Hindawi
Applied Computational Intelligence and Soft Computing
Volume 2018, Article ID 8909357, 12 pages
https://doi.org/10.1155/2018/8909357

Research Article

Performance Assessment of Multiple Classifiers Based on Ensemble Feature Selection Scheme for Sentiment Analysis

Monalisa Ghosh and Goutam Sanyal

Department of Computer Science and Engineering National Institute of Technology, Durgapur, West Bengal, India

Correspondence should be addressed to Monalisa Ghosh; monalisa_05mca@yahoo.com

Received 16 May 2018; Accepted 14 August 2018; Published 1 October 2018

Academic Editor: Shyi-Ming Chen

Sentiment classification or sentiment analysis has been acknowledged as an open research domain. In recent years, an enormous research work is being performed in these fields by applying various numbers of methodologies. Feature generation and selection are consequent for text mining as the high-dimensional feature set can affect the performance of sentiment analysis. This paper investigates the inability or incompetency of the widely used feature selection methods (IG, Chi-square, and Gini Index) with unigram and bigram feature set on four machine learning classification algorithms (MNB, SVM, KNN, and ME). The proposed methods are evaluated on the basis of three standard datasets, namely, IMDb movie review and electronics and kitchen product review dataset. *Initially, unigram and bigram features are extracted by applying n-gram method. In addition, we generate a composite features vector CompUniBi (unigram + bigram), which is sent to the feature selection methods Information Gain (IG), Gini Index (GI), and Chi-square (CHI) to get an optimal feature subset by assigning a score to each of the features. These methods offer a ranking to the features depending on their score; thus a prominent feature vector (CompIG, CompGI, and CompCHI) can be generated easily for classification. Finally, the machine learning classifiers SVM, MNB, KNN, and ME used prominent feature vector for classifying the review document into either positive or negative.* The performance of the algorithm is measured by evaluation methods such as precision, recall, and F-measure. Experimental results show that the composite feature vector achieved a better performance than unigram feature, which is encouraging as well as comparable to the related research. The best results were obtained from the combination of Information Gain with SVM in terms of highest accuracy.

1. Introduction

Expeditious growth of the user-generated content on the web requires the generation of an efficient algorithm for mining important information. This situation enhances the importance of text classification whose aim is to categorize the texts into relevant classes according to their contents. In current years sentiment mining has been receiving a lot of attention from researchers as a most active research area in natural language processing. Sentiment mining or analysis is the process of determining the emotional tones behind a series of words.

We were motivated to this work because researches on sentiment analysis are growing to a great extent and attracting wide ranges of attention from academics and industries as well. Understanding emotions, analyzing situations and sentiments linked with it, is human's natural ability. However, how to empower the machines to do the same thing remains a very crucial and important question to be explored and answered. Sentiment analysis offers a huge scope in effective analysis of the attitude, behavior, likes, and dislikes of an individual. Signal processing and AI both have conducted the evolution of advanced intelligent systems that aim to detect and process dynamic information contained in multimodal sources.

There are several areas such as marketing, politics, and news analytics which benefit from the result of sentiment analysis. These solutions can be roughly categorized into

machine learning approach and lexicon-based approach to solve the problem of sentiment classification. The former approach is applied to classify the sentiments based on trained as well as test datasets. The second category does not require any prior training dataset as it performs the task by identifying a list of words and phrases that consists of a semantic value. It mainly concentrates on the patterns of unseen data. There are few researchers who applied hybrid approaches by combining both methods, machine learning and lexical, to improve the sentiment classification performance.

This field becomes more challenging due to the fact that many demanding and interesting research problems still exist in this field to solve. Sentiment-based analysis of a document is quite tough to perform in comparison with topic based text classification. The opinion words and sentiments always vary with situations. Therefore, an opinion word can be considered as positive in one circumstance but may become negative in some other circumstance.

Sentiment classification [1] process has been divided into three levels: document level, sentence level, and feature level. The entire document is classified in document level, based on the positive or negative opinion expressed by the authors. Sentiment classification at the sentence level considers the individual sentence to identify whether the sentence is positive or negative. In feature level, classify the sentiment with respect to the specific aspects of entities. Aspect level sentiment classification requires deeper analysis on features, mainly which are expressed implicitly and are usually hidden in a large text dataset. During this study, the focus has been on feature level sentiment classification.

The main contributions of this paper can be stated as follows:

(i) In this present work, we investigate the performance of the different combination of feature selection methods (FSM) such as n-gram, Info Gain, Gini Index, and Chi-square. After completion of preprocessing on a large high-dimensional movie review dataset, primarily unigram and bigram feature sets are extracted. Further we create a combined feature vector with unigram and bigram features.

(ii) Next, we applied feature selection methods (FSM) to get an optimal feature subset by assigning a weight to each feature. Finally, we trained the supervised classifiers SVM, MNB, KNN, and ME with these optimal feature vectors for classifying the review document.

(iii) We carried out experiments considering the 10-fold cross validation, as product review dataset consists of separate files for positive and negative reviews but training and testing data are not isolated. For movie review dataset, we noticed that the distribution is suboptimal since the training samples are not sufficient according to 25000 testing reviews. Finally, to improve the performance of classifier we decided to use cross validation for movie as well as product review datasets.

(iv) The effectiveness of classification algorithm is evaluated in terms of F1-score, precision, and recall.

The rest of the paper is constructed as follows: Section 2 consists of the existing literature that can be related to our approach. Then Section 3 describes the approaches used in this paper for polarity detection. Section 4 explains methodology that includes features and proposed feature selection technique. The detail regarding implementation of proposed classification algorithm is discussed in Section 5. The particulars about experiments and results are expounded in Section 6. Finally, Section 7 concludes with a discussion of the proposed method and with ideas on future steps.

2. Related Work

In current years, sentiment analysis of social media content has become most sought area among researchers because the numbers of product review sites, social networking sites, blogs, and forums are growing enormously. This field mainly utilizes supervised, unsupervised, and semisupervised technique for sentiment prediction and classification task. In this section we provide a brief overview of the previous studies regarding supervised multiple machine learning (ML) algorithm. All the previous research works related to machine learning classifiers for sentiment analysis have been discussed in Table 1.

