ISSN 2319 – 7595
Volume 13, No.2, March – April 2024
International Journal of Information Systems and Computer Sciences
Available Online at http://warse.org/IJISCS/static/pdf/file/ijiscs011322024.pdf
https://doi.org/10.30534/ijiscs/2024/011322024

Unendorsed Machine Approaches of Learning in the field of Analysis of Sentiment over Unstructured Bigdata

Sharon Susan Jacob[1]
[1]Assistant Professor,
School of Data Analytics
Mahatma Gandhi University,
Kottayam, Kerala, India
sharonsusanjacob@gmail.com

Received Date : February 20, 2024 Accepted Date : March 26, 2024 Published Date : April 07, 2024

ABSTRACT

Terabytes of data are created every day by modern information systems and new technology. It takes a lot of work across many levels to get useful information out of these massive datasets for decision-making. Social media, as well as other Internet-based applications, have been a major source of big data in recent years. When it comes to social media, Twitter is a household name across the globe. Unstructured data can be found in abundance on Twitter and other social media platforms. An innovative way to examine the emotions expressed on social media is through the use of sentiment analysis clustering techniques. Unsupervised machine learning methods for sentiment analysis of unstructured bigdata is discussed here. The Meanshift clustering method and the Bisecting K means algorithm is then compared using the metrics of precision, recall, and accuracy, among other things the f1 score. Python programming and Pyspark are employed for data analysis.

Key words: Bigdata, Bisecting K Means, Meanshift, PySpark, Analysis of Sentiment

1. INTRODUCTION

Using the term "Big Data," we are referring to the amount of information that can be mined from a large volume of data. It's a collection of complex data sets that can't be processed by traditional data processing tools. It is possible to divide data into three types: structured, semi-structured, and unstructured. Only 15% of our data is semi-structured, and only 5% of our data is structured; the vast majority of our data falls into the unstructured category. Thanks to social media and its corresponding applications, millions of people can express and spread their views on a subject, as well as show their feelings by liking or disliking content. Big social data is an expression used to explain data that is exalted in volume, high in velocity, exalted in variety, exalted in value, and highly variable as a result of all of these continuous actions on social media. In general, this type of data refers to a large set of opinions that can be processed to identify people's digital tendencies. Big social data has piqued the interest of several researchers who hope to better understand and predict human behaviour across a wide range of contexts by utilising this data source. Text analysis is one method that can be used to help with this type of processing [24]. The text makes up the vast majority of internet data, making text analysis an essential tool for gauging public sentiment and opinion.

Analysis of sentiments expressed in the form of reviews, blogs, news articles and comments and feedback is a form of computational sentiment analysis. This includes sentiments about electronic products and movies, public or private services, individuals, issues, events, topics, and their attributes. As a result, there is a lot of activity in this area right now.

Bisecting K Means and the Meanshift algorithm were used to evaluate sentiment analysis performance. Metrics such as accuracy, precision, recall, and F1 score are used to compare the two approaches. Python programming and PySpark, an Apache Spark and Python combination, were used to analyse the data.

1.1 The Worker's Sense of Purpose

Several organisations store data in a variety of formats, not just tabular data in rows and columns. Data is no longer confined to a structured format as a result of the latest developments in the digital world. There is a significant amount of real-world data generated in the form of unstructured data. Traditional data models are unable to deal with unstructured, heterogeneous, and complex data in the context of big data, and this is their main limitation. Information retrieval is a critical component of big data analysis. Data volumes and formats are making it difficult to use previous methods and solutions for storing and retrieving information. A semantic representation is needed for the proper and efficient use of massive amounts of data.

2. RELATED WORKS

To quote University of Pennsylvania economist and statistician Francis Diebold, Recent and unprecedented advances in data recording and storage technology are largely to blame for the recent and unprecedented growth in the quantity (and sometimes quality) of data that has come to be known as "bigdata." It is no longer useful to measure a sample size in terms of the number of observations, but rather in terms of the number of megabytes. We're not

surprised to see data pile up at rates as high as several terabytes every single day.

