Mehran University Research Journal of Engineering and Technology
Vol. 40, No. 3, 630 - 644, July 2021
p-ISSN: 0254-7821, e-ISSN: 2413-7219
DOI: https://doi.org/10.22581/muet1982.2103.16

Sentiment Analysis based on Soft Clustering through Dimensionality Reduction Technique

Saba Akmal[1a], Hafiz Muhammad Shahzad Asif[1b]

RECEIVED ON 03.02.2019, ACCEPTED ON 16.12.2020

ABSTRACT

Clustering based sentiment analysis confers new directions to analyze real-world opinions without human participation and pre-tagged training data overhead. Clustering based techniques do not rely on linguistic information and more convenient as compared to other traditional machine learning techniques. Combining the dimensionality reduction techniques with clustering algorithms highly influence the computational cost and improve the performance of sentiment analysis. In this research, we applied Principal Component Analysis technique to reduce the size of features set. This reduced feature set improves binary K-means clustering results of sentiments analysis. In our experiments, we demonstrate the performance of the clustering system with a reduced feature set to provide high-quality sentiment analysis. However, K-mean clustering has its own limitations such as hard assignment and instability of results. To overcome the limitation of traditional K-means algorithm we applied soft clustering (Expectation maximization algorithm) approach which stabilizes clustering accuracy. This approach allows a soft assignment to cluster documents. Consequently, our experimental accuracy is 95% with standard deviation rate of 0.1% which is sufficient to apply the clustering technique in real-world applications.

Keywords: **Sentiment Analysis, Dimensionality Reduction, Principal Component Analysis, Gaussian Mixture Model, Expectation Maximization, Supervised Learning, Symbolic Technique, Unsupervised Learning**

1. INTRODUCTION

In this advanced era of technology, due to the wide use of World Wide Web, a lot of data is available over the internet. Volume of online sources is increasing beyond storage capacity [1]. Most of the data is available in the form of text, video, and images. This textual information is categorized into two types, facts, and opinions. Facts are the objective point of view of an event and property. Opinions are the subjective point of view that people describe their feelings and expressions. Researchers are actively investigating automatic techniques for text processing to extract useful information and improve e-commerce, science, society and national security.

From last few decades, discussion forums, blogs, news articles and social media websites (face book, twitter, blogs, Linked-In *etc.*) are growing rapidly. These sites are the main source of textual information in form of discussions, comments, and opinions over the internet. The world has been transformed into digital era due to the explosion of World Wide Web. A huge number of users are generating dynamic content on these social sites. Most of the research is already conducted on these social media sites such as twitter reviews [2] face book posts [3] and comment analyzer [4]. People are expressing their reviews about different articles, TV serials, movies and products. A different group of people has a different opinion about products, goods and scientific articles [5]. According to survey [6] and [7] most consumers conduct the online search at least

[1] Department of Computer Science, University of Engineering and Technology Lahore, 54890, Pakistan.
Email: [a]sabaakmal23@yahoo.com (Corresponding Author), [b]shehzad@uet.edu.pk

once before purchasing a product. More than 81% analysis reports [8] illustrate that the product reviews or sentiments have a great influence on purchase [9] and merchandising [10]. Before World Wide Web, people had limited choice for making a decision by themselves, their friends and families. However, analyzing opinions, and monitoring user-generated content is still not a simple task due to diverse nature of practical applications, different languages and the huge volume of textual data. Most of blogs and tweets have some hidden opinions. Thus, automated opinion or sentiment analysis is crucial task of natural language processing. Currently, Sentiment Analysis (SA), computational linguistics and text mining are growing disciplines of Information Retrieval (IR) and natural language processing.

A bulk of unstructured and subjective information in online reviews requires some statistical approach to analyze textual data. Researchers are implementing automatic techniques such as supervised machine learning and symbolic technique for opinion processing [11]. The supervised machine learning technique achieves very high accuracy but requires human involvement such as a large amount of benchmark dataset. This dataset consumes a lot of time to train the classification model. On the other hand, the symbolic technique gives limited accuracy and does not include human involvement. Its performance depends on the scoring method.

A novel clustering-based technique for sentiment analysis is proposed by Li and Liu [12] to overcome major issues in traditional supervised machine learning and symbolic techniques. Clustering method neither requires human involvement and prior knowledge to explore important features from unstructured data nor rejects linguistic information [13]. However, clustering method is still not proved comparable to traditional supervised learning technique. Li and Liu [12] used k-means algorithm for sentiment cluster in two different groups (positive and negative). K-means clustering results are unstable due to the random selection of initial centroids. All sentiments not only belong to true positive or negative class but some opinions are neutral in context [14]. The aim of our research is to further enhance clustering approach to improve the accuracy of unsupervised learning models in the domain of sentiment analysis and address the impact of dimensionality reduction techniques with soft clustering approaches.

Rest of paper is organized as follows. In Section 2 we present literature review of sentiment analysis. Performance analysis of existing clustering approach with problem statement is described in Section 3. Our proposed approach of producing more efficient and accurate results is described in Section 4. Step by step experimental analysis and results are described in Section 5. In Section 6, discussion and evaluation of results are presented and finally Section 7 presents conclusion and further research directions.

2. LITERATURE REVIEW

Semantic orientation is main research objective of sentiment analysis (identifying the polarity of an opinion positive or negative) at phrase, sentence or document level [15]. It determines the reviewer's attitude towards the discussed topic. A sentiment contains opinionated words to express opinion polarity. For example, good, gorgeous and amazing are positive opinionated words while bad, ugly and poor are negative words. In natural language, usually, adjectives and adverbs are considered as opinionated words. These conjoined adjectives or adverbs help to identify the opinion class of sentiment through classification and other log-linear regression techniques [16].

The main stream approaches conducted to analyze the sentiments are symbolic techniques, and the supervised and unsupervised machine learning techniques [12]. Supervised machine learning and symbolic techniques are traditional approaches for sentiment analysis and used by many researchers in text mining domain.

2.1 Supervised Learning Techniques

The first sentiment analysis was conducted by Pang *et al.* [17] on a movie reviews dataset. They developed a classification model of supervised machine learning technique to classify reviews in negative and positive classes. They applied three classification models, namely Support Vector Machine (SVM), Naive Bayes (NB) and Maximum Entropy (ME) model with different classification features. Their experiments achieved accuracy of 77% and best accuracy of 77.7% for NB and ME respectively and SVM gave the best accuracy of 75.1%. SVM with NB features (NBSVM) produced fine results for sentiment analysis but it used an interpolation parameter which was the main drawback of NBSVM [18]. Recent few studies showed that NB produced more than 80% accuracy results after applying sentiment analysis on movie reviews dataset [19]. Multivariate Bernoulli Naive Bayes is another variant of NB Classifier but it only performs better with unigram feature [19].

K-nearest neighbor is nonparametric instanced-based learning model. This model assigns an output class membership to a given test sample by voting its nearest k neighbors of training data samples for optimal solution [20]. SA using K-NN classifier yielded an accuracy of 69.81% for movie review dataset [19]. Kataria and Singh [20] used K-NN using Genetic Algorithm (GA) that led higher classification accuracy up to 90% and reduced training size [20].

