meanshift clustering technique is used to find out whether there is any improvement in clustering.

1. Sentiment Clustering using Bisecting K Means Algorithm

Bisecting k-means is a hybrid approach between Divisive Hierarchical Clustering and K-means Clustering. Instead of partitioning the data set into K clusters in each iteration, bisecting k-means algorithm splits one cluster into two sub clusters at each bisecting step (by using k-means) until k clusters are obtained. Here, instead of using 3 clusters for data point updation, 6 clusters are used to find the actual cluster.Analysis with this method yields 57% of accuracy, 73% of precision, 40% of recall and 52% of F1 score.

2. Sentiment Clustering using Basic Meanshift Algorithm

Mean Shift is a hierarchical clustering algorithm. Mean-shift algorithm basically assigns the datapoints to the clusters iteratively by shifting points towards the highest density of datapoints i.e. cluster centroid. Analysis with this method provides 77% of accuracy, 86% of precision, 65% of recall and 74% of F1 score.

3. Sentiment Clustering using modified Meanshift Algorithm based on Gaussian mixture model

This is our proposed method for improving the performance of clustering over Twitter Bigadata.

In the existing Meanshift clustering algorithm, we used a flat kernel for finding out the weight of each neighborhood point. But in our proposed algorithm, a gaussian is used for finding out the weight. Using Gaussian mixture model (GMM) we found out the cluster centers of data points. Once the GMM algorithm converges, we will get some cluster centers for these data points. These cluster centers and datapoints together are used for estimating the kernel bandwidth. According to Meanshift Gaussian, we pass this kernel bandwidth as a parameter to the Meanshift algorithm. The gaussian kernel has two parameters, i) Distance calculated using Euclidian distance between the neighbor and current data point. ii) Kernel bandwidth, that was estimated from gaussian cluster centers and data points. The algorithm converges when all the meanshift values are found out. These meanshift values are reversed transformed by joining the meanshift value with tf-idf which will more accurately find out whether the tweet is of which cluster i.e., positive, negative or neutral.

Steps in proposed work

Step1: Input Dataset
Step2: Create RDD
Step3: For each sentence, calculate the Tf-Idf value
Step4: Apply Gaussian mixture model on the dataset
Step5: Estimate the kernel bandwidth parameter by using data centers identified by Gaussian mixture model
Step6: Pass the estimated kernel bandwidth parameter to Meanshift clustering
Step7: Apply Gaussian mixture model based Meanshift clustering on the dataset
Step8: Join the new meanshift values with tf-idf
Step9: Performance measurement

IV. EXPERIMENTAL ANALYSIS & RESULTS

In this section, we illustrate experimental results of our proposed Gaussian mixture model based Meanshift clustering algorithm. This yields increasingly good results on dataset. The result of this analysis is then compared with other two techniques viz., Bisecting K means Clustering and the basic Meanshift Clustering. When comparing, our proposed method was the one that obtained the most robust results. The metrics used here for evaluating the performance are *accuracy*, *precision*, *recall* and *F1-score*.

$$Accuracy = \frac{tp+tn}{tp+tn+fp+fn} \quad (3)$$

$$Precision = \frac{tp}{tp+fp} \quad (4)$$

ISSN- 2394-5125 VOL 7, ISSUE 15, 2020

$$Recall = \frac{tp}{tp+fn} \quad (5)$$

$$F1\ Score = 2 * \frac{(recall*prcision)}{(recall+prcision)} \quad (6)$$

where,
$\mathbf{t_p}$ = true positive which correctly predicted positive values
$\mathbf{t_n}$ = true negative which correctly predicted negative values
$\mathbf{f_p}$= false positive which is falsely predicted positive class
$\mathbf{f_n}$ = false negative which falsely predicted negative class.

