Sentiment Clustering of Movie Review Data UsingUnsupervised Machine Learning Algorithm

[1]Abinash Tripathy, [2]Panchananda Jha, [3]Rugada Vaikunta Rao, [4]Ch.Srinivasu

[1]Department of Computer Science and Engineering, Raghu Engineering College, Visakhapatnam
[2,3]Department of Mechanical Engineering, Raghu Engineering College, Visakhapatnam
[4]Department of Electronics and Communication Engineering, Raghu Engineering College, Visakhapatnam

Abstract:

It is observed that consumer often share their opinion, views or feeling about any term used on social networkthrough reviews, comments. These reviews are often writtenin natural language and mostly unstructured. Thus, to obtainany meaningful information from these reviews, it needs to beprocessed. The reviews obtained are mostly not tagged as whetherit is of positive or negative in nature. so, in order to processit unsupervised machine learning techniques are implemented.Clustering method are applied to analyse the reviews by makingcluster of reviews. In this paper, four different unsupervisedclustering techniques i.e., K-Means, mini batch K-Means, AffinityPropagation and DBSCAN are applied to analyse the moviereviews. Different performance evaluation parameters are usedto evaluate the performance of these techniques.

Keywords: Unsupervised Machine Learning, Sentiment Analysis, Clustering Techniques, Performance evaluation parameters

1. Introduction

Sentiment analysis aims at analysing the reviews, comments and opinions on a particular topic, event or product.Sentiment clustering helps to partition the review data into different sets that are meaningful and relevant for the purpose.During the process of clustering, the natural language informationis considered and based on that the clustering of datais performed [1].

Machine leaning algorithms are very often helpful tocluster and predict whether a document represents positive ornegative sentiment. Those algorithms are categorized as twotypes known as supervised and unsupervised machine learningalgorithms. Supervised algorithm uses a labelled dataset where each document of training set is labelled with appropriate sentiment. Whereas unsupervised learning algorithms include unlabelled data set where text is not labelled with appropriate sentiments[2]. This study is concerned with unsupervised learning techniques on a case study of unlabelled movie reviewdata.The movie reviews are written in natural language whichare mostly unstructured. This unstructured data needs to be converted to meaningful data in order to apply machine learning algorithms.

In this study, an attempt has been made to transform thetextual movie reviews to a numerical matrix where each columnrepresents the identified features and each row representsa particular review. The matrix is given as input to machine learning algorithm in order to train the model. This model isthen tested and different performance parameters are studied. The results obtained are critically examined on the basis of comparison with existing literature.

The following paper is organized as follows: Section 2 provides a brief idea about the present literatures; Section 3suggests the detailed methodology adopted by the proposedalgorithms; Section 4 explains the proposed approach withresult; Section 5 gives a comparison of obtained results with other literatures and finally section 6 concludes the paper alongwith scope for future work.

2. Related Work

Pang et al., has proposed classification of document based on sentiment analysis of on-line movie review data using three machine learning methods such as maximum entropyclassification, Naive Bayes and support vector machine [3].

Jain et al implemented different method of data clusteringsuch as Hierarchical Clustering Algorithms, Partitioned Algorithms, Mixture-Resolving and Mode-Seeking Algorithms, Nearest Neighbour Clustering, Fuzzy Clustering also artificialneural network [4].

Li and Liu proposed a clustering-based approach for sentimentanalysis of text document by TF-IDF weighting scheme, importing score of term and voting mechanism [1]. Also, evaluation of three different methods, Symbolic technique, supervised learning and clustering-based approach has been done.

Ma et al., used different clustering method for online review sentiment analysis has done for comparison of different clustering algorithm with different weighting schemes on six different data set and result obtained in terms of accuracy [5].

Scully et al., proposed a modified method of K-meansknown as Mini-Batch K-Means for clustering purpose. Here inthis algorithm computational time gets decreased but qualityof result gets deteriorated [6].

Guan et al., proposed a new clustering algorithm forclustering of text document that is seed affinity propagation (SAP). It reduces the computing complexity of text clustering and improves the accuracy. Also, a new similarity measurement is proposed, which is extension of cosine coefficient, capturing structural information of text [7].