Subjectivity and objectivity analysis in a sentiment is another step towards accuracy improvement, categorizing opinion polarity by identifying subjective sentences in a review and removes extra and misleading text. Pang *et al.* [17] used a cut based technique to determine clause of subjective terms in a sentiment. They trained simple Naive Bayes classifier from 5000 subjective and 5000 objective sentences and extracted subjectivity (PS_{ij}) and objectivity (PO_{ij}) probabilities of reviews sentences. Association graph was built to associate relatedness of one sentence with the other. A minimum cut based technique was used on the associated graph to remove extra and misleading (objective) sentence from review to enhance the sentiment classification results [17]. This cut-based subjectivity detection increased accuracy to 86.4% and 86.15% for NB and SVM respectively. In 2005 Pang and Lee [21] further extended their research to classify multi-class (Positive, negative and neutral) based on author's rating [21]. Light weight discourse analysis [22] is another direction which analyzes the twists and turns in a sentiment [23]. A summary of the most important articles in opinion mining is presented in Table 1.

2.2 Symbolic Techniques

The symbolic technique is a process of applying polarity score to each term of a sentiment. This score indicates the intensity of sentiment term in an opinion. The average sum of all terms score gives whole document score. Scoring method directly influences the performance of the symbolic technique. A simple scoring method firstly used for the symbolic technique

Table 1: Summary of the most important recent research articles in opinion mining

Approach	Technique	Year	Accuracy	Efficiency	Drawbacks
Supervised Machine Learning	Naïve Bays, SVM, ME	2016	86.23%, 88.94%, 88.48%	Slow	Requires Human Participation, Prior Knowledge, training overhead, high-dimensional data, time Complexity
	Naïve Bays, KNN	2016	82.4%, 69.8%		
	Naïve Bays, SVM	2017	87.06%, 97.25%		
Symbolic Technique	SentiWord-Net (corpus and dictionary based)	2011	76.37%	Fast	Unable to handle semantic dependency of words, efficiency is domain dependent, Requires a large size dictionary context for fine grind analysis Fast
	PMI	2010	65.5%		
	WordNet (dictionary-based)	2014	63.01%		
Un-Supervised Machine Learning	K-mean	2014	89%	Fast	Instability of results, Outliers issue in clusters, number of clusters mostly unknown
	Bisecting K-mean	2017	68.2%		

is score applying to each term through human participation [24] using word bank. But this technique is highly dependent on human understanding and background domain knowledge.

WordNet based score is calculated for only adjective terms of reviews [25]. Graph of adjective terms generated with nodes as terms and values at vertices shows the synonyms relation between words. Two baseline words ‘good’ and ‘bad’ are used as negative and positive reference terms respectively. The distance of word from these reference terms expresses terms direction towards negative or positive intensity. Based on this reference distances $\mathrm{D}\,(\text{‘good’}, \mathrm{w})$ and $\mathrm{D}\,(\text{‘bad’}, \mathrm{w})$ an evaluation method is used to calculate final word score as

$$\mathrm{EVA}\,(\mathrm{W}) = \frac{\mathrm{D}\,(\text{‘good’},\mathrm{w}) - \mathrm{D}\,(\text{‘bad’},\mathrm{w})}{\mathrm{D}\,(\text{‘good’},\text{’bad’})} \tag{1}$$

This Evaluation function produces word score in the interval of [-1, 1]. Turney [27] achieved 70% Accuracy based on this WordNet scoring method. Semantic relatedness in Senti-WordNet for a large number of synset of senti-words can further increase scoring accuracy [26]. Turney [27] used Pointwise Mutual Information (PMI) for terms score based on assumption that if two words occur more frequently then they are most similar [28]. The PMI for two words w1 and w2 can be calculated as

$$\mathrm{PMI}(\mathrm{w1}, \mathrm{w2}) = \frac{\mathrm{P}(\mathrm{w1},\mathrm{w2})}{\mathrm{P}(\mathrm{w1})\mathrm{P}(\mathrm{w2})} \tag{2}$$

Turney applied this scoring technique on different reviews topic and achieved 65.83% accuracy for movie reviews and 84% for automobile data. Studies show that the accuracy is reduced when this PMI technique applies to an unbiased data as compared to a biased dataset. This traditional machine learning technique requires a lot of human understanding for Natural Language Processing (NLP) and background domain knowledge. Consequently, the learning models are language and domain-specific [29].

2.3 Unsupervised Learning Techniques

Instead of the two above discussed traditional techniques, Li and Liu [12] applied a clustering-based technique in the domain of sentiment analysis. They applied basic k-means clustering algorithm for opinion analysis on movie review data (IMDB). They randomly selected 300 positive and 300 negative reviews to obtain fair distribution and applied the Porter Stemmer and Stanford Part Of Speech (POS) tagger. Since adjectives and adverbs are good opinion indicators [30], they selected more than 6000 adjectives and adverbs for further analysis. They built a term-incidence and term-frequency matrix of adjectives and adverbs as a feature set from movie reviews. They achieved 55% accuracy from term-frequency and term presence matrix. These Clustering results were very unstable with Standard Deviation ranging from 2% to 4%. In the same publication, Li and Liu applied Term Frequency-Inverse Document Frequency (TF-IDF) weighting on term-document matrix:

$$\mathrm{TF-IDF\ Weight} = \mathrm{tfi} \times \log\left(\frac{\mathrm{D}}{\mathrm{dfi}}\right) \tag{3}$$

where tf_i is frequency of term t_i in document d_i in **D** and df_i is frequency of terms in whole dataset **D**. TF-IDF weights dramatically increased average accuracy up to 72% and 73% for frequency and presence of data respectively. But accuracy fluctuation rate highly increased. The standard deviation for term-frequency data reached to 4.02% and 6.2% for the presence of data. In order to overcome instability issue, they used a voting mechanism for final sentiment class identification. For example, they ran clustering algorithm N times, and identified the two classes (positive and negative) by using the following formula (equation (4)):

$$\left\{\mathrm{S(class)} = \right. \\ \left\{\begin{matrix} \mathrm{Positive, if\ N(Classj = positive)} \geq \mathrm{N(Classj = negative)} \\ \mathrm{Negative, if\ N(Classj = positive)} < \mathrm{N(Classj = negative)} \end{matrix}\right\} \tag{4}$$

Since term scores highly impact to identify the polarity of document hence they further applied a score based symbolic technique, the previously discussed score using Word-NET technique to enhance further accuracy results. They [12] calculated the score using geodistic distance in equation (5) of an adjective from reference word ‘good’ and ‘bad’ and finally got polarity weight towards reference words using experimental threshold equation described below.

$$\mathrm{w} = \begin{cases} 1.2(\mathrm{X}-1)\times 0.02,\ \mathrm{x}\leq 8 \\ 1-(\mathrm{X}-1)\times\ 0.1, 8<\mathrm{x}\leq 11 \end{cases} \tag{5}$$

By combining term score w with already discussed clustering method they built a hybrid system for sentiment identification which significantly increased the accuracy rates up-to 77% on average with low standard deviation rate. Consequently, use of adjective-terms decreased dimensionality overhead as well as term score influenced the accuracy fluctuation rate.