Pang et al. employed three different ML algorithms such as SVM, NB, and ME. They considered bag of word framework with n-gram features such as unigram, bigram, and their combination. The performance of SVM algorithm was convincing according to their analysis. Research work of Dave et al. [2] used some tools for analyzing the reviews from Amazon and CNET for classification. They select bigram and trigram features using n-gram model and some scoring methods are applied finally to determine whether the review holds positive or negative opinion. SVM and NB classifier were implemented for sentence level classification with the accuracy of 87.0.

The movie reviews dataset IMDb was used in a study by Annett & Kondrak, 2008 [3]. They adopted lexical resource WordNet for sentiment extraction. Different classifiers such as SVM, NB, and alternating decision tree used for review classification and more than 75% accuracy was achieved.

In some cases [4], SVM classifier separately is unable to provide a satisfactory performance for small datasets, but the combination of SVM with NB classifier performs surprisingly well by integrating the advantages of both classifiers.

Zhang et al. [5] proposed a classification approach of Chinese reviews on clothing products. They applied word2vec and SVM^{pref} technique while word2vec helped to capture the semantic features based on semantic relationship. SVM^{pref} is nothing but an alternative structural formulation of SVM optimization problem for binary classification. They achieved good outcomes of this combination for sentiment classification. Mouthami et al. [6] proposed new approach as sentiment fuzzy classification algorithm on the movie review dataset to improve the classification accuracy. Preprocessing method tokenization, stop word removal, TF-IDF, and POS

Table 1: Research work related to machine learning classifiers for sentiment analysis.

Author/Year	Technical Approach	Accuracy in %	Dataset domain
Pang et al. (2002)	Applied N-gram model with NB, SVM, ME	77.4 – 82.9	Internet Movie Database (IMDb)
Dave et al. (2003)	Used N-gram model for feature extraction with SVM, NB classifier	87.0	Product review from Amazon & CNET
Annett & Kondrak (2008)	Considered WordNet as Lexical resource with SVM, NB, Decision Tree classifier	75.0	Movie reviews (IMDb)-1000 (+) and 1000 (-) reviews
Ye et al. (2009)	NB, SVM classifier used for classification	85.14	Travel Blogs
Mouthami et.al (2013)	TF-IDF and POS tagging with fuzzy classification algorithm	87.4	Movie review dataset
Zha et al. (2014)	SVM, NB, ME classifier adopted with evaluation matrices F1-Measure	83.0- 88.43	Customer reviews (feedback)
Habernal et al. (2014)	n-gram and POS related features & emoticons are selected using MI, CHI, OR, RS method. Classifier ME and SVM used for classification.	78.50	Dataset from social media
Zhang *et.al.* (2015)	Use word2vec for features with SVM classifier for classification	89.95- 90.30	Chinese review dataset
Luo et.al. (2016)	first transform the text into low dimensional emotional space (ESM), next implement SVM, NB, DT classifier.	63.28 – 79.21	Stock message text data

tagging are used for initial pruning. Fuzzy rules [7–9] are implemented with different algorithms in various fields of data mining domain. In [10] they researched on travel blogs and applied various machine learning algorithms, NB and SVM, by considering the n-gram model to obtain the feature set. In this study, SVM worked best with 85.14% accuracy.

The feature selection stage mainly helps in refining features, which are considered as input for classification task. Feature selection is definitely a beneficial task considered by Narayanan et al. [11] based on the experimental result. They have applied only Mutual Information feature selection method with Naive Bayes (NB) classifier in the domain of movie review.

Dey et al. focused on quick detection of sentiment content of online movie reviews and hotel reviews. The statistical method Chi-square test has been used to find positive information and negative score for each feature and create a word dictionary by summarizing information score. The classifiers KNN and NB were applied with detailed explanation, where NB produces better accuracy than KNN classifier for movie review dataset.

Amolik et al. [12] proposed a model for sentiment prediction of more than 21,000 tweets by applying the machine learning classifiers SVM and NB. Feature vectors were also made competent to handle the problem of repeating characters in Twitter. They achieved higher accuracy with SVM (75%) in comparison with NB (65%) by using evaluation matrices precision and recall. A huge number of research papers with different ML classifiers, namely, Naive Bayes (NB) [6, 13], Support Vector Machine (SVM) [4, 5, 14], Maximum Entropy [15–17], and decision trees [3, 13, 18] have been used mostly to build classification model in different domain.

3. Proposed Approach

The proposed classification method is summarized in several steps as described below:

(1) Data Collection: In this work, movie review database (IMDb) and product review (electronics, kitchen) database are considered to solve the problem regarding sentiment classification.

(2) Preprocessing: This technique is required to remove noisy, inconsistent, and incomplete data by considering tokenization, stop words removal, and stemming method.

(3) Feature Extraction and Selection: Initially, to create a feature vector with numeric value, the frequency count of each unigram and bigram was performed. The machine learning classifier needs numerical matrices to perform sentiment classification. The frequency score of each feature from combined feature set (unigram + bigram) is computed and only those features which are considered are those having a value greater than 5. Further, this reduced feature set is sent to the feature selection methods IG, CHI, and Gini Index. Feature selection methods IG, Chi-square, and Gini Index are used to assign a particular weight to each individual feature and create a list of top-ranked features.

(4) Classification: Finally, train the supervised machine learning classifiers SVM, MNB, KNN, and ME with the different feature vector for classification the dataset.