Yang and Ng proposed a new scalable distance based clustering (SDC) algorithm, which is found out to be betterthan DBSCAN. It forms a smaller number of relevant clusters, basedon density-reachability criteria. Also, SDC and DBSCAN are evaluated based on micro-accuracy and macro accuracy [8].

3. Methodology Used

Sentiment clustering is a process of grouping of the datainto different clusters. The number of clusters created onany dataset varies depending upon the requirement. In this paper, two different clusters are considered, i.e., one for the positive and other for negative cluster. Again, the reviews are in natural language; hence they need to be processed properly. As machine learning techniques mostly are applied on numerical data, these movie reviews are need to converted into numerical vectors for machine learning processing.

The vectorization of textual data to numerical vector is done using following methodologies.

- **Count Vectorizer (CV):** This process of vectorization mainly depends upon the occurrence of any feature or words. It does not depend upon the number of times a feature occurs in the text. Thus, it generates a sparsematrix where the occurrence of any feature representsby ‘1’ and non-occurrence by ’0’[9]. The concept ofCV can be explained using following example:

  Calculation of CountVectorizer Matrix: suppose wehave three different documents containing following sentences.

  “Book is interesting”.
  “Book is Awful”.
  “Book is good”.

  Matrix generated of size 3*5 because we have 3documents and 5 distinct features. Thematrix willlook like given in Table 1.

  Table 1: Matrix generated under CountVectorizer Scheme

	Feature 1	Feature 2	Feature 3	Feature 4	Feature 5
Sentence 1	1	1	1	0	0
Sentence 2	1	1	0	1	0
Sentence 3	1	1	0	0	1

  Each 1 in a row corresponds to presence of a feature and 0 represents absence of a feature from particular document.

- **Term Frequency - Inverse Document frequency (tf -idf):** Unlike the CV, where the frequency if the features are not considered, tf -idf concerned aboutthe frequency of a word not only in particular reviewbut also in the total review set. This score helps inbalancing the weight between most frequent or generalwords and less commonly used words. Term frequency calculates the frequency of each token in the review;but this frequency is offset by frequency of that tokenin the whole corpus[9]. tf - idf value shows the importance of a token to a document in the corpus.

The process of tf - idf can be explained usingfollowing example:

For example: a movie review contains 100 words,wherein the word Awesome appears 10 times. The termfrequency (i.e., tf) for Awesome then (10 / 100) =0.1. Again, suppose there are 1 million reviews in thecorpus and the word Awesome appears 1000 times inwhole corpus. Then, the inverse document frequency(i.e., idf) is calculated as log(1,000,000 / 1,000) = 3.
Thus, the tf - idf value is calculated as: 0.1 * 3 =0.3.

After the dataset in converted into a matrix on numbers, then it is given input to the machine learning algorithms for clustering. The different machine learning algorithms used in this paper are explained as follows:

1. **K-Means:** This algorithm is simple and fast for computation of clustering. In this algorithm initial cluster centre are assigned randomly which have a great impact on result formed[10]. The process of k-means clustering can be explained as follows:

   - A dataset $D = \{d_1,\ d_2, \dots\ ,\ d_n\}$is consisting of ‘indifferent data point or features.
   - In k-means, the number of clusters are definedbefore the processing starts. Here in this casetwo clusters are defined i.e., positive and negativecluster.
   - The squared Euclidean distances between the featuresand the centroid (cluster centre) are foundout. This value is known as clustering error andvaries upon the centre of cluster.
   - This error can be found out using following equation:

$$E\ (C_1, C_2, \dots, C_m) = \sum_{i=1}^{N}\sum_{k=1}^{M} I(d_i \in C_k)\ ||d_i - C_k)|| \qquad (1)$$

   Where, $E(C_1, C_2, \dots, C_m)$ is the error found outfor different cluster, $I(d_i) = 0$ if D is positive and 0 if D is negative.$||d_i - C_k)||$ finds out thedistance between the features and the centre.

   - Depending up on the distance of the data point form the centroid, the centroid is changed until the optimum result obtained where the data points make a cluster near centroid.