Clustering based sentiment analysis approach was further improved as compared to the above-discussed results [14]. The Opposite Opinion Content Processing (OOCP) technique was applied to identify opposite opinion sentence in reviews. Negation words were used (not, nor, neither) with some discoursed relation (use the conjunction like but, although, unlike) in sentences to identify opposite opinions in a sentiment. The sentence polarity was inverted if it contained any negation words to extract opposite opinions in sentiment. Although OOCP technique did not highly impact on fluctuation rate, there was an increase of 1% in accuracy.

Since, in Natural Language (NL) each sentence in a review document does not contain opinionated words but some extra and misleading text may be present in documents to describe a reviewer’s complete attitude. The authors of reference [31] applied a cut based subjectivity extraction technique to remove non-opinion contents from review documents. This cut based classifier calculated the subjectivity probability $P_i(S)$ and objectivity $P_i(O)$ using the Naive Bayes model trained on 5000 subjective and 500 objective sentences. These probabilities were further used to build association graph G[V,T] where V are vertices (sentences) and T are edges (association probabilities) between nodes of V adjective node (Obj), subjective node (Sub) in G. A minimum cut was applied on G to partition review sentiments as objective and subjective. These two enhancements improved the accuracy of k-means clustering technique up to 4% as applied in the previous publication and finally achieved 88.7% and 88.9% for term frequency and term presence respectively.

Apart from improvements in binary clustering method, they further applied three class clustering techniques to find the neutral opinion in movie review dataset. Modified voting mechanism was used to identify positive, negative and neutral opinion (equation (6)).

$$\mathrm{S(class)} = \begin{cases} \mathrm{Pos,\ if\ N\ (Classj\ =\ Pos)\ >\ T} \\ \mathrm{Neg, if\ N\ (Classj\ =\ Neg)\ >\ T} \\ \mathrm{Neutral,\ Otherwise} \end{cases} \tag{6}$$

Li and Liu achieved the maximum of 60% accuracy with a balanced review class dataset which is quite higher than the baseline (33%) of the three class sentiment classification techniques.

Current techniques are focused towards binary class identification such as a positive and negative opinion. Neutral opinion analysis is currently growing research area. Identifying an opinion as strong negative, positive and neutral opinion is conducted by Nakov *et al.* [32].

Recently, Ma *et al.* [33] investigated different initialization procedures and weighting methods to improve k-mean’s clustering performance on different polarity datasets [33]. They achieved best accuracy result for IMDB review dataset, through weighting method of DPH Divergence From Randomness (DPH-DFR). The weight w_{ij} for term t_i in document length dl_j is described as,

$$\mathrm{wij} = \begin{cases} 0, \qquad \text{if } \mathrm{tf_{ij}} = 0 \\ \dfrac{(1-\frac{\mathrm{tftf_{ij}}}{\mathrm{dl_j}})^2}{\mathrm{tfij+1}} \times \left\{\mathrm{tf_{ij}} \times \log\dfrac{\mathrm{tf_{ij}}\times \mathrm{avg\ dl}}{\mathrm{dl_j}}\right\} \\ +0.5\ \times \log\left[2\pi\ \times \mathrm{tf_{ij}} \times \left(1-\dfrac{\mathrm{tf_{ij}}}{\mathrm{dl_j}}\right)\right], \mathrm{x}\geq 0 \end{cases} \tag{7}$$

For document clustering, they applied k-means algorithm to DPH-DFR matrix and finally achieved an average accuracy rate of 68.2%. In their experiments, the K-mean algorithm performed better than other clustering approaches with the best accuracy of 78%.

The most recent research articles in the domain of sentiment analysis in Movie Review domain are summarized in Table 1. A comparison is described in terms of different techniques adopted, their efficiency and drawbacks.

Recent researches do not apply dimensionality reduction techniques to overcome computational cost and improve the performance of Sentiment Analysis. In real-world problems, performance and effectiveness of algorithm is the goal of the analysis. In addition, the randomness of the k-means algorithm makes it mathematically not suitable model. Initialization is another issue of the k-means clustering algorithm. Due to the instability of K-means results, multiple iterations of clustering algorithm [12] are required for stable outcomes which are not suitable for real-world large size cluster dataset. All these issues directed towards new research trend in unsupervised sentiment analysis.

Existing clustering approach is applicable in the real-world analysis where there is no previous knowledge about patterns inside samples and its grouping. However, this approach is not mature as compared to Classification technique and requires further improvement to enhance the performance of clustering based sentiment analysis. In this research, Firstly, we have applied dimensionality reduction technique for improving the performance of traditional k-means clustering approach. Secondly, we have applied another probabilistic clustering (soft clustering) technique to overcome the issues of the K-means algorithm. We have applied Soft clustering techniques to extract more stable clustering results.

3. PROPOSED SOLUTION

This section describes our organization and methodology to speed up analysis approach. Our workflow is described in Fig. 1. We used Pong Lee Internet Movie Review Database (IMDB) (http://www.cs.cornell.edu/people/pabo/-movie-review-data/, the URL contains hyphens only around the word “review”) because of its extensive use in recent research of sentiment analysis domain [8,12,14,21,31]. IMDB review dataset consists of 2000 review documents including 1000 Positive and 1000 Negative movie reviews. Due to computational cost and space complexity, we have randomly selected 600 documents, 300 positive and 300 negative respectively, from IMDB review documents. This selection highly decreases data dimensions issue in experiments.

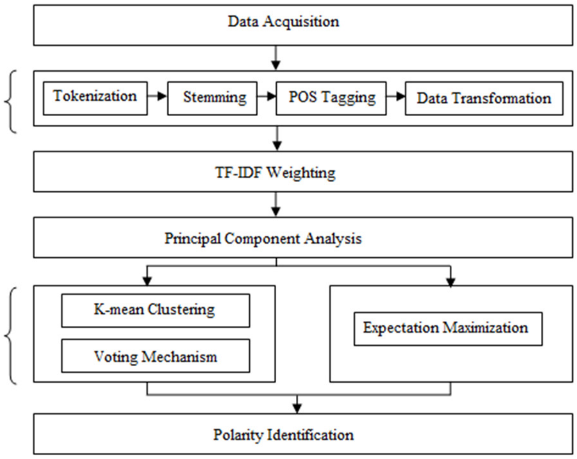

Fig. 1: Architectural Flow of Proposed Solution

3.1 Pre-Processing

To transform textual data in the statistical form to apply ML algorithms we applied basic NLP task and obtained a well-organized bag of words for term-document matrix generation. We used c# libraries for text pre-processing. We parsed review documents and generated a list of tokens for potential possessing. Further, we used maximum entropy Stanford Part of Speech (POS) Tagger [34] to tag list of the bag of words. Stanford tagger has accuracy up to 78% to correctly apply POS tags. Since adjectives play a key role for sentiment polarity identification hence we have selected only adjectives for future analysis. A Term-Document Matrix (TDM) is the suitable approach to convert textual data in a statistical layout.