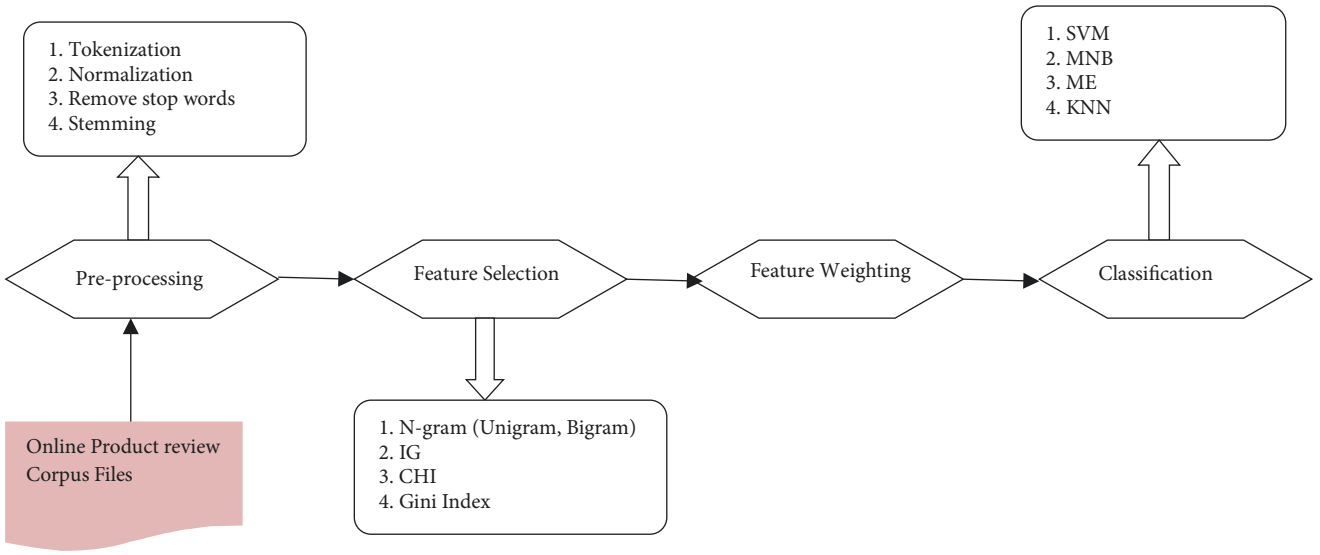

Figure 1: The architecture of a proposed framework for sentiment classification task.

4. Methodology

Text classification as a research field was introduced a long time ago [19]; however, sentiment-based categorization was initiated more recently [2, 16, 20].The ultimate purpose of this research work is to investigate the performance of various machine learning classifiers (MLC) with three combined feature sets. The whole process can be completed in four steps: data acquisition, preprocessing, feature selection, and classification. A general overview of proposed framework is introduced in Figure 1, and the following subsections present a detailed description about each preliminary function.

4.1. The Data: Dataset Preparation. We conducted experiments on movie review dataset, which were prepared by Pang & Lee, 2004 [20]. This study uses movie review and product review dataset (electronics and kitchen) to perform sentiment classification task. The movie review dataset https://www.kaggle.com/nltkdata/movie-review#movie_reviews.zip is one of the popular benchmark datasets, which has been exploited by several researchers in order to analyze the experimental outcomes. The standard movie review dataset consists of overall 2000 reviews where 1000 reviews are tagged as positive and 1000 are negative. The amazon products review dataset http://www.cs.jhu.edu/~mdredze/datasets/sentiment/processed_acl.tar.gz provided by Blitzer et al. (2007) is considered for an investigation and we adopted the dataset of electronics and kitchen domain from the corpus produced by Blitzer et al. Each domain of this corpus has 1000 pos+ and 1000 neg-labeled reviews.

The preprocessing is approved to prepare these three datasets for experiment.

4.2. Preprocessing

(i) **Tokenization or segmentation:** It can be accomplished by splitting documents (crawled reviews) into a list of tokens such as word, numbers, and special characters, making the document ready to be used for further processing.

(ii) **Normalization:** This process converts all the word tokens of a document into either lower case or upper case because most of the reviews consist of both cases, i.e., lowercase and uppercase characters. As a result, tokens (shifted into a single format) can easily be used for prediction.

(iii) **Removal of stop words:** Stop words are very common and high-frequency words. This process was carried out by removing frequently used stop words (prepositions, irrelevant words, special character, and ASCII code), new lines, extra white spaces, etc. to enhance the performance of feature selection technique.

(iv) **Stemming:** It is the process of transforming all the tokens into their stem or root form. Stemming is a swift and easy approach that makes the feature extraction process more effortless.

4.3. N-Grams. N-gram model consists of a contiguous sequence of n words from a given review dataset. Models are generally used with 1-gram sequence, 2-gram sequence, and 3-gram sequence, and sometimes the sequence can be extended.

Example

Text data: "something is better than nothing."

(1 gram or n=1) Unigrams: "something", "is", "better", "than", "nothing".

Table 2: Notation use for feature selection.

Symbol	Description
$P(c_i)$	Probability that a document d is in class c_i
P(f)	Probability that document d contains feature f
$P(\overline{f})$	Probability that a document d does not contain feature f
$P(c_i/f)$	Probability that document d contains feature f in class c_i
$P(c_i/\overline{f})$	Probability that document d does not contain feature f in class c_i

(2 gram or n=2) Bigrams: "something is", "is better", "better than", "than nothing".

(3 gram or n=3)Trigrams: "something is better", "is better than", "better than nothing".

4.4. Feature Selection. Feature selection method (FSM) is an essential task to enhance the accuracy of sentiment classification process. Generally, FSMs are statistically represented by the relationship between feature and class category. The performance of the classifier is mostly dependent on the feature set; if the feature selection method performs well, then the simplest classifier may also give a good accuracy through training. These FSMs are often defined by some probabilities to realize the theoretical analysis of these probabilistic methods. We use a list of notations, which is depicted in Table 3.

Analytical information from the training data is required to determine these probabilities and notations about the training data listed in Table 2, given as follows:

we denote by $C_{i=1}^{m} = \{c_1, c_2, \ldots c_m\}$ the set of classes.

4.4.1. Information Gain (IG). This statistical property is used as an effective solution for feature selection. IG method is used to select important features based on the class attribute rules of features classification. The IG value of each term can measure the number of bits of information acquired for class prediction by knowing the presence or absence of that term in the document [21]. The IG value of a certain term or feature is calculated by the following equation

$$\begin{aligned} \mathrm{IG}(\mathrm{f}) = & \left\{ -\sum_{i=1}^{m} P(c_i) \log P(c_i) \right\} \\ & + \left\{ P(f) \left[\sum_{i=1}^{m} P(c_i \mid f) \log P(c_i \mid f) \right] \right\} \\ & + \left\{ P(\overline{f}) \left[\sum_{i=1}^{m} P(c_i \mid f) \log P(c_i \mid f) \right] \right\} \end{aligned} \quad (1)$$

and it is defined as follows.

$$\begin{aligned} \mathrm{IG}(\mathrm{f}) = & \left\{ -\sum_{i=1}^{m} \frac{N_i}{N_{all}} \log \frac{N_i}{N_{all}} \right\} \\ & + \left(\sum_{i=1}^{m} \frac{W_i}{N_{all}} \right) \left[\sum_{i=1}^{m} \frac{W_i}{W_i + X_i} \log \frac{W_i}{W_i + X_i} \right] \\ & + \left(\sum_{i=1}^{m} \frac{Y_i}{N_{all}} \right) \left[\sum_{i=1}^{m} \frac{Y_i}{Y_i + Z_i} \log \frac{Y_i}{Y_i + Z_i} \right] \end{aligned} \quad (2)$$

IG offers a ranking of the features depending on their IG score; thus a certain number of features can be selected easily.