The authors [15] discussed the various big data analytics platforms that are available. There is a direct correlation between a task's nature and complexity and the computing platform of choice. The adoption of computing platforms for performing the analysis prompted discussions on issues such as real-time access or archival processing, scalability, and the need for higher computing capabilities in the future. Several other closely related technologies, such as "MapReduce," "Spark," "HIVE," and so on, were also discussed, as were the applications of each. Finally, a theoretical comparison of the advantages and disadvantages of these platforms was presented in their work.

Damir Demirovi [1] provided an explanation of the meanshift algorithm's implementation in great detail. An explanation of the meanshift algorithm's workings is provided, along with the algorithm's analysis and implementation with clear comments.

Comaniciu et al. [2], presents that scale-space can be solved in two ways. Adaptive normalised density gradient estimation is used in the first instance of this problem. In terms of performance, variable bandwidth mean-shift is found to be superior to fixed bandwidth.

To get around the local maximum density, Meng et al. [3] proposes a new strategy that incorporates various adaptive bandwidth strategies and a bidirectional adaptive bandwidth mean shift.

Clustering and score representation were proposed by M. Farhadloo et al. [4] to perform multiclass sentiment analysis for the English language. The model made use of sentiment analysis at the aspect level. A bag of nouns was preferred to a bag of words in order to improve clustering results, score representation, and more accurate sentiment recognition.

Chunxu Wu [5] proposed a framework for synthesising semantic orientations that are not determined by WordNet. The sentiment of people's opinions can be gleaned from this method through the application of semantic closeness measures. Such measures are used when there is insufficient relevant information to guide reviews.

Text documents can be classified according to the polarity of their content using an automated, unsupervised sentiment analysis system, according to Sharma et al. [6]. Classifying negative and positive sentiment words from a collection of documents is done using this system.

Musto et al. [8] proposed an approach that began by breaking the tweet into small-scale phrases, such as those indicated by part signs in the content. An additional micro-phrase was created whenever the part signal appeared in the text and extracted sentiment words from document collections, classifying them according to their polarity. The part signal consisted of punctuation and conjunctions.

When analysing social media data, researchers used emotional signals to determine how people felt. The term "emotional signal" refers to any data that has a strong correlation or association with the polarities of human emotions. X.Hu et al. [9].

The emotional intensity was estimated using lexicon-based methods by Paltoglou and Thelwall [10]. For sentiment polarity classification, as well as for detecting subjective texts expressing an opinion, this method was a good fit.

The relationships between words in a sentence were completely ignored in Turney's [11] methodology to sentiment exploration, which he called the "bag-of-words" approach. The sentiment of each word was calculated individually and then aggregated using aggregation functions to determine the overall sentiment of the sentence.

Georgescu et al. [12] proposed an MS method that uses spatially coherent hash tables and a fast nearest neighbour scan. Feature space elements in close proximity to each other are used to estimate the mean value of the locality-sensitive hashing algorithm.
For example, C Xiao and his colleagues proposed an MS approach that uses a smaller feature space [14]. Adaptive KD-tree in a high-dimensional affinity space is used to create this reduced feature space, which is then used for adaptive clustering of the original dataset.

3. METHODOLOGY

Unendorsed machine learning approaches were tested for their ability to analyse sentiment in unstructured Twitter data. Bisecting K means and Meanshift clustering methods were used to evaluate the performance of sentiment analysis on unstructured Twitter big data. The work clusters Twitter feeds into three categories: positive, negative, and neutral, based on a given set of Bigdata. Unsupervised methods such as Bisecting K Means and Meanshift can be evaluated in accordance with the subsequent conceptual picture, which shows the steps involved in the evaluation of performance over unstructured Twitter big data.