3.2 TF-IDF Weighting

We also applied term frequency- inverse document frequency (TF-IDF) weighting scheme to highlight term importance within a document and whole corpus.

$$\mathrm{TF-IDF\ Weight} = \mathrm{tfi} \times \log(\mathrm{D}/\mathrm{dfi}) \quad (8)$$

We applied this technique to our TDM and generated a new document matrix, TF-IDF for term frequency and TP-IDF for term presence for both versions of data.

3.3 Dimensionality Reduction

In real-world applications, the dataset is large and consists of a huge number of adjectives. Further processing of these adjectives is a more crucial task. We can reduce data matrix size overhead by applying dimensionality reduction technique to reduce TF-IDF or TP-IDF matrix size. We applied the Principal Component Analysis (PCA) technique for dimensionality reduction which reduces high dimensional dataset to lower dimensions without losing any semantic relatedness between correlated variables by revamping these variables in new space [35]. PCA helps to express data in form of similarities and differences by an orthogonal transformation of original data. PCA is based on Singular Value Decomposition (SVD) factorization (described in Equation (9) to obtain low-rank approximation of original dataset.

$$X = S\Sigma V^T \qquad (9)$$

where S is $m \times m$ matrix of orthonormalized largest eigenvectors of XX^T and V is a $n \times n$ matrix of orthonormalized eigenvectors of X^TX. $\sum$ is a diagonal $m \times n$ matrix whose diagonal values are square root of non-negative real eigen values of XX^T. V^T is the transpose of V.

where

$$XX^T = (V\Sigma V^T)(V\Sigma S^T) = V\Sigma^2 V^T \qquad (10)$$
$$X^TX = (V\Sigma S^T)(S\Sigma V^T) = S\Sigma^2 S^T$$

PCA (Table 2) transforms the set of input features X = {X1, X2, . . . ,X_n} to a new space Y = {Y1,Y2, . . . ,Y_n} where the Y’s are containing higher variance information of original data (Principal Components). For a given original matrix X if S’T is feature vector then transformed Principal Components can be computed as Y= $S^{'T}X$ and by selecting first k rows from Y we have projected n dimensions reduced to just k dimensions. PCA depends on two basic moments, mean of adjusted data and Covariance matrix. Mean vectors for original matrix X can be calculated as

$$X = \frac{1}{m}\sum_{i=1}^{m} X^i \qquad (11)$$

Covariance matrix can be calculated as

$$C = \frac{1}{m}\sum_{i=1}^{m}\left(X^i - (X)\right)(X^i - (X))^T = \frac{1}{m}XX^T \qquad (12)$$

Table 2: Basic Steps for Principal Component Analysis

1.	Adjust data to zero means dataset.
2.	For original matrix X find Covariance matrix Y
3.	Calculate normalized eigenvectors and eigen values of Y.
4.	Sort the eigen vectors according to eigen values in descending order.
5.	Take transpose of eigenvectors result in F feature vector.
6.	Multiply F with transpose of zero mean adjusted dataset to obtain PCA Components

We used simple MATLAB function PCA() to obtain Principal Components (PC) from original matrix. After applying PCA to TF-IDF and TP-IDF matrices we contracted new transformed matrix called as TF-IDF_{PCA} and TP-IDF_{PCA}. This technique highly reduces our matrix to very few dimensions (less than number of documents used for experiments). We randomly select a different set of principal components and examine clustering process to achieve the best number of PCs for accurate and stable results.

3.4 K-Means Clustering

K-means algorithm is the most popular clustering algorithm used to natural K groups in dataset due to its easy implementation. Corresponding to Lee's experiments [12] we used cosine distance function to calculate distance in the k-means algorithm to diminish variable length issue among documents. We applied voting method previously used to stabilize k-mean's result in equation (6).

The k-means algorithm is mathematically not suitable clustering model. It is a hard clustering approach where each sample is associated with only one cluster often not acceptable. Due to a random selection of initial centroids, it produces more unstable results. Unlike this traditional clustering approach, soft clustering model correlates each sample to probabilities of association to a cluster (soft assignment.

3.5 Expectation Maximization

Expectation Maximization (EM) algorithm is based on Gaussian Mixture Model of probability distributions. For space of k random variable X, the estimated Probability Distribution Function (PDF) for a random variable can be described as

$$\mathrm{PDF}\left({}^{X}/_{\mu,\sigma}\right) = \frac{1}{\sigma\sqrt{2\pi}} e^{\frac{-(x-\mu^2)}{\sigma^2}} \qquad (13)$$

The Gaussian mixture model for a given density function is described as follows

$$P(X) = \sum_{i=1}^{m} w_i \times \mathrm{PDF}\left({}^{X}/_{\mu,\sigma}\right) \qquad (14)$$

where w_i is mixture component weight with constraint summation $\sum_{i=1}^{m} w_i = 1$.

EM algorithm is (a soft version of the k-means algorithm) an iterative algorithm to find the maximum likelihood of posterior probabilities from initial few data points. This algorithm consists of two steps *i.e.* Expectation step and maximization step. Expectation step computes latent variables (Mean (μ), Covariance (σ) and Weight (w) values and maximization step updates our learning model function through new computed (in expectation step) values of latent variables. For a k-dimensional data with N samples, Expectation step follows.

$$\tau(z_{nk}) = \frac{w_k \mathrm{PDF}\left({}^{X_n}/_{\mu,\sigma}\right)}{\sum_{i=1}^{k} w_i\, \mathrm{PDF}\left({}^{X_i}/_{\mu_i,\sigma_i}\right)} \qquad (15)$$

and the latent variable mean (μ), covariance (σ) and mixture component weight (w) can be computed in maximization step as

$$\mu_k^{new} = \frac{\sum_{n=1}^{N} \tau(z_{nk}) x_n}{N_k} \qquad (16)$$

$$\sigma_k^{new} = \frac{1}{N_k} \sum_{n=1}^{N} \tau(z_{nk})(x_k - \mu_k^{new})(x_k - \mu_k^{new})^T \qquad (17)$$

These new latent variables are used in expectation step iteratively. In our experiments, we used MATLAB functions for EM algorithm. Since EM algorithm requires an initial start for Expectation step, we used k-means clustering output as an initial start for EM algorithm.

3.6 Label Identification from EM Probabilities

Unlike traditional k-means clustering, EM computes probabilities for predicted labels associated with a data sample. Consequently, it provides soft membership for sample documents. We associate a clustering document to a cluster for which document contains the highest association probability.

4. EXPERIMENTAL RESULTS

In this section, we describe experiments performed for clustering based sentiment analysis. We perform step by step experiments to analyze the performance behavior of dimensionality reduction technique and impact of the soft clustering algorithm in sentiment analysis.

4.1 Data Acquisition

In our experiments we used two processed versions of raw reviews of IMDB dataset after a lightweight pre-processing. IMDB contains a processed version of raw reviews after removing some noise and extra HTML tags (by four different authors) that is available online. Authors processed this version V1 dataset and labeled them according to the rating scale. They further applied Subjectivity extraction technique [31] to scaled Version V1' dataset. The extracted dataset after applying subjectivity summarization is available on the same site.