4.4.2. Chi-square (χ^2). Chi-square (χ^2) is a very commonly applied statistical test, which can quantify the association between the feature or term f and its related class C_i. It tests a null-hypothesis that the two variables feature and class are completely independent of each other. The CHI value of feature f for class C_i is higher, and the closer relationship exists between the variables feature f and class C_i. The features with the highest χ^2 values for a category should perform best for classifying the documents. The formulation of this method is as follows.

$$\chi^2(\mathrm{f}, c_i) = \frac{N_{all..} (W_i Z_i - Y_i X_i)^2}{(W_i + Y_i) \cdot (X_i + Z_i) \cdot (W_i + X_i) \cdot (Y_i + Z_i)} \quad (3)$$

It can also be defined by considering Y_i as (N_i-W_i) and Z_i as (N_{all}-N_i-X_i) and the above formula is rewritten as follows.

$$\begin{aligned} & \chi^2(\mathrm{f}, c_i) \\ & = \frac{N_{all} \cdot [W_i (N_{all} - N_i - X_i) - (N_i - W_i) X_i]^2}{N_i \cdot (N_{all} - N_i) \cdot (W_i + X_i) \cdot [N_{all} - (W_i + X_i)]} \end{aligned} \quad (4)$$

4.4.3. Gini Index (GI). Gini Index measures the feature's ability to discriminate between classes. This method was mainly proposed to be used for decision tree algorithm based on an impurity split method. The main principle of Gini Index is to consider S as a dataset of the sample having m number of different classes $C_{i=1}^{m} = \{c_1, c_2, \ldots c_m\}$. According to the class level, the sample set can be splitted into n subset(*Si*, i=1, 2. . .n).The Gini Index of the set S is

$$Gini\ Index(S) = 1 - \sum_{i=1}^{n} P_i^2 \quad (5)$$

where probability P_i of any sample belongs to class C_i and can be computed by S_i/S [22]. Gini Index for a feature can be estimated independently for binary classification. We adopted Gini Index Text (GIT) method for calculating the feature score, which was introduced by Park et al. [23]. This algorithm was enhanced to overcome the limitations of Gini Index method.

According to previous notation defined in Table 3, we can compute the Gini Index for a feature f of document d belonging to class C_i.

$$GIT_{wi}(f, C_i) = P(C_i \mid f)^2 \quad (6)$$

$$GIT_{Xi}(f, C_i) = \left| \frac{P(Ci \mid f)\, 2}{\log_2 P(f)} \right| \quad (7)$$

5. Classification

5.1. Naive Bayes (NB). Naive Bayes classification method is used for both classification and training. The fundamental theory of NB classifier is based on the independence

Table 3: Notation use for feature selection.

Symbol	Description
N_{all}	The total no. of documents in training dataset
N_i	No. of documents in class c_i
W_i	No. of documents in class c_i containing feature f
X_i	No. of documents not in class c_i but containing feature f
$Y_i = N_i\text{-}W_i$	No. of documents in class c_i not containing feature f
$Z_i = N_{all}\text{-}N_i\text{-}X_i$	No. of documents neither in class c_i nor containing the feature f.

assumption, where the joint probabilities of features and categories are used to roughly calculate the probability score of categories of a given document. It is a simple probabilistic classifier that helps in classifying a document d_r, out of classes $c_i \in C$ ($C_{i=1}^{m} = c_1, c_2, \ldots c_m$). The best class returns in NB classification and is the most probably or maximum posterior (MAP) class C_{map}.

$$C_{map} = \text{argmax}\, P(c_i)\, P(d_r \mid c_i) \quad c_i \in C \tag{8}$$

where the class $P(c_i)$ can be estimated by dividing the number of documents of class c_i by the total number of documents. $P(d_r \mid c_i)$ indicated the number of occurrences of the feature in document d_r belonging to class c_i. The probability value $P(c_i \mid d_r)$ will be computed for each possible class, but $P(d_r)$ does not change for each class. Thus we can drop the denominator.

We thus select the highest probable classes' c_{map} of given document d by calculating the posterior probability of each class.

There are several Naive Bayes variations. In this paper, we consider the Multinomial Naive Bayes classifier.

Multinomial Naïve Bayes (MNB). The multinomial Naive Bayes model [24] is typically used for discrete counts. We consider MNB classifier for text classification task, where a document d is represented by a feature vector ($f_1, f_{2\ldots}$, fn) with the integer value of word frequency in the given document. For multinomial NB model, the conditional distribution $P(d \mid c_i)$ of document d given the class c is as follows.

$$\begin{aligned} \text{Multinomial} \quad P(d_r \mid c_i) &= P((f_1, f_2, \ldots, f_n) \mid c_i) \\ &= \prod_{1 \le j \le n} P(f_j \mid c_i) \end{aligned} \tag{9}$$

The final equation with Bayes' rules the highest probable classes by a Naive Bayes classifier as follows.

$$C_{map} = \text{argmax}\, \widehat{P}(c_i) \prod_{1 \le j \le n} \widehat{P}(f_j \mid c_i) \quad c_i \in C \tag{10}$$

Now, to estimate the probability $\widehat{P}(f_j \mid c_i)$ we consider the feature as a word appears in the document's bag of words. Thus we will compute $\widehat{P}(wj \mid c_i)$ by considering N_{jr} as the number of occurrences of word wj in documents d_r from class c_i among all words in all documents of class c_i. Then the estimated probability of a document given its class is given as follows:

$$P(d_r \mid c_i) = \left(\sum_{j}^{|v|} N_{jr} \right)! \prod_{j=1}^{|v|} \frac{\widehat{P}(w_j \mid c_i)^{N_{jr}}}{N_{jr}!} \tag{11}$$

where v is the union of all the word types in all classes.