The results are demonstrated in the table below:

Table 1Metric Representation of Big Data (Twitterdata)

Technique	*Accuracy*	*Precision*	*Recall*	*F1 score*
Bisecting K means	0.5778	0.7361	0.4052	0.5226
Meanshift Clustering	0.7775	0.8695	0.6568	0.7483
Meanshift with Gaussian	0.8737	0.9022	0.7047	0.7913

Our results demonstrated that the accuracy or the percentage of correctly clustered instances obtained with Bisecting K Means clustering technique on Big data is 57.7 % where as Meanshift clustering produces 77.7% accuracy. The proposed method Gaussian mixture model based Meanshift clustering outperforms these two methods with an overall accuracy of 87.3%. The precision, recall and F1 score measures obtained by the proposed method are 0.90, 0.70 and 0.79 whereas Bisecting K Means produced 0.73 Precision, 0.40 Recall and0.52 F1 score value. Meanshift clustering produced Precision as 0.86, Recall as 0.65 and 0.74 F1 score value. However, even better results are achieved when using our proposed algorithm.

Performance of the methods Bisecting K means, normal Meanshift and Meanshift with Gaussian Mixture Model are represented in the graph below:

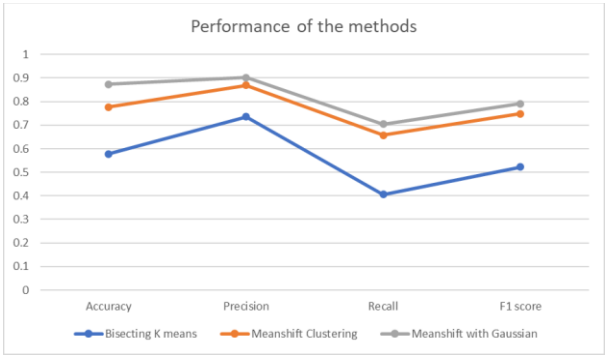

Fig.2Performance of GMM based Meanshift clustering

Comparison of the Accuracy, Precision, Recall and F1 score values obtained by the proposed method and other two methods are represented in the graph below:

ISSN- 2394-5125 VOL 7, ISSUE 15, 2020

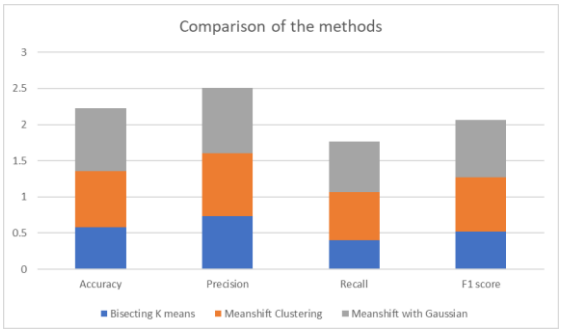

Fig.3Comparison of the GMM based Meanshift with Bisecting K means and normal Meanshift techniques

V. CONCLUSION

Bigdata comes from myriad different sources and one of the important aspects found in sets of bigdata is sentiment analysis. All the information gathered can reveal how people are feeling about a brand, if any potential issues may arise etc. Twitter sentiment analysis systems allow to sort large sets of tweets and detect the polarity of each statement automatically. This work aimed to develop an effective clustering technique to extract emotions from Bigdata of twitter feeds. Here, an improved version of Meanshift clustering algorithm i.e., Gaussian mixture model based Meanshift clustering is used to make the clustering more effective than the existing ones. Based on the experimental analysis, the proposed Gaussian mixture model based Meanshift clustering approach is proven to be useful in performing high quality results in the domain of twitter sentiment analysis. The Proposed approach has been tested on lakh tweets and also made a comparison on the performance of the proposed work with existing meanshift clustering and Bisecting K means algorithm. The result indicates that the proposed Gaussian mixture model based Meanshift clustering methodprovides the accuracy of 87%, that is more clustering accuracy than existing meanshift and Bisecting K means algorithms.

VI. REFERENCES

[1] Damir Demirović, *An Implementation of theMean Shift Algorithm*, Image Processing On Line,9 (2019),pp.251–268.

[2] D. Comaniciu, V. Ramesh, and P. Meer, The variable bandwidth mean shift and data driven scale selection, inEighth IEEE International Conference on Computer Vision (ICCV), vol. 1, IEEE, 2001,pp. 438–445.

[3] F. Meng, H. Liu, Y. Liang, L. Wei, and J. Pei, A bidirectional adaptive bandwidth mean shift strategy for clustering, in IEEE International Conference on Image Processing (ICIP), 2017, pp. 2448–2452.