4.2 Experiment 1: Preliminary investigation

We applied TF-IDF weighted mechanism on sub-set of IMDB dataset (Scale version V1 and Subjectivity extracted (V1') provided by Pong and Lee). We have Computed TF-IDF and TP-IDF matrix for term frequency and term presence respectively. Matrix dimensions are 600-by-1223 for 600 randomly selected movie review documents.

Table 3: Preliminary Results without PCA

	TF	TP	TF'	TP'
No. of DOC	600			
No. of Features	1223 adjectives			
No. of iterations	30			
Average	73.30 %	76.78 %	85.76 %	86.51 %
Maximum	77.83 %	82.50 %	87.83 %	89.50 %
Minimum	68.00 %	71.16 %	81.83 %	83.67 %
STD	2.84%	2.36%	1.25%	1.49%

We applied K-means algorithm with voting method (used in Gang Lee's research [14]) to achieve natural groups of sentiments. We repeatedly applied clustering step to acquire set of results and computed an average accuracy. The standard deviation rate in this accuracy set (Table 3) describes the efficiency and consistency of our clustering models in Fig. 2.

4.3 Experiment 2: Overview of Accuracy enhancement for K-mean algorithm through PCA

After applying PCA to TF-IDF and TP-IDF matrix we contracted new transformed matrix called as TF-IDF$_{PCA}$ and TP-IDF$_{PCA}$. This technique highly reduces the matrix to very few dimensions (less than number of documents used for experiments).

We have randomly selected a different set of Principal components and examined clustering process to achieve the best number of PCs for accurate and stable results. We again applied K-means Algorithm to these reduced dimensions PC matrices. After 30 times applying K-means approach, we obtained results of applying clustering procedure (denoted as TF_{P-K}, TP_{P-K} or Scaled V1 data and TF'_{P-K}, TP'_{P-K} for Subjectivity Extracted version) displayed in Table 4. We obtained a significant improvement in results (denoted as TF'_{P-K}, TP'_{P-K}. Thus processed data produced more accurate and stable results as compared to base-line approach.

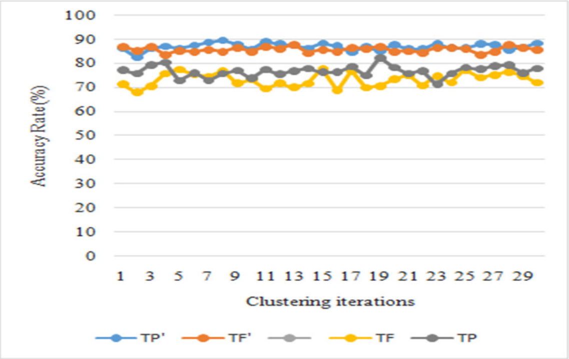

Fig. 2: Preliminary Results of Accuracy without PCA

Table 4: Results of K-mean algorithm after applying PCA

	TF_{P-K}	TP_{P-K}	TF'_{P-K}	TP'_{P-K}
No. of DOC	600			
No. of Features	599 PCA components			
No. of iterations	30			
Average	77.91 %	85.05 %	94.74 %	96.32 %
Maximu m	79.83 %	87.16 %	95.33 %	96.83 %
Minimum	76.5%	82.67 %	94%	95.16 %
STD	0.83%	1.09%	0.32%	0.41%

The result of applying the k-means algorithm to Principal Components of term Frequency and term presence is shown in Fig. 3.

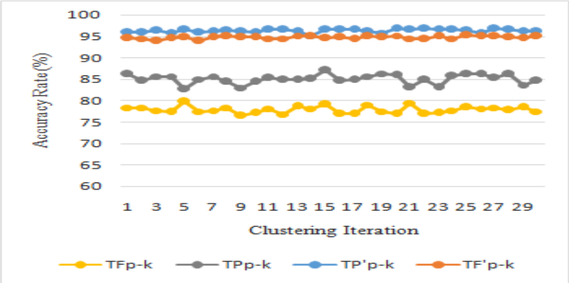

Fig. 3: Accuracy results of K-mean algorithm after applying PCA.

4.4 Experiment 3: Performance Enhancement of sentiments analysis through EM algorithm

In our third experiment, we have applied EM algorithm on reduced matrices. The results of EM algorithm after applying PC matrices on both versions of data (V1 and V1') are described in Table 5.

The result of applying the k-means algorithm to Principal Components of term Frequency and term presence is elaborated in Fig. 4.

5. Discussion and Evaluation of Results

5.1 Impact of PCA on Performance of K-means Algorithm

According to Ding and He [36], PCA is a continuous solution for binary clustering. According to their theorem 2.2, the Objective function for k-means clustering is to maximize least sum of square error between clustered points and centroids. PCA provides rotation of original data in new space and maximizes variations in initial few dimensions. In addition, Nepoleon and Pavalakodi [37], give evidence that initial centroids for k-means clustering. PCA is a non-deterministic method for k-means initialization which outperforms for the large dataset (Survey by Celebi *et al.*) [38]. We experimentally proved above described theorems.

A comparison of baseline results with PCA-Kmeans in Fig. 5 and Fig. 6 describe improvements in clustering accuracy throu6gh dimensionality reduction. PCA proportionally improves clustering accuracy for TF-IDF and TP-IDF matrices.

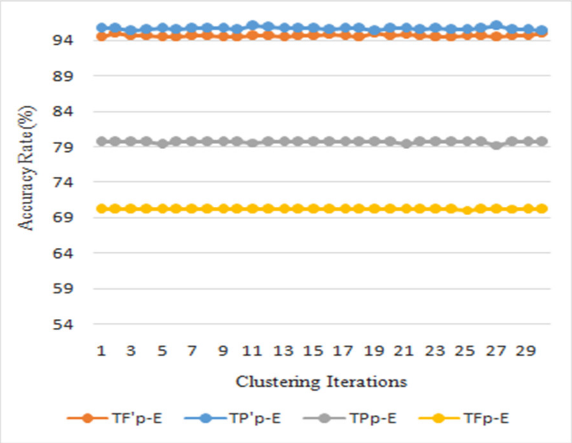

Fig. 4: Accuracy Results of EM algorithm after applying PCA

Table 5: Results of EM Algorithm after applying PCA

	TP$_{P-E}$	**TF$_{P-E}$**	**TF'$_{P-E}$**	**TP'$_{P-E}$**
No. of DOC	**600**			
No. of Features	599 PCA Component			
No. of iterations	30			
Average	79.66 %	70.16 %	94.65 %	94.68 %
Maximum	79.67 %	70.26 %	95.00 %	96.00 %
Minimum	79.21 %	70.16 %	94.50 %	95.33 %
Standard devotions	0.12%	0.02%	0.15%	0.16%

For IMDB scaled Version V1 data, Maximum accuracy rate for TP$_{P-K}$ is 87.1% which is more accurate from baseline results 82.5% of TP in Fig. 5. For subjectivity extracted scaled Version V1 data, Maximum accuracy achieved after dimensionality reduction is 96.83% for TP'$_{P-K}$ is a significant improvement in clustering results as compared to 89.66% baseline results in Fig. 6. Furthermore, Low standard deviation rate demonstrates the significance of PCA before K-means clustering technique.