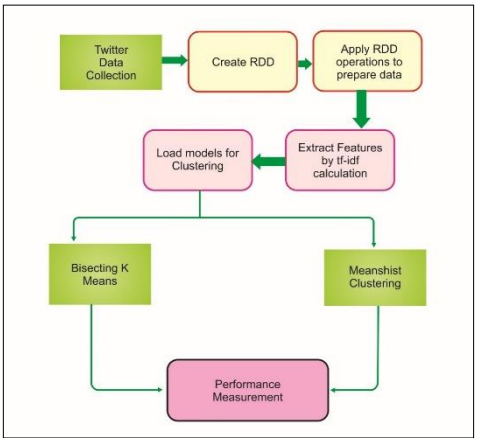

Figure 1: System Block Diagram

3.1 Dataset

This experiment made use of Twitter data that was downloaded from Kaggle. There are 100.000 tweets in the dataset. The dataset is based on the School of AI-Algiers challenge [18].

3.2 Construction of Resilient Distributed Datasets (RDDs)

A Spark programme begins with the creation of a SparkContext object, which tells Spark how to connect to a cluster of machines. First, a SparkConf object must be built that contains information about the application in order to create a SparkContext.

Once the PySpark has been installed and configured, the RDD is created. RDD can be generated in two ways: from already-existing storage or from outside sources. Existing storage uses a driver programme to generate RDDs, which can be parallelized if desired. The RDD is built using a local file system, HDFS, HBase, or any other data source that supports the Hadoop Input Format.

The RDD used in this study is generated from a dataset located on the file system rather than a database.

3.3 Application of Resilient Distributed Datasets (RDD) operations

RDD operations, such as map, flatMap, filter, and collect, were used to prepare the data following the creation of RDD..

- **map()**

Map() was used to lowercase every word in a document that was passed in. When an RDD needs to be transformed by a map transformation, it is useful.

- **flatMap()**

The results of map() are flattened using the flatMap() transformation because the function returns a nested list.

- **filter()**

Because stop words don't add much value to a piece of writing, they must be eliminated. Stopwords are removed using the filter() transformation.

collect()

The action collect() was used to return all of the data from the RDDs to the driver programme. In order for the data to be collected, it must fit in a machine and be transferred to the driver.

Preparing the data is the first step in extracting the features.

3.4. Extracting the Feature

Term Frequency-Inverse Document Frequency was used to extract the features (TF-IDF). It measures the importance of a term in a document in relation to other terms in the same document. Term frequency (TF) and inverse document frequency (IDF) are used to calculate the frequency of a term (IDF). There are many ways of determining how important a word or phrase is in a given document, but TF is one of the most commonly used. Using IDF, you can determine how frequently a term appears throughout all the documents that make up a given corpus. HashingTF was used to implement the TF in this case. Fixed-length features vectors were generated from the sequence of terms (the output of a tokenizer). Feature hashing was used to transform the terms into indices of fixed length. As an input, IDF used the HashingTF feature vector and scaled it according to the frequency of the terms used in each document

3.5. Sentiment Analysis using Bisecting K Means

To categorise a collection of instances, it's one of the most widely-used clustering algorithms (clusters). Another K means variant is the Bisecting K means algorithm [17]. If you're looking for a way to divide a dataset into K sub-clusters rather than K clusters in each iteration, you can use the Bisecting K Means algorithm. However, the clustering of bisecting K means is different from that of regular K means. If you compare bisectional k means to regular k means, you'll see that: K Means Bisecting K Means produce clusters of the same size and are more efficient than regular K means when the size of K is large.

Performing sentiment clustering based on pseudo with the Bisecting K Means algorithm is given in the Table 1 below.

Table 1: Pseudocode of Bisecting K Means algorithm

1. Input dataset
2. Initialize the list of clusters to contain the cluster consisting all points
3. *repeat*
4. Remove a cluster from the list of clusters
5. {Perform several "trial" bisections of the chosen cluster}
6. *for i*=1 to number of trials *do*
7. Bisect the selected cluster using K means
8 *end for*
9. Select the two clusters from the bisection with the lowest SSE
10. Add these two clusters to the list of clusters
11. *until* Until the list of clusters contains K clusters
Output: A set of K clusters

The algorithm uses the K-means class of data to perform the bisecting. When the class first meets, the number of clusters is predetermined to be three. Until the desired number of clusters is achieved, it is repeated. By starting with the highest SSE cluster and then dividing it into two new clusters, the algorithm selects two clusters with the lowest combined SSE at the end of each cycle.