2. **Mini Batch K-means:** Mini-Batch K-Means is modified form of K-Means Method. Its uses smaller subset to decrease the processing time and trying to increase optimize solution[6]. Each subset is randomly created in every iteration. To find the Local solution of problem, mini batch reduces the computation. But the result obtained is observed it is not better than the standard algorithm. The algorithm has basically two steps. In first step, from the dataset, different samples are selected randomly to create mini-Batch. Those mini-Batch created are allocated to nearest centroid. In next step centroid gets updated. For each sample the above step is repeated. For eachsubset of data in mini-Batch, centroid get updated byaverage of sample data and all previous sampled datain that particular centroid. This helps in decreasing therate of change of centroid over time. All those steps arerepeated till fixed number of iterations are reached.

   The mini batch k-means is an optimization problem to findout the set of clusters C, to minimize over a set of dataX with an objective function as follows:

$$\min \sum_{x \in X} ||f(C, x)||^2 \qquad (2)$$

   Where f(C, x) returns the nearest cluster centre to x using Euclidean distance.

3. **Affinity propagation:** This algorithm finds the similarity between pair of input data point. Several messages are exchanged between data points until the best set of exemplars comes out. Here exemplar refers to representativeof each cluster [11]. The approach adopted by the method can be explained as follows:

   The dataset $D = \{d_1, d_2, \dots, d_n\}$ is the 'n' different data elements or features. 'S' be the function that represents the similarity between two data points, where, $S(x_i, x_j) > (x_i, x_k)$iff x_i is more similar to x_j than x_k. The algorithm moves forward with updating the message passing steps, thus creating two different matrices i.e., "Responsibility matrix" and "Availability matrix". All these matrices are initially set to zero and then updated as the processcontinues. The responsibility matrix R has values r(i, k)that quantifies as to how serves as the exemplar for x_k, relative to other candidate exemplars for x_j. The matrix can be updated as follows:

$$r\,(i,k)\ \leftarrow s\,(i,k) - \max\{\,a(i,k^{`}) + s(i,k^{`})\} \qquad (3)$$

   The "availability" matrix A contains values a(i, k) that represents as to how "appropriate" it would be for x_i to pick x_k as its exemplar, taking into account of other points' preference for x_k as an exemplar. The matrix canbe updated as follows:

$$a\,(i,k)\ \leftarrow \min(0, r(k,k)) + \sum_{i^{`} \notin \{i,k\}} Max\ \left(0,\ \ r(i^{`},k)\right)\ \ if\ i \neq k \qquad (4)$$

$$a\,(k,k)\ \leftarrow \sum_{i \neq k} Max\ \left(0, r(i^{`},k)\right) \qquad (5)$$

4. **DBSCAN:** Clustering of data in DBSCAN algorithm is formed based on density of data. Clusters are separated between high density and low density [12]. The cluster formed can be in any shape due to this mechanism. Where, as in K Means clustering algorithm, cluster found is assumed mostly to be in convex shaped. Area which has high density is considered to be main component of this algorithm, also called core samples. The clusters formed are set of core samples and non-core samples. Where core samples are near to each other and non-core samples areclose to core sample, but do not belong to core samples.There are two parameters, those are minmunsampleandeps. Higher value of minimum samples or lower value of eps indicates high density necessary to form cluster.

In order to validate the result obtained by the system, it is compared by some performance evaluation parameters. The different performance evaluation parameters used in this paper to evaluate the performance of the clustering algorithms are described as follows:

- **Homogeneity:** The data point that belongs to single class must be assigned to single cluster in order to satisfy homogeneity criteria [13], which means it must have zero entropy. In other words, inside a singlecluster only one class has to be there. Homogeneity can be calculated as:

$$H = \begin{cases} 1\ if\ H(C,k) = 0 \\ 1 - \dfrac{H(C|H)}{H(C)}\ \ else \end{cases} \qquad (6)$$

  where,

$$H(C|K) = -\sum_{K=1}^{|K|}\sum_{C=1}^{|C|} \frac{a_{ck}}{N} \log \frac{a_{ck}}{\sum_{C=1}^{|C|} a_{ck}} \qquad (7)$$

$$H(C) = -\sum_{c=1}^{|c|} \frac{\sum_{k=1}^{|k|} a_{ck}}{N} \log \frac{\sum_{k=1}^{|k|} a_{ck}}{N} \qquad (8)$$