5.2 Performance comparison of EM and K-means Algorithm

EM algorithm is not based on any cluster distance function. Alternatively, it computes probability based on Gaussian distributions for each observation and assigns it to the respective cluster group.

The objective function of EM algorithm is to maximize overall dataset probabilities for final cluster solution. Due to probabilistic nature, EM algorithm provides review labels as probabilities associated with each cluster. The decision for cluster membership through these probabilities is more accurate as compared to the voting method previously devised for k-means clustering.

EM algorithm (soft clustering approach) is a probabilistic framework that provides a mathematically principled way to understand and addresses the limitations of K-means. Consequently, we apply EM algorithm for a fine grinds opinion analysis.

According to standard deviation rates (shown in Fig. 7), accuracy results for EM algorithm are more stable (relatively straight line) as compared to k-means clustering technique. Furthermore, since in this technique we do not apply Voting mechanism, it executes as fast as compared to k-means approach.

In our experiments, documents which contain equal posterior probability for both clusters (negative and positive) are denoted as multi-class documents. There are four such documents in 600 reviews that are classified as negative as well as positive simultaneously. In these review documents, the ratio

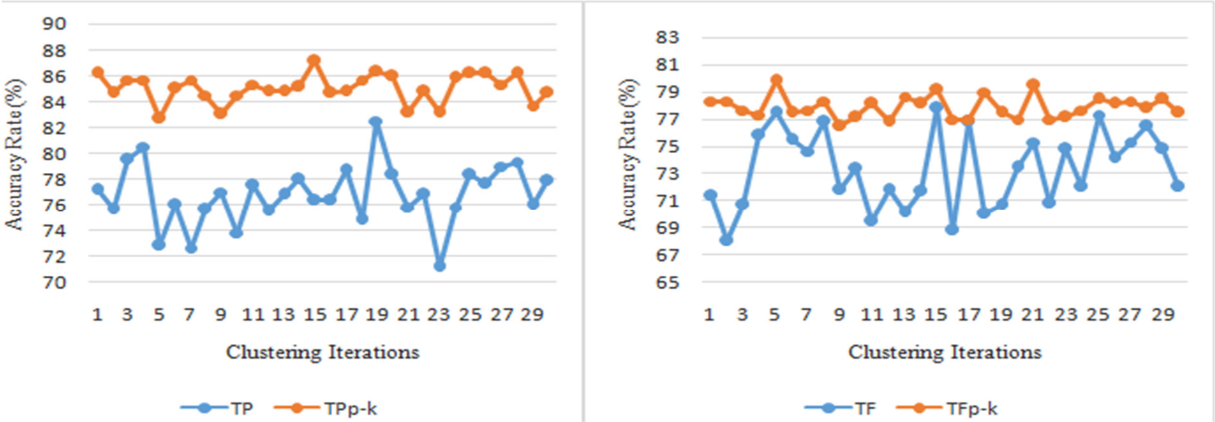

Fig. 5: {(i) Comparison of Term Presence Scaled V1 data before and after applying PCA(ii) Comparison of Term Frequency Scaled V1 data before and after applying PCA

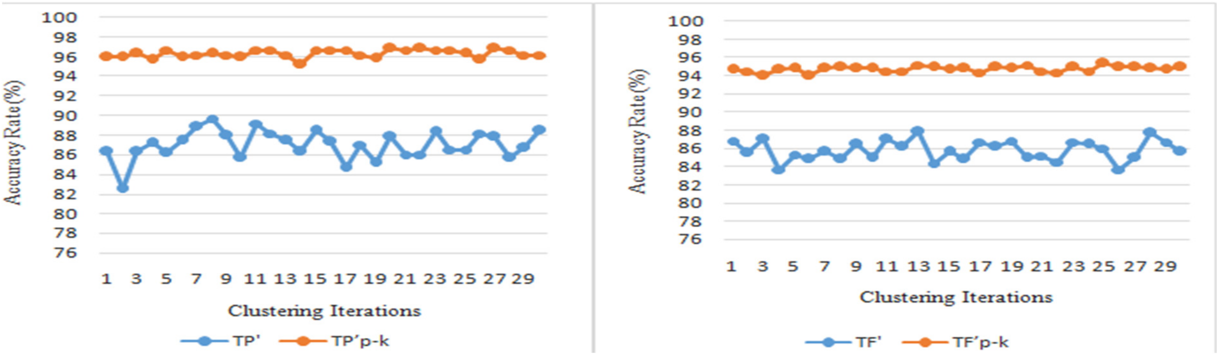

Fig. 6: {(i) Comparison of Term Presence Subjectivity extracted data before and after applying PCA(ii) Comparison of Term Frequency Subjectivity extracted data before and after applying PCA

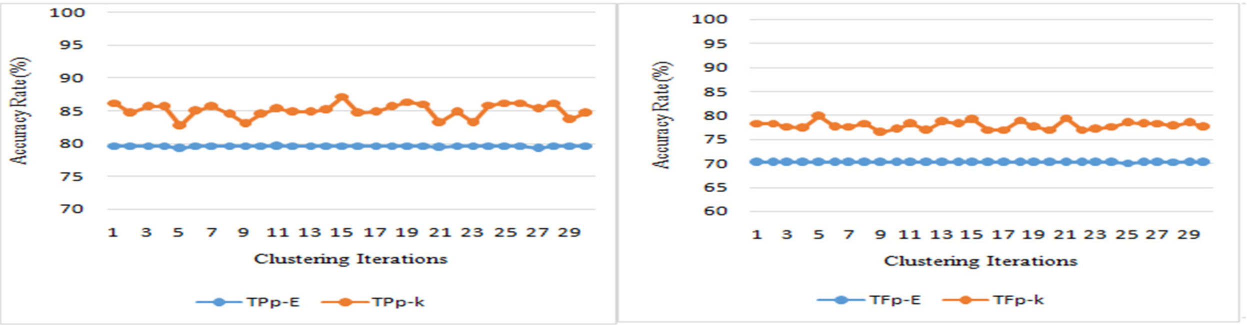

Fig. 7: (i) Comparison between Kmean and EM algorithm on Term Presence Scaled version V1 data. (ii) Comparison between Kmean and EM algorithm on Term Frequency Scaled version V1 data.