The algorithm begins training with a single cluster of all points. K-means is used to bisect each cluster until there are k total clusters or none of them are divisible, whichever comes first. It is more efficient to group together the bisecting steps of clusters on the same level. When all divisible clusters on the bottom level generate more than k leaf clusters, the larger clusters are prioritised. Using a trained model, an algorithm compares a point to the root cluster node's cluster centres. It is then compared to the child cluster centres of the root's nearest offspring to see if any differences exist. The algorithm repeats itself until it reaches a node in the cluster of leaf nodes. In the end, the point is awarded to the leaf cluster that is closest.

3.6. Sentiment Analysis using basic Meanshift clustering algorithm

Classifying a dataset's cluster structure by looking for areas with high data densities is an obvious choice. A kernel density estimation can be used to find clusters using meanshift algorithms. Non-parametric mode finding can be accomplished using the mean shift algorithm. Hill-climbing algorithm on the density defined by a finite mixture or an estimate of the kernel density [16] can be used to describe it. Mean shift is a non-parametric clustering method based on ideas proposed by Fukunaga and Hostetler [14].

Things needed before starting to run Meanshift on a set of data points *X*:

The neighbours of a point x are found using a function N(x). Points within a certain distance of each other are considered to be neighbouring. In most cases, the distance metric is used. Euclidean Distance

- The neighbours of a point x ∈ are X are found using a function N(x). Points within a certain distance of each other are considered to be neighbouring. In most cases, the distance metric is used. Euclidean Distance
- A kernel *K (d)* is made use in Meanshift where d is the distance between two datapoints. In this existing methodology A Flat kernel was used.

Performing sentiment clustering based on the pseudo code with the basic Meanshift clustering algorithm is given in Table 2.

Table 2: Pseudocode of basic Meanshift Clustering algorithm

1. Input set of data points *X*
2. Start with the data points assigned to a cluster of their own
3. For ***each datapoint*** *x ∈ X, find the neighbouring points N(x) of x.*
4. For each datapoint x ∈ X, calculate the mean shift m(x) from this equation:

$$m(x) = \frac{\sum_{x_i \in N(x)} K(x_i - x) x_i}{\sum_{x_i \in N(x)} K(x_i - x)} \quad (1)$$

5. For each datapoint x ∈ X, update x ← m(x).
6. *The process will be iterated and moved to the higher density region.*
7. Repeat from 3 for n_iteations or until the points are almost not moving or not moving.
8. Output: Three clusters of data points

The algorithm iteratively assigns each data point to a cluster centroid, using the majority of nearby points as a compass. Each data point will get closer to the cluster centre with each iteration because that is where the majority of points are located. Upon completion of the algorithm, a cluster is formed, with each point being assigned a new location within it.

4. EXPERIMENTAL ANALYSIS

As each data point is assigned to a cluster centroid using the majority of nearby points as a compass, the algorithm iteratively assigns each data point Iteration after iteration will bring each data point closer to the cluster centre, which is where the vast majority are. A cluster is formed once the algorithm has been completed and each point has been given a new address.

4.1 Evaluation Measures

For sentiment analysis of unstructured big data, accuracy, precision, recall, and f1-score are some of the evaluation measures used.

4.2 Comparative Analysis & Results

Unsupervised machine learning methods on unstructured big data were evaluated using two existing clustering methods, which were compared.

The effectiveness of unsupervised learning methods was evaluated using f1-score and other external validation metric measures. There was a 57% difference in accuracy between the Bisecting K Means and the basic Meanshift clustering methods.