- **Completeness:** From all given classes, all data points must be member of same cluster in order to satisfy the criteria of completeness. If the result is perfectly complete, it means that all data points from differentclasses are skewed into single cluster mentioned in [13]. Completeness can be calculated as:

$$H = \begin{cases} 1 & if\ H(K,C) = 0 \\ 1 - \dfrac{H(K|C)}{H(K)} & else \end{cases} \quad (9)$$

where,

$$H(K|C) = -\sum_{C=1}^{|C|}\sum_{K=1}^{|K|} \frac{a_{ck}}{N} \log \frac{a_{ck}}{\sum_{K=1}^{|K|} a_{ck}} \quad (10)$$

$$H(K) = -\sum_{K=1}^{|K|} \frac{\sum_{C=1}^{|C|} a_{ck}}{N} \log \frac{\sum_{K=1}^{|C|} a_{ck}}{N} \quad (11)$$

- **V measure:** V-Measure is the weighted harmonic mean of homogeneity and completeness. It evaluates how successfully criteria of completeness and homogeneity are fulfilled, described in [13]. It's anentropy-based measurement. It is calculated by

$$V_\beta = \frac{(1 + \beta)\, x\, h\, x\, c}{(\beta\, x\, h) + c} \quad (12)$$

where h indicates homogeneity and c indicates completeness

- **Adjusted Rand Index:** Rand index in clustering is measurement of similarity of data cluster[14]. Adjusted Rand Index is another form of Rand index. In rand index the value obtained lies between 0 and 1 but in case of adjusted rand index values can be negative in case when index value is less than expected index.From mathematical point of view, it is similar to accuracy, but it is only applicable when there is noclass label on data.

Given a set S of v elements, and two cluster of these points, namely $x_1, x_2, \dots, x_n$ and $y_1, y_2, \dots, y_r$ theoverlapping of X and Y betweencan be summarized in a contingency Table 2. Where each entry v_{ij} denotes the number of objects in common between x_i and y_j.

$$ARI = \frac{i - e_i}{m_i - e_i} \quad (13)$$

Table 2: Contingency Table

	Y_1	Y_2	Y_3	…	Y_r	Sum
X_1	v_{11}	v_{12}	v_{13}	…	v_{1r}	p_1
X_2	v_{21}	v_{22}	v_{23}	…	v_{2r}	p_2
X_3	v_{31}	v_{32}	v_{33}	…	v_{3r}	p_3
…	…	…	…	…	…	…
X_n	v_{n1}	v_{n2}	v_{n3}	…	v_{nr}	p_n
Sum	q_1	q_2	q_3	q_4	q_5	

$$ARI = \frac{\sum_{ij}(v_{ij}) - \frac{\sum_i(p_i)\sum_i(q_i)}{v}}{\frac{1}{2}[\sum_i(p_i) + \sum_j(q_j)] - \frac{\sum_i(p_i)\sum_i(q_i)}{v}} \quad (14)$$

Where, i represent index, e_i indicate expected index and m_i indicates maximum index.

- **Silhouette Coefficient:** It represents the comparison of tightness and separation of cluster[15]. It shows which data point lies inside the cluster and which data points lies somewhere in between clusters. Mathematically silhouette coefficient can be defined as

$$s(i) = \frac{b_i - a_i}{\max(a_i\ ,\ b_i)} \quad (15)$$

Or

$$s(i) = \begin{cases} 1 - \frac{a_i}{b_i} & if\ a_i > b_i \\ 0 & if\ \ a_i = b_i \\ \frac{b_i}{a_i} - 1 & if\ b_i > a_i \end{cases} \quad (16)$$

Where, i indicates each data point, ai indicates the average dissimilarity of data within a cluster and bi indicates the lowest average dissimilarity of other cluster where i does not belong to. So, $-1 \leq s(i)\ \leq 1$.

In this paper, Internet Movie Database (IMDb) is considered for sentiment analysis. It consists 12500 positive labelled test reviews, 12500 positive labelled train reviews. Similarly, 12500 negative labelled test reviews, 12500 positive labelled trainreviews. Apart from labelled supervised data, an unsupervised data set also contains 50000 reviews.