The probability of w_j in c_i is estimated from training dataset and it is defined as follows.

$$\widehat{P}(wj \mid c_i) = \frac{\text{count}(w_j c_i)}{\sum_{w \in V} count(w, c_i)} \tag{12}$$

5.2. Support Vector Machine (SVM). Support Vector Machines (SVMs) are supervised learning model introduced [25] for binary classification in both linear and nonlinear versions. Generally, datasets are nonlinearly inseparable, so the main aim of the SVM classifier is to catch the best accessible surface to make separation between positive and negative training samples based on empirical risk (training set and test set error) minimization principal. SVM method can try to define a decision boundary with the hyperplanes in a high-dimensional feature space. This hyperplane separates the vectorized document into two classes as well as determining a result to make a decision based on this support vector [26]. The optimization problem of SVM can be minimized as follows.

Given N linearly separable training set with feature vector x of d dimension, for dual optimization where $\alpha \in R^N$ and $y \in \{1, -1\}$, the solution of SVMs (dual) can be minimized as follows.

$$\begin{aligned} \vec{\alpha}^{*} &= \text{argmin} \left\{ -\sum_{i=1}^{n} \alpha_i + \sum_{i=1}^{n} \sum_{j=1}^{n} \alpha_i \alpha_j y_i y_j \left\langle \vec{x_i}, \vec{x_j} \right\rangle \right\} \\ &\text{Where,} \sum_{i=1}^{n} \alpha_i y_i = 0; \; 0 \le \alpha_i \le C \end{aligned} \tag{13}$$

The classical SVM seems to be able to separate the linear dataset with a single hyperplane, which can separate two classes. For nonlinear dataset where more than two classes are to be handled, kernel functions are used in that situation to lay out the data to a higher dimensional space in which it is linearly separable.

5.3. K-Nearest Neighbor. To identify the class of unknown samples, the KNN algorithm works by inspecting the K-Closest instances in the training dataset and making a

prediction based on to which the majority of its "closest neighbors" belong. KNN algorithm is one of the simplest and effective algorithms, being commonly used for classification and regression. KNN first trained the system with existing review dataset to predict the test samples category.

The classification process of sample S using KNN algorithm is defined as follows [27]:

(i) Suppose there are total N training samples of i categories *(C1, C2...Ci)* and m dimensional feature vector obtained by applying different feature selection method. We prepare the sample S in the form of the vector (s1, s2...sm) as all training samples.

(ii) Calculate the similarities between all training samples and S. Considering the jth training sample dj $(dj_1, dj_{2,\ldots} dj_m)$ estimate the similarity $SIM(S, d_j)$ as follows.

$$SIM\left(S, d_j\right) = \frac{\sum_{i=1}^{m} S_i.d_{ji}}{\sqrt{\sum_{i=1}^{m} S_i^{\,2}} \cdot \sqrt{\sum_{i=1}^{m} d_{ji}^{\,2}}} \tag{14}$$

(iii) Select k samples which are larger than N similarities of $SIM\ (S, d_j)$, where $j= 1,2, N$, and consider them as k-nearest neighbors of sample S. Calculate the probability of S of each category with the following formula.

$P(S,C_i) = \sum_d SIM(S, d_j).y(d_j.C_i)$ where $y(d_j.C_i)$ is attribute function of different category with the following condition.

$$y\left(d_j, C_i\right) = \begin{Bmatrix} 1 & d_j \in C_i \\ 0, & otherwise \end{Bmatrix} \tag{15}$$

Finally, predict the category of sample S with largest $P\ (S,C_i)$

5.4. Maximum Entropy (ME). This is a probabilistic classifier usually used in various NLP applications. This classification technique provides the anticipation that a document belongs to a specific class given a framework to maximize the entropy of the classification document [28]. ME does not make any hypothesis that features conditionally independent of each other, such that the result is more reliable than NB. This classifier needs more time to train than NB classifier as it solves the optimization problem to estimate the parameters of the model. In order to handle the classifier Max Entropy, we should select a feature to set the constraints. For the purpose of text classification, we consider word count as a feature. The ME value can be expressed by exponential form as follows:

$$\mathrm{P_{ME}}\,(\mathrm{c} \mid \mathrm{d}) = \frac{1}{\boldsymbol{z}\,(d)} \exp\left(\sum_i \lambda_{i,c} f_{i,c}\,(d, \mathrm{c})\right) \tag{16}$$

where $\mathrm{P_{ME}}(\mathrm{c} \mid \mathrm{d})$ refers to the probability of document 'd' of class 'c' and z(d) is a normalized function. $\lambda_{i,c}$ indicates the feature-weight parameters to be estimated, if $\mathrm{f_{i,c}}$ (d, c) is the function for feature $\mathrm{f_i}$, and class c feature/class function $\mathrm{f_{i,c}}$ (d, c) can be defined as follows:

$$\mathrm{f_{i,c}}\,(\mathrm{d}, \mathrm{c}) = \begin{Bmatrix} 1 & \mathrm{Ni}\,(\mathrm{d}) > 0,\ c' = c \\ 0 & otherwise \end{Bmatrix} \tag{17}$$

where $\mathrm{f_{i,c}}$ (d, c) is the function for feature $\mathrm{f_i}$, and class c $\mathrm{N_i}$(d) indicates the occurrence of feature 'i' in document 'd'. The feature-class pair which occurs very frequently in document 'd', having high frequency, is the strong indicator for class c. The function which holds a strong orientation will be set to 1; otherwise it will be 0.