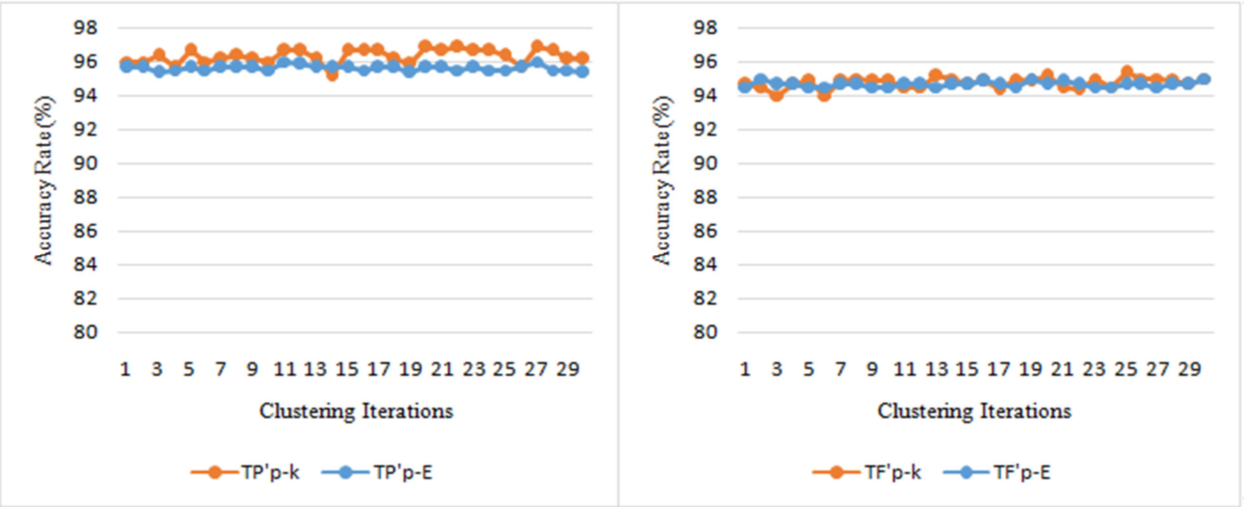

Fig. 8 (i) Comparison between Kmean and EM algorithm on Term Presence of Subjectivity data. (ii) Comparison between Kmean and EM algorithm on Term Frequency of Subjectivity data.

of positive and negative opinions is almost balanced.

5.3 Significance Testing

The experimental results gathered through statistical inference allow us to assess evidence in favor of our proposed technique hypothetically. To verify the impact of dimensionality reduction technique on clustering performance, we applied significance testing to our experimental results. In our Case, pre-tagged dataset is available, hence we are able to apply McNemar's Test [39] to support our proposed clustering approach.

5.4 Performance analysis of K-means clustering with PCA Hypothesis

We validate the effectiveness of k-means clustering results after applying PCA technique through the following hypothesis.

H0 = PCA-K_means is less accurate than baseline approach.
H1 = PCA-K_means Performs better (higher accuracy) than baseline approach.

5.4.1 Evaluation

Let $M_{Baseline}$ is baseline results and $M_{PCA-Kmean}$ are our proposed results. McNemar's statistics to highlight significance of $M_{PCA-Kmean}$ are displayed in Table 6.

h=1 indicates rejection of the null hypothesis ($M_{PCA-Kmean}$ is less accurate than $M_{Baseline}$ accuracies) at 0.5 significance level. Asymptotic p-value < 0.5 addresses a strong evidence to reject the null hypothesis. Consequently, Alternative hypothesis is accepted, which indicates that $M_{PCA-Kmean}$ is statistically significant. Furthermore, Clustering Error rate summarizes that $M_{PCA-Kmean}$ is more accurate than $M_{Baseline}$.

Based on above experiments we compared our clustering results (in Table 7) with previous K-means clustering approaches. We examined that TF-IDF weighting is not enough able to achieve stable results and we need to use some hybrid approaches [12,14]. Li and Liu [12] hybrid technique is effective to achieve stable results but the computational cost for term score calculation is one-off task. Subjectivity extraction has positive impact on binary clustering accuracy [14]. Bisecting K-means [33] does not produce stable and most accurate results.

Table 6: Difference between PCA-Kmean and Baseline (McNemar Test)

Variable	Value
H	1 (Logical)
P	7.5171e-11
$E_{Baseline}$	0.1283
$E_{PCA-Kmean}$	0.0317

Table 7: Comparison of existing clustering results with proposed approach

Year	Algorithm	Feature Set	Best Accuracy	Average Accuracy	Standard Deviation
2012	K-means without voting and hybrid approach	TF-IDF and TP-IDF approach	78%	77%	0.40%
2014	K-means with voting and hybrid approach	TF-IDF and TP-IDF with subjectivity extraction approach	89%	88%	0.50%
2017	Bisecting K-means clustering approach	DPH_DFR	78%	68.20%	
Our Results	K-means with voting	TF-IDF and TPIDF and PCA	96%	95%	0.33%
	GMM without voting	TF-IDF and TPIDF and PCA	96%	95.20%	0.17%

Results show that by applying Principal component analysis before k-means clustering significantly improves clustering accuracy as compared to existing Hybrid approach (Lee and Liu [12]). Hence PCA provides a continuous solution for k-means clustering (Ding's and He's Theorem 2.2 proved experimentally).

Although, our K-means clustering results are close near and better to supervised machine learning results, however, Voting method for K-means clustering forces an overhead for large dataset processing. We applied EM algorithm (soft clustering technique) which produces more stable and accurate results as compared to our k-means clustering consequences. In addition, it speeds p clustering process because EM does not require any voting method to stabilize clustering results.

6. CONCLUSION AND FUTURE DIRECTION

The contribution of our paper is to apply the dimensionality reduction technique into clustering-based sentiment analysis approach to enhance its performance. Currently, by applying PCA we are able to produce higher accuracy in binary sentiment analysis.

Dimensionality reduction through PCA reduces the size of feature set and computational cost. PCA before clustering process provides more accurate and stable results because its first principal components help to select initial centroids for K-means algorithm. We have also verified that voting mechanism overhead may be reduced by applying Gaussian mixture model, a Soft clustering technique. In addition, EM algorithm highly improves the stability of clustering results.

This clustering based approach is language independent and more efficient as compared to classification approaches. Accuracy rate and stability of results are good enough to apply this clustering technique in real-world applications. In future, different clustering techniques can be used to apply in multi-class sentiment analysis for fine grid opinion mining and prediction of review rating.

ACKNOWLEDGEMENT

The authors are thankful to the officials of Department of Computer Science and Engineering UET Lahore for providing research and supportive environment for completion of this research work.

REFERENCES

1. Kaisler S., Armour F., Espinosa A.J., Money W., "Big Data: Issues and Challenges Moving Forward", *Proceedings of the 46th Hawaii International Conference on System Sciences (HICSS)*, pp. 995-1004, Wailea, HI, USA, 2013.
2. Trinh S., Nguyen L., Vo M., "Combining Lexicon-Based and Learning-Based Methods for Sentiment Analysis for Product Reviews in Vietnamese Language". In Lee R. (Ed.):

Computer and Information Science. Studies in Computational Intelligence, Vol. 719, Springer, 2017.