4. Proposed Approach

The stepwise elaboration of the approach is described as follows:

1. The reviews in dataset obtained are written in natural language which contains absurd information that needs to be removed before the process of clustering started. The unwanted information are as follows:

    - **Stop words:** These words have no effect to the calculation of sentiment values thus they mustbe removed. The words are like “I, it, this”.
    - **Special character and numeric values:** The specialcharacters like “%,$,” and numeric valuesmust be removed as they have no role to playwith the sentiment value evaluation.

2. After the unwanted information removal, the next step is to convert the text reviews into numerical vector. Different methods used for conversion of text data into numerical vectors are CV and tf -idf. In this paper, the tf - idf is used for conversion of text data into numerical data.

3. After the text data is converted into numerical vectors, they are given input to the unsupervised machine learning algorithms to obtain the clustering of reviews. The algorithms can be described as follows:

    - **K-Means:** This algorithm is simple and fast for computation of clustering. In this algorithm, initial cluster centres are assigned randomly which have a great impact on result formed. The distance of data points is calculated form the centre and based on it the clustering is done.
    - **Mini batch K-Means:** Its uses smaller subsetto decrease the processing time and tries toincrease optimize solution. In each step a randomsubset of total data is considered andwith change in result the centre changes to getoptimum value.

- **Affinity propagation:** This algorithm finds thesimilarity between pair of input data point.Several messages are exchanged between datapoints until the best set of exemplars comesout. Here exemplar refers to representative ofeach cluster.

- **DBSCAN:** Clustering of data in DBSCAN algorithm is formed based on density of data. Clusters are separated between high density and low density.

4. After the different machine learning algorithms are implemented, they are evaluated using different performanceevaluation parameters. The result obtainedare shown in following Table 3 as below.

Table 3: Performance Evaluation after Clustering

	Algorithms Used			
	K- Means	Mini K-means	Affinity Propagation	DBSCAN
Homogeneity	0.745	0.626	0.912	0.953
Completeness	0.764	0.675	0.854	0.883
v-measure	0.754	0.65	0.882	0.917
ARI	0.834	0.704	0.85	0.95
Silhouette	0.007	0.006	0.111	0.004

It can be observed from the table III, that the DBSCAN method shows the best result as compared to other three methods. It can also be found out that the values of homogeneity, completeness, v-measure and ARI are close to 1, whereas the value of Silhouette coefficient is close the zero i.e., the parameters other than Silhouette coefficient must be higher to shows the better accuracy and the silhouette coefficient valuemust be low enough which shows the error rate.

The DBSCAN method shows a better result in comparedto other methods because in this method, theanalysis is mainly based on the density or distributionof the data element. On the other hand, in the case ofk-means and mini batch k-means the analysis is based on the distance of the data points from the centroidwhich is ever changing until the optimum result isobtained. Thus, in these cases the result found out to be less accurate. Even in case of Affinity Propagation, where message transmission between the data points carried out and the comparison between them indicates the centre and associated cluster. Thus, the DBSCANmethod shows better result in comparison with othermethods as it works on distribution of the data pointsthat helps to ultimate cluster making.

The following Figure 1 shows the output after the clustering algorithms run on IMDb dataset.