[4] Mohsen Farhadloo, Erik Rolland," Multi-Class Sentiment Analysis with Clustering and Score Representation", IEEE 13th International Conference on Data Mining Workshops, pp. 904-912, 2013.

[5] C. Wu, L. Shen, and X. Wang, "A new method of using contextual information to infer the semantic orientations of context dependent opinions," in Artificial Intelligence and Computational Intelligence, 2009. AICI'09. InternationalConference on, 2009, vol. 4: IEEE, pp. 274-278.

[6] R. Sharma, S. Nigam, and R. Jain, "Opinion mining of movie reviews at document level," arXiv preprintarXiv:1408.3829, 2014.

[7] R. Sharma, S. Nigam, and R. Jain, "Polarity detection at sentence level," International Journal of Computer Applications, vol. 86, no. 11, 2014.

[8] C. Musto, G. Semeraro, and M. Polignano, "A comparison of lexicon based approaches for sentiment analysis of microblog posts,"Information Filtering and Retrieval, vol. 59, 2014.

[9] X. Hu, J. Tang, H. Gao, and H. Liu, "Unsupervised sentiment analysis with emotional signals," in Proceedings of the 22nd international conference on World Wide Web, 2013: ACM, pp. 607-618.

[10] G. Paltoglou and M. Thelwall, "Twitter, MySpace, Digg: Unsupervised sentiment analysis in Socialmedia,"ACM Transactions on Intelligent Systems and Technology (TIST), vol. 3, no. 4, p. 66, 2012.

[11] P. D. Turney, "Thumbs up or thumbs down?: semantic orientation applied to unsupervised classification of reviews," in Proceedings of the 40th annual meeting on association for computational linguistics, pp. 417–424,
Association for Computational Linguistics, 2002.

[12] B. Georgescu, I. Shimshoni, P. Mee, "MeanShift based clustering in high dimensions: A texture classification example". In ICCV, pp. 456–463, 2003.

ISSN- 2394-5125 VOL 7, ISSUE 15, 2020

[13] C Xiao, M Liu, "Efficient mean-shift clustering using gaussian KD-tree". Comput Graph Forum29 (7), 2065–2073 (2010).

[14] Y. Cheng, "MeanShift, Mode Seeking, and Clustering", IEEE Trans. PAMI, vol. 17, no. 8, pp. 790-799, 1995.

[15] C. Yang, R. Duraiswami, D. DeMenthon, L. Davis, "Mean-shift analysis using quasi-Newton methods", In: Proceedings of the International Conference on Image Processing, vol. 3, pp. 447– 450 (2003)

[16] Bernice Purcell "The emergence of "big data" technology and analytics" Journal of Technology Research 20137. Garlasu, D.; Sandulescu, V.; Halcu, I. ; Neculoiu, G.;,(17-19 Jan. 2013),"A BigData implementation based on Grid Computing", Grid Computing.

[17] Xu, R. and Wunsch II, D. Survey of clustering algorithms, Transactions on Neural Networks, 16(3), (2005), 645-678.

[18] Neethu M S and Rajasree R, "Sentiment Analysis in Twitter using Machine Learning Techniques" 4th ICCCNT 2013 July 4 - 6, 2013, Tiruchengode, India IEEE – 31661.

[19] H Guo, P Guo, H Lu, "A fast MeanShift procedure with new iteration strategy and re-sampling".Proceedings of the IEEE International Conference Systems, Man and Cybernetics SMC '06, 2385– 2389 (2006).

[20] Conover MD, Goncalves B, Ratkiewicz J, A, Menczer F. Predicting the political alignment of twitter users. In: 2011 IEEE third conference on privacy,security, risk and trust and 2011 IEEE third international conference on social computing. 2011.p. 192–9.

[21] https://www.kaggle.com/youben/twitter- sentiment-analysis/data

[22] https://www.linkedin.com/pulse/understanding- tf-idf-first-principle-computation-apache- asimadi/

[23] https://brilliant.org/wiki/gaussian-mixture-model/