Meanshift clustering precision is (0.73) for Bisecting K Means, while the basic Meanshift clustering precision is (.86). In the same way, the recall measures obtained by these methods are 40 and 65 per cent, respectively. For the basic Meanshift and Bisecting K Means clustering methods, the f1-score values are (.52) and (.74). Based on evaluation metrics, the following table shows a side-by-side comparison.

Table 3: Evaluation Metric Scores of the Methods

Clustering 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

The basic Meanshift clustering method has greater accuracy (77 percent), precision (.86), recall (.65), and f1 score (.74) than the Bisecting K Means method.Graphs showing the measurement results are shown in the following figures 2 and 3:

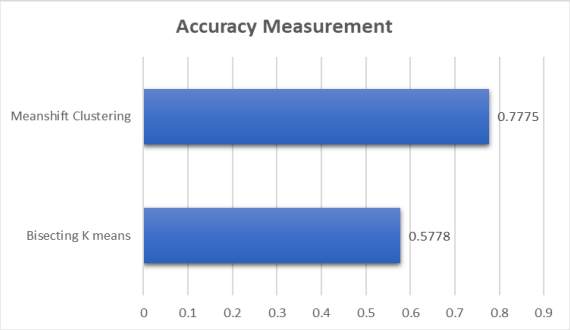

Figure 2: The plot representing accuracy values of the unsupervised methods on twitter bigdata

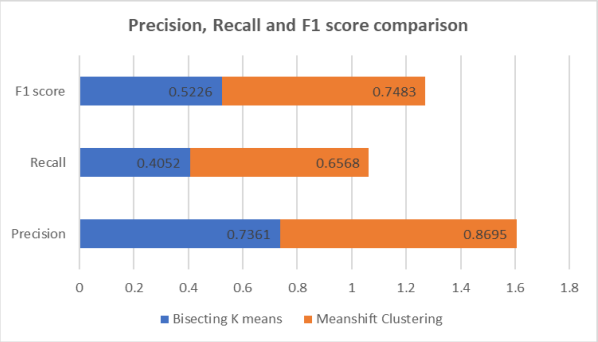

Figure 3: The precision, recall and f1 score plot of unsupervised methods on twitter bigdata

5. CONCLUSION AND FUTURE WORK

Unsupervised machine learning methods were used to test sentiment analysis on Twitter's massive data set. Data analysis was carried out using Python and PySpark. The performance of sentiment analysis was assessed using bisection of K means and the Meanshift algorithm. For example, these methods were compared based on metrics such as precision, recall and the F1 score. The basic Meanshift clustering method outperformed the Bisecting K means method by a margin of 77 percent. The accuracy, recall, and f1 score of this study were all excellent (all above 0.86). Meanshift clustering's efficiency can be improved in the future by adjusting the bandwidth parameter. Other big data tools can help speed up data storage and processing.

REFERENCES

[1] Damir Demirović, An Implementation of the Mean Shift Algorithm, Image Processing OnLine, 9, pp. 251–268, 2019

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

[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), pp. 2448–2452, 2017.

[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, AICI'09. International Conference on, 2009, vol. 4: IEEE, pp. 274-278, 2009.

[6] R. Sharma, S. Nigam, and R. Jain, "Opinion mining of movie reviews at document level," arXiv preprint arXiv: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 social media," 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.

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

[14] K. Fukunaga and L. Hostetler, The estimation of the gradient of a density function, with applications in pattern recognition, IEEE Transactions on Information Theory, vol.21, pp.32–40, 1975.

[15] Singh, D., & Reddy, C. K.,"A survey on platforms for big data analytics", Journal of Big Data, vol.2(1), Article No.8, 2015.

[16] M.A. Carreira-Perpinan´, "A review of mean-shift algorithms for clustering", arXiv preprint arXiv:1503.00687, 2015.

[17] Liu, G., Huang, T., & Chen, H., "Improved bisecting K-Means clustering algorithm", Computer Applications and Software, 2, 2015.

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