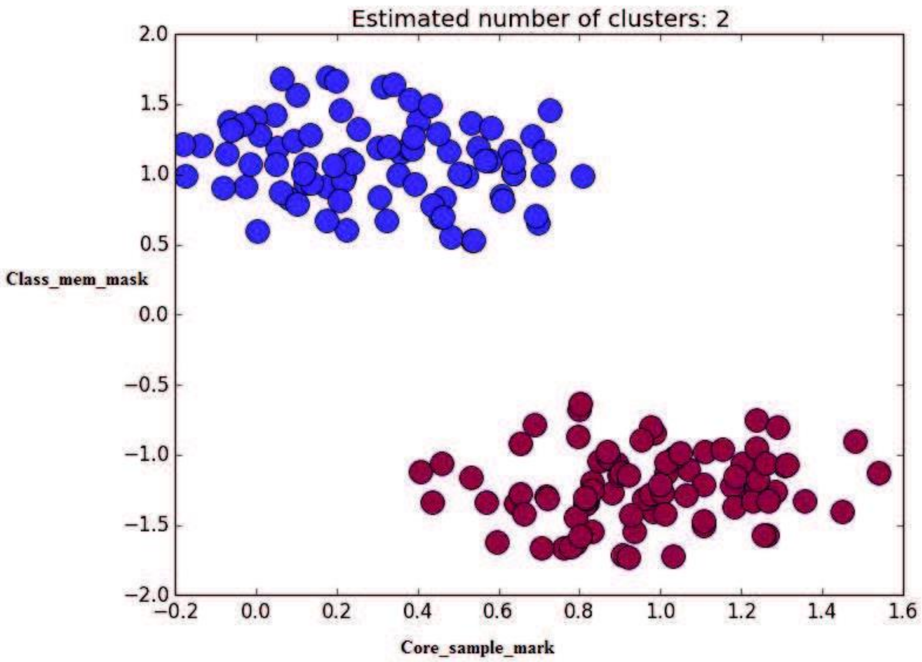

Figure 1: Clustered formed after analysis

It can be seen in figure 1 that, two different clusters are created, i.e., represented with blue colour and other with red colour. The X-axis represents the Core sample mask i.e., the elements of the samples are represented on the X-axis on the other hand the y-axis represents the Class mem mask thatrepresent which element belongs to positive cluster by placingthe data points above the ‘0’ point and the negative data pointsare placed below the ‘0’ point. Again, if the graph checkedproperly, in the y axis, clusters are divided into two groups onewhich is below the zero value and other above the zero value.The data elements that are present before the zero value in Yaxis represent the cluster for positive elements on the otherhand the data elements below the zero value to represent thecluster of negative data elements. The figure 1 is a consolidatefigure for the all the proposed algorithms.

The following figure 2 shows a comparative analysis ofthe performance evaluation parameter of proposed algorithms.This figure shows a graphical comparison of the valuesobtained after the clustering process i.e., the performanceevaluation parameters. It can be observed that the homogeneityvalue varies in a range of 0.75 to 0.95, the completeness valuevaries in between 0.76 to 0.88, the v-measure varies in between0.75 to 0.91, ARI otherwise known to be accuracy varies from83 to 95 %, Finally silhouette value varies from 0.007 to 0.004.

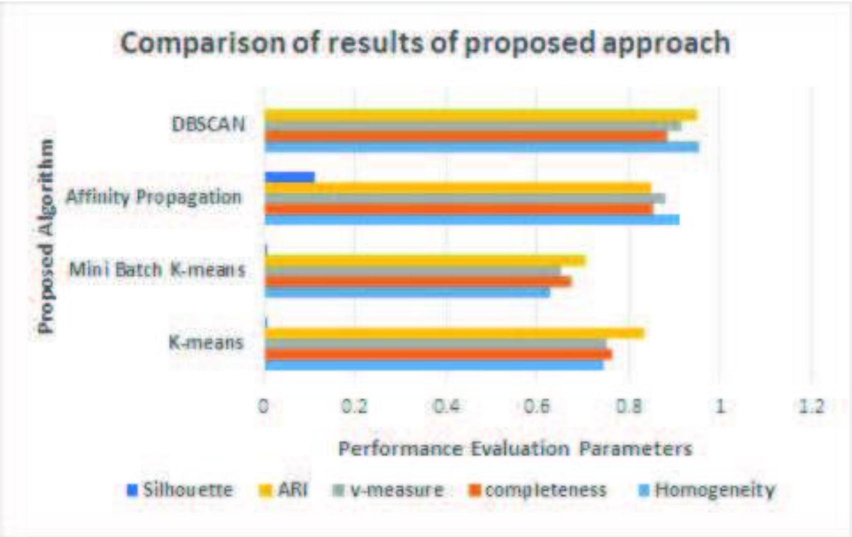

Figure 2: Comparison of performance of machine learning techniques

5. Comparative Study

The following Table 4 compares the proposed result with the existing results.

Table 4: Comparative analysis of the obtained result

	Authors					
Methods	Li and Liu	Balbantaray et.al	Chaturvedi et.al.	Sureka and Punitha	Scully et.al.	Proposed Approach
K- Means	77.17 – 78.33	66.67				83.40
Mini K-means					65.38	70.4
Affinity Propagation			75.06			85
DBSCAN				91.66		**95.2**

In the above Table 4, it can be viewed that Li and Liu[1] have obtained an accuracy for K-Means algorithm in between 71.77%-78.33% whereas Balabantaray et al. [16] have found out an accuracy of 66.67%. But in proposed approach the accuracy achieved is 83.44%. In case of Mini-Batch K-means, Sculley et al.[6] have achieved an accuracy of 65.38%, But in proposed approach the accuracy achieved is 70.04%. In case of Affinity propagation, Chaturvedi et al., [17], have obtained an accuracy of 75.06% whereas the proposed method shows a result of 85%. In case of DBSCAN the accuracy obtained by Sureka and Punitha is 91.66 % but as per the proposed approach the accuracy is 95.20 %.