3. Junaid S. M., Jaffry S. W., Yousaf M. M., Aslam L., Sarwar S., "Sentiment Analysis and Opinion Mining-A Facebook Posts and Comments Analyzer", *Techical Journal*, University of Engineering and Technology, Taxila, pp. 89-104, Vol. 22, No. 2, 2017.
4. Kim J., Hastak M., "Social Network Analysis", *International Journal of Information Management*, pp. 86-96, Vol. 38, 2018.
5. Gao J., Zhou T., "Evaluating user reputation in online rating systems via an iterative group-based ranking method," *Physica A: Statistical Mechanics and its Applications*, pp. 546-560, Vol. 473, 2017.
6. Rainie L., Hitlin P., "The use of online reputation and rating systems", *Pew Internet and American Life Project*, Vol. 20, 2004.
7. Athow D., "Online Consumer-Generated Reviews have Significant Impact on Offline Purchase Behavior", *ITProPortal*, 2007.
8. Pang B., Lee L.J., "Opinion mining and sentiment analysis", *Foundations and Trends® in Information Retrieval*, Vol. 2, pp. 1-135, 2008.
9. Li Q., Xing J., Chong W., Liu O, "The Impact of Big Data Analytics on Customers' Online Behaviour", *Proceedings of the International Multitopic Conference of Engineers and Computer Scientisits*, Vol. II, Hong Kong, March 15-17, 2017.
10. Lak P., Turetken O., "The Impact of Sentiment Analysis Output on Decision Outcomes: An Empirical Evaluation", *AIS Transactions on Human-Computer Interaction*, Vol. 9, No. 1, pp. 1-22, 2017.
11. Boiy E., Hens P., Deschacht K., Moens M-F, "Automatic Sentiment Analysis in On-line Text", *ELPUB*, pp. 349-360, 2007.
12. Li G., Liu F., "Application of a clustering method on sentiment analysis," *Journal of Information Science*, Vol. 38, pp. 127-139, 2012.
13. Xu R., Wunsch D., "Survey of clustering algorithms", *IEEE Transactions on Neural Networks*, Vol. 16, pp. 645-678, 2005.
14. Li G., Liu F., "Sentiment analysis based on clustering: a framework in improving accuracy and recognizing neutral opinions", *Applied Intelligence*, Vol. 40, pp. 441-452, 2014.
15. Indurkhya N., Fred J. Damerau, *Handbook of natural language processing*, Vol. 2, CRC Press, 2010.
16. Hatzivassiloglou V., McKeown K.R., "Predicting the semantic orientation of adjectives," *Proceedings of the Eighth Conference on European Chapter of the Association for Computational Linguistics*, pp. 174-181, Madrid, Spain, 1997.
17. Pang B., Lee L., Vaithyanathan S., "Thumbs up?: sentiment classification using machine learning techniques", *Proceedings of the ACL-02 Conference on Empirical Methods in Natural Language Processing,* pp. 79-86, Philadelphia, USA, Vol. 10, July 2002.
18. Wang S., Manning C.D., "Baselines and bigrams: Simple, good sentiment and topic classification," *Proceedings of the 50th Annual Meeting of the Association for Computational Linguistics: Short Papers-* pp. 90-94, Vol. 2, 2012.
19. Dey L., Chakraborty S., Biswas A., Bose B., Tiwari S., "Sentiment Analysis of Review Datasets Using Naive Bayes and K-NN Classifier," *arXiv preprint arXiv:1610.09982*, 2016.
20. Kataria A., Singh M. D., "A review of data classification using k-nearest neighbour algorithm", *International Journal of Emerging Technology and Advanced Engineering*, Vol. 3, pp. 354-360, 2013.
21. Bo Pang B., Lee L., "Seeing stars: Exploiting class relationships for sentiment categorization with respect to rating scales", *Proceedings of the 43rd Annual Meeting on Association for Computational Linguistics*, pp. 115-124, 2005.
22. Taboada M., Voll K., and Brooke J., "Extracting sentiment as a function of discourse structure and topicality", *Technical Report,* Simon Fraser Univeristy School of Computing Science, 2008.
23. Mukherjee S., Bhattacharyya P., "Sentiment Analysis in Twitter with Lightweight Discourse Analysis", pp. 1847-1864, Proceedings of *the 24th International Conference on Computional Linguistics (COLING'2012)*, Mumbai, India, 8-15 December, 2012.

24. Cesarano C., Dorr B.J., Picariello A., Reforgiato D., Sagoff A., Subrahmanian V.S., *Proceedings of the AAAI Spring Symposium on Computational Approaches to Analyzing Weblogs*, pp. 21-36, 2004.
25. Turney P.D., "Thumbs up or thumbs down?: semantic orientation applied to unsupervised classification of reviews", *Proceedings of the 40th Annual Meeting on Association for Computational Linguistics*, pp. 417-424, 2002.
26. Herring S.C., *The Encyclopedia of Applied Linguistics*, Wiley, 2012.
27. Turney P., "Mining the web for synonyms: PMI-IR versus LSA on TOEFL", *Proceedings of the Twelfth European Conference on Machine Learning*, pp. 491-502, Freiburg, Germany, 2001.
28. Kamps J., Marx M., Mokken R.J., De Rijke M., "Using WordNet to Measure Semantic Orientations of Adjectives.", *Proceedings of the Fourth International Conference on Language Resources and Evaluation*, pp. 1115-1118, Vol. 4, Lisbon, Portugal, May 2004.
29. Choi Y., Kim Y., Myaeng S-H, "Domain Specific Sentiment Analysis Using Contextual Feature Generation", *Proceedings of the 1st International CIKM Workshop on Topic-Sentiment Analysis for Mass Opinion*, pp. 37-44, Hong Kong, China, November 2009.
30. Benamara F., Cesarano C., Picariello A., Recupero D.R., Subrahmanian V.S., "Sentiment analysis: Adjectives and adverbs are better than adjectives alone.", *Proceedings of the International Conference on Weblogs and Social Media*, pp. 203-206, Boulder Co., USA, 2007.
31. Pang B., Lee L., "A sentimental education: Sentiment analysis using subjectivity summarization based on minimum cuts", *Proceedings of the 42nd Annual Meeting on Association for Computational Linguistics*, p. 271, Barcelona, Spain, 2004.
32. Nakov P., Ritter A., Rosenthal S., Sebastiani F., Stoyanov V., "Sentiment Analysis in Twitter.", *Proceedings of the 10th International Workshop on Semantic Evaluation,* pp. 1-18, San Diego, California, USA, 2016.
33. Ma B., Yuan H., Wu Y., "Exploring performance of clustering methods on document sentiment analysis," *Journal of Information Science*, Vol. 43, pp. 54-74, 2017.
34. Toutanova K., Manning C.D., "Enriching the knowledge sources used in a maximum entropy part-of-speech tagger", *Proceedings of the 2000 Joint SIGDAT conference on Empirical methods in natural language processing and very large corpora: held in conjunction with the 38th Annual Meeting of the Association for Computational Linguistics,* Vol. 13, pp. 63-70, 2000.
35. Bishop C., *Pattern Recognition and Machine Learning*, Springer-Verlag, New York, 2006.
36. Ding C., He X., "K-means clustering via principal component analysis," in *Proceedings of the Twenty-First International Conference on Machine Learning*, Banff, Canada, 2004.
37. Napoleon D., Pavalakodi S., "A new method for dimensionality reduction using K-means clustering algorithm for high dimensional data set", *International Journal of Computer Applications*, Vol. 13, pp. 41-46, 2011.
38. Celebi M.E., Kingravi H.A., Vela P.A., "A comparative study of efficient initialization methods for the k-means clustering algorithm", *Expert Systems with Applications*, Vol. 40, pp. 200-210, 2013.
39. McNemar Q., "Note on the sampling error of the difference between correlated proportions or percentages", *Psychometrika*, Vol. 12, pp. 153-157, 1947.