6. Experiments and Results

6.1. Evaluation Parameters. The performance of supervised ML algorithm can be evaluated based on the term or elements of confusion matrix on a set of test data. The confusion matrix consists of four terms, namely, True Positive (TP), False Positive (FP), True Negative (TN), and False Negative (FN). According to the value of these elements, the evaluation matrices precision, recall, and accuracy are determined to estimate the performance score of any classifier.

$$\textbf{Precision } (\boldsymbol{\pi}): \quad \frac{\boldsymbol{TP}}{\boldsymbol{TP} + \boldsymbol{FP}} \tag{18}$$

$$\textbf{Recall } (\boldsymbol{\rho}): \quad \frac{\boldsymbol{TP}}{\boldsymbol{TP} + \boldsymbol{FN}} \tag{19}$$

$$\textbf{F-Score:} \quad \frac{\boldsymbol{2 * Precision * Recall}}{\boldsymbol{Precision + Recall}} \tag{20}$$

6.2. Results and Discussion. The following experimental results help in the study of the effects of an individual as well as a combination of the different feature selection methods on the performance of the classifier. This result clearly shows how each classifier behaves with different feature selection methods. In this section, an in-depth investigation was carried out to measure the effectiveness of the proposed approach, i.e., to compare the performance of the four supervised classifiers SVM, MNB, KNN, and ME based on the combination of the different feature selection methods.

Tables 4–6 display the performance of machine learning methods SVM, MNB, KNN, and ME with respect to different feature selection methods. The method IG performed well in comparison with other FSMs.

The result for movie review dataset in Table 4 indicates that the ComopIG and CompGI are some good choices among various feature selection methods. The composite feature set CompUniBi (unigram + bigram) provides very convincing results with IG and GI method, while the result of unigram and bigram features is not able to give a satisfactory output individually.

While comparing the performance of classification algorithms in present work, SVM produces the best result with a movie as well as a kitchen dataset. The highest F-score obtained by SVM 90.39 with CompGI method is represented in Table 4. The resulting values demonstrate that feature selection method CompIG also performed well with SVM

Table 4: Results of the proposed model for movie review data set.

Method	Classifier											
	SVM			MNB			KNN			ME		
	Prec	Recall	F-Score	Prec	Recall	F -Score	Prec	Recall	F –Score	Prec	Recall	F-Score
Unigram	86.6	82.4	84.44	85.7	82.4	84.01	84.1	80.2	82.10	84.7	83.1	83.89
Bigram	87.4	84.8	86.08	82.1	80.7	81.39	83.0	80.8	81.88	82.4	80.0	81.18
Unigram+ Bigram	86.2	83.2	84.67	86.2	83.4	84.77	78.6	76.1	77.32	86.6	82.8	84.65
CompIG	92.5	88.1	90.24	88.9	87.2	**88.04**	87.4	83.6	**83.45**	84.2	87.1	85.62
CompCHI	90.0	87.3	88.62	86.6	84.7	85.63	86.2	82.4	80.25	87.6	84.5	86.02
CompGI	91.2	89.6	**90.39**	88.0	84.5	86.21	87.3	84.9	79.08	85.9	88.4	**87.13**
Average	89.4	86.4	87.86	86.25	83.81	84.69	84.43	81.33	82.84	85.23	84.31	84.74

Table 5: Results of the proposed model for electronics review data set.

Method	Classifier											
	SVM			MNB			KNN			ME		
	Prec	Recall	F- Score	Prec	Recall	F-Score	Prec	Recall	F-Score	Prec	Recall	F-Score
Unigram	86.9	83.2	85.0	85.7	86.6	86.14	82.3	80.1	81.18	82.1	79.8	80.93
Bigram	81.5	80.1	80.79	82.5	80.2	81.33	84.5	81.2	82.81	78.2	76.2	77.18
Unigram+ Bigram	87.2	84.4	85.77	85.2	87.4	86.28	86.2	82.7	80.41	83.1	80.4	81.72
CompIG	88.3	87.1	**87.69**	89.2	86.0	**87.96**	87.7	84.5	**85.07**	85.7	81.9	83.75
CompCHI	86.7	82.5	84.54	87.9	85.4	86.63	83.9	81.2	82.52	86.2	83.6	84.88
CompGI	88.1	84.8	86.41	88.2	85.7	86.93	88.6	83.4	84.92	87.3	85.5	**86.39**
Average	86.6	83.5	85.01	86.3	85.4	86.33	85.53	82.18	83.81	83.76	82.10	84.75

Table 6: Results of the proposed model for kitchenware review data set.

Method	Classifier											
	SVM			MNB			K NN			ME		
	Prec	Recall	F-Score	Prec	Recall	F-Score	Prec	Recall	F-Score	Prec	Recall	F-Score
Unigram	85.4	80.2	82.71	82.1	80.7	81.39	83.9	81.4	82.63	84.0	81.2	82.57
Bigram	85.6	84.1	84.84	81.4	78.8	80.07	82.5	78.9	80.65	81.2	83.4	82.28
Unigram+ Bigram	88.7	86.2	**87.43**	86.3	84.9	**85.59**	84.6	83.2	83.89	83.3	84.5	83.89
ComopIG	87.9	85.6	**86.73**	86.8	83.1	84.86	84.7	82.6	83.63	85.1	83.7	84.39
CompCHI	86.8	85.9	85.34	84.0	82.5	83.24	81.1	83.2	82.13	84.4	85.5	84.94
CompGI	83.2	82.10	83.13	86.2	83.8	84.98	85.4	87.7	**86.53**	85.2	87.2	**86.18**
Average	85.75	83.11	85.40	84.46	82.3	83.35	83.84	81.18	82.48	83.86	84.25	84.04

classifier for all three review datasets, but the results of using CHI with SVM are not impressive.

According to Table 5, the classifiers SVM, MNB, and ME provide quite impressive results with 10-fold cross validation technique and the maximum accuracy 87.96 got by MNB classifier for electronics review dataset. The MNB performs surprisingly well for sentiment analysis in many previous studies. NB method is a simple and popular classification technique, although the conditional independence assumption is harsh. However, in our investigation, MNB is next best to SVM in performance for movie review dataset with F-score 88.04. In all three datasets, MNB classifier maintained consistently high performance throughout the whole work.

As reported in Tables 4–6, the F-score value obtained using combination (unigram +bigram) is comparatively better than that obtained using unigram or bigram individually. If we consider the results of kitchen review dataset based on Table 6, the F-score for combined feature list of unigram and bigram increased from 82.71 to 87.43.

The feature selection method IG and Gini Index with composite feature set produce the best classification results with more or less every classifier because they eliminate the irrelevant and noisy features at primary stage and consider only top-ranked features. They chose the features based on their importance to the class level attribute. The best performance of the KNN classifier is achieved with review

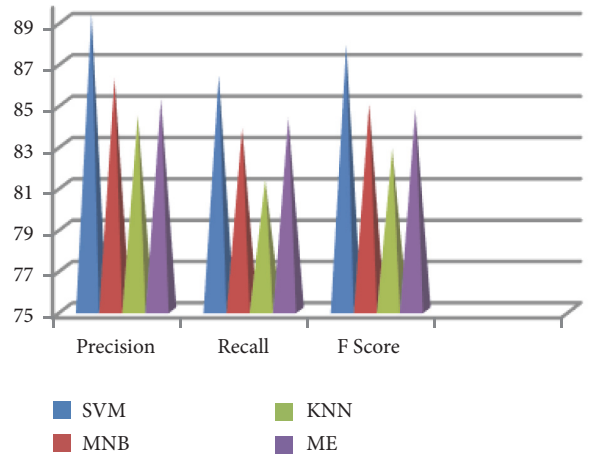

Figure 2: Comparison of classifier performance for **movie review** dataset.

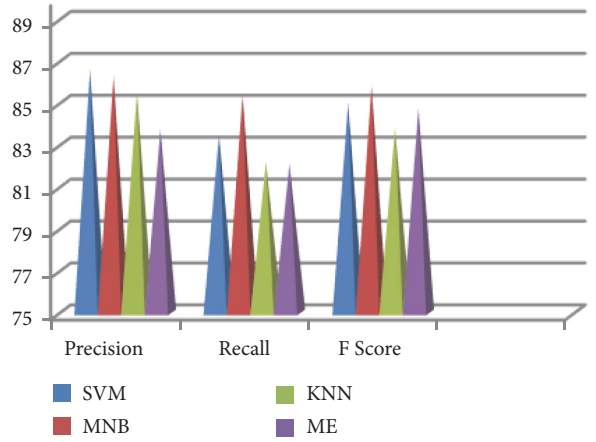

Figure 3: Comparison of classifier performance for **electronics product review** dataset.

dataset of kitchenware (86.3%) when the Gini Index method is being used.

ME classifier got the maximum F-score of 87.13 for Chi-square method. When we consider the domain electronics and kitchen, the F-score for ME classifier reduced to 86.39 and 86.18, respectively.

In order to investigate Figures 2–4, if we compare the classifiers performance, SVM outperforms the other classification methods, MNB, KNN, and ME. According to the average value of precision, recall, and F-score value, we estimate the results of three algorithms on testing dataset. The highest average value 87.86 is portrayed in Figure 2 for movie review dataset. Figure 3 indicates that MNB classifier secured the minimum average value 86.33 for electronics database. According to Figure 4, the resulting value 85.40 as an utmost average score is obtained by classifier SVM for kitchen review dataset.

6.3. Performance Evaluation. This section compares the accuracy of proposed approach with other existing

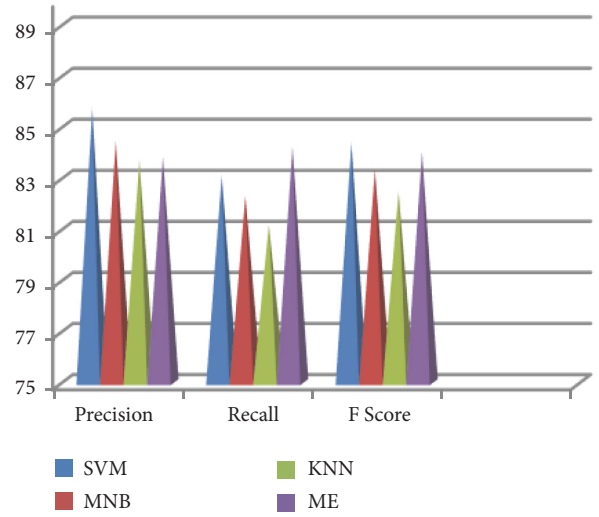

Figure 4: Comparison of classifier performance for kitchen review dataset.

approaches considered such as IMDb dataset. This comparison was carried out according to the accuracy value that these methods achieved. The adopted approach, i.e., the combination of different feature selection methods, produces a better result in comparison with the result obtained by applying individual feature selection method in previous research approaches shown in Table 7.

In an experimental study performed by Pang et. al. [20] on sentiment analysis, they have used SVM, NB, and ME classifier with n-gram technique of unigram and bigram as well as their combination on movie review database (IMDb). They got the accuracy of 82.7, 81.2, and 81.0 for the classifiers SVM, NB, and ME, respectively.

Agarwal et al. [29] have proposed a hybrid method combining rough set theory and Information Gain for sentiment classification. These methods are evaluated on four standard datasets such as movie review (IMDb) and product (book, DVD, and electronics) review dataset. SVM and NB classifiers are used with 10-fold cross validation for classifying sentiment polarity of review documents. F1-measure value is considered as a performance measure with maximum 87.7 and 80.9 for SVM and NB classifier.

Kalaivani et al. [30] examined how classifiers SVM, NB, and KNN work with different feature sizes of movie review dataset. Feature selection method Information Gain (IG) was applied to select top p% ranked features to train the classifier. In this work, SVM approach outperformed the Naive Bayes and KNN approaches with highest accuracy of 81.71. The experimental result reported the precision and recalled the value for positive and negative corpus separately.

In [31], the investigation by Tripathy et al. employed machine ML classifiers, namely, NB, SVM, ME, and SGD, to perform sentiment classification of online movie reviews [23] with n-gram techniques. The performance evaluation can be done by parameters such as precision, recall, F-measure, and accuracy.

Table 7: Comparison of performance of proposed approach with different literature using movie review dataset.

	Dataset	Feature Selection Method	Classifier	Performance
Pang et al.	Internet movie database (IMDb)	N-gram features	SVM NB ME	82.9 (Accuracy) 81.5 81.0
Agarwal et al.	Movie (IMDb) Product (book, DVD, Electronics)	N-gram, IG,RSAR, Hybrid (IG+RSAR)	SVM NB	87.7 (F-measure) 80.9
Al-Moslmi et al.	Movie Reviews in the Malay Language	IG, CHI, Gini Index	SVM NB KNN	85.33(F-measure) 80.88 74.68
Kalaivani et al.	Movie (IMDb)	----------	SVM NB KNN	81.71 72.55 98.70
Tripathy et al.	Movie (IMDb)	N-Gram features	SVM ME NB SGD	88.94 88.48 86.23 85.11
Our Approach	Movie (IMDb)	N-gram, Combination of Unigram & bigram with IG, CHI, Gini Index	SVM MNB KNN ME	90.39 (F-measure) 88.04 86.03 87.13

The results in comparing with our approach show that FSMs have a great impact on the classification performance. The feature ranking techniques (Information Gain, Chi-square, and Gini Index method) improve classification performance over no feature selection.

Al-Moslmi et al. [32] studied feature selection methods effects on machine learning approaches in Malay sentiment analysis. It was demonstrated that improved feature selections resulted in better performance in Malay sentiment-based classification. The author approached three feature selection methods (IG, Gini Index, and CHI) to enhance the performance of three machine learning classifiers (SVM, NB, and KNN). A dataset of 2000 movie reviews is crawled from several web contents in Malay language. The results showed that the combination of SVM classifier and IG-base method established the best classification algorithm, with an accuracy of 85.33% and feature size of 300. Authors have also reported that use of the FSMs yields improved results compared to those from the original classifier.

7. Conclusion

Sentiment analysis is one of the most challenging fields involved with natural language processing. It has a wide range of applications like marketing, politics, and news analytics, and all these areas benefit from the result of sentiment analysis.

The aim of this paper is to explore the ability of statistical feature selection methods such as IG, Chi-square, and Gini Index to improve the classification performance of four machine learning algorithms SVM, MNB, ME, and KNN for sentiment classification. First, we applied n-gram (unigram, bigram) method on noise free preprocessed dataset and obtained a combined feature set as CompUniBi, fed to the feature selection methods IG, CHI, and GI to get an optimal feature subset. These methods offer a ranking to the features depending on their score; thus a prominent feature vector (CompIG, CompGI, CompCHI) can be generated easily for classification. Finally, the classifiers SVM, MNB, KNN, and ME machine learning used prominent feature vector for classifying the review document into either positive or negative.

The performance of sentiment analysis is evaluated on three different domain datasets: movie, electronics, and kitchen review, and the effectiveness of classification algorithm is estimated in terms of F-measure, precision, and recall. As discussed in Section 6.2. The composite feature set of unigram and bigram produce very convincing results. Specifically, it is clear that SVM performed better in terms of higher accuracy (90.24) than MNB, ME, and NN on composite IG (CompIG) feature vector, while MNB classifier delivers performance 88.04 when used with fewer features. These empirical experiments show that the proposed method is highly effective and encouraging.

In the future, our aim is to improve the performance of sentiment classification by expanding the amount of experimental data. We are also planning for future to merge the traditional machine learning method with deep learning techniques to tackle the challenge of sentiment prediction of massive amount of unsupervised product review dataset.

Data Availability

The movie review dataset https://www.kaggle.com/nltkdata/movie-review#movie_reviews.zip is one of the popular benchmark datasets used in our research work. We

adopted the dataset http://www.cs.jhu.edu/~mdredze/datasets/sentiment/processed_acl.tar.gz of electronics and kitchen domain from the corpus produced by Blitzer et al.

Conflicts of Interest

The authors declare that they have no conflicts of interest.

References

[1] B. Liu, "Sentiment analysis and subjectivity," in *Invited Chapter for the Handbook of Natural Language Processing*, N. Indurkhya and F. J. Damerau, Eds., 2nd edition, 2010.

[2] K. Dave, S. Lawrence, and D. M. Pennock, "Mining the peanut gallery: opinion extraction and semantic classification of product reviews," in *Proceedings of the 12th International Conference on World Wide Web (WWW '03)*, pp. 519–528, Budapest, Hungary, May 2003.

[3] M. Annett and G. A. Kondrak, "Comparison of sentiment analysis techniques: Polarizing movie blogs," *Advances in Artificial Intelligence*, pp. 25–35, 2008.

[4] A. Tripathy, A. Agrawal, and S. K. Rath, "Classification of sentimental reviews using machine learning techniques," *Procedia Computer Science*, vol. 57, pp. 821–829, 2015.

[5] D. Zhang, H. Xu, Z. Su, and Y. Xu, "Chinese comments sentiment classification based on word2vec and SVMperf," *Expert Systems with Applications*, vol. 42, no. 4, pp. 1857–1863, 2015.

[6] K. Mouthami, K. N. Devi, and V. M. Bhaskaran, "Sentiment analysis and classification based on textual reviews," in *Proceedings of the 2013 International Conference on Information Communication and Embedded Systems, ICICES 2013*, pp. 271–276, India, February 2013.

[7] H.-Y. Wang and S.-M. Chen, "Evaluating students' answerscripts using fuzzy numbers associated with degrees of confidence," *IEEE Transactions on Fuzzy Systems*, vol. 16, no. 2, pp. 403–415, 2008.

[8] S.-M. Chen, A. Munif, G.-S. Chen, H.-C. Liu, and B.-C. Kuo, "Fuzzy risk analysis based on ranking generalized fuzzy numbers with different left heights and right heights," *Expert Systems with Applications*, vol. 39, no. 7, pp. 6320–6334, 2012.

[9] S.-M. Chen, Y.-C. Chang, and J.-S. Pan, "Fuzzy rules interpolation for sparse fuzzy rule-based systems based on interval type-2 gaussian fuzzy sets and genetic algorithms," *IEEE Transactions on Fuzzy Systems*, vol. 21, no. 3, pp. 412–425, 2013.

[10] Q. Ye, Z. Zhang, and R. Law, "Sentiment classification of online reviews to travel destinations by supervised machine learning approaches," *Expert Systems with Applications*, vol. 36, no. 3, pp. 6527–6535, 2009.

[11] V. Narayanan, I. Arora, and A. Bhatia, "Fast and accurate sentiment classification using an enhanced naïve Bayes model," in *Intelligent Data Engineering and Automated Learning – IDEAL 2013*, vol. 8206 of *Lecture Notes in Computer Science*, pp. 194–201, Springer, Berlin, Heidelberg, Germany, 2013.

[12] A. Amolik, N. Jivane, M. Bhandary, and M. Venkatesan, "Twitter sentiment analysis of movie reviews using machine learning technique," *International Journal of Engineering and Technology*, pp. 2038–2044, 2016.

[13] B. Luo, J. Zeng, and J. Duan, "