6. Conclusion

In this paper, four different algorithms are implemented for clustering of text document. From the obtained result, DBSCAN is the best suited for clustering of text document. Also, Mini-Batch K-Means algorithm has the less execution time than K-Means algorithm but accuracy of Mini-Batch K-Means gets reduced.

In future these algorithms can be implemented using different weighting schemes such as BM25, DPH DFR and H LM for increasing the accuracy of result. Also, several different clustering algorithms may be implemented to achieve better results.

References

[1] G. Li and F. Liu, "Sentiment analysis based on clustering: a framework in improving accuracy and recognizing neutral opinions," *Applied Intelligence,* vol. 40, no. 3, pp. 441-452, 2014.

[2] Y. Singh, P. K. Bhatia and O. Sangwan, "A review of studies on machine learning techniques," *International Journal of Computer Science and security,* vol. 1, no. 1, pp. 70-84, 2007.

[3] B. Pang, L. Lee and S. Vaithyanathan, "Thumbs up?: sentiment classification using machine learning techniques," in *ACL-02 conference on Empirical methods in natural language*, 2002.

[4] A. K. Jain, M. N. Murty and P. J. Flynn, "Data clustering: a review," *ACM computing surveys (CSUR),* vol. 31, no. 3, pp. 264-323, 1999.

[5] B. Ma, H. Yuan and Q. Wei, "A comparison study of clustering models for online review sentiment analysis," *Web-Age Information Management, Springer,* pp. 332-337, 2013.

[6] D. Sculley, "Web-scale k-means clustering," in *Proceedings of the 19th international conference on World wide web*, 2010.

[7] R. Guan, X. Shi, M. Marchese, C. Yang and Y. Liang, "Text clustering with seeds affinity propagation,," *IEEE Transactions on Knowledge and Data Engineering,* vol. 23, no. 4, pp. 627-637, 2011.

[8] C. C. Yang and T. D. Ng, "Analyzing and visualizing web opinion development and social interactions with density-based clustering," *IEEE Transactions on Systems, Man and Cybernetics,* vol. 41, no. 6, pp. 1144-1155, 2011.

[9] A. Tripathy, A. Agrawal and S. K. Rath, "Classification of sentimental reviews using machine learning techniques," in *Procedia Computer Science*, 2015.

[10] A. Likas, N. Vlassis and J. J. Verbeek, "The global k-means clustering algorithm," *Pattern recognition,* vol. 36, no. 2, pp. 451-461, 2003.

[11] B. J. Frey and D. Dueck, "Clustering by passing messages between data points," *Science,* vol. 315, pp. 972-976, 2007.

[12] M. Ester, H.-P. Kriegel, J. Sander and X. Xu, "A density-based algorithm for discovering clusters in large spatial databases with noise," *KDD,* vol. 96, pp. 226-231, 1996.

[13] A. Rosenberg and J. Hirschberg, "V-measure: A conditional entropy based external cluster evaluation measure," in *EMNLP-CoNLL*, 2007.

[14] W. M. Rand, "Objective criteria for the evaluation of clustering method," *Journal of the American Statistical association,* vol. 66, pp. 846-850, 1971.

[15] P. J. Rousseeuw, "Silhouettes: a graphical aid to the interpretation and validation of cluster analysis," *Journal of computational and applied mathematics,* vol. 20, pp. 53-65, 1987.

[16] R. C. Balabantaray, C. Sarma and M. Jha, "Document clustering using k-means and k-medoids," *arXiv preprint arXiv:1502.07938,* 2015.

[17] A. Chaturvedi, K. Barse and R. Mishra, "Affinity propagation based document clustering using suffix tree," *International Journal of Engineering Research and Technology,* vol. 3, no. 1, 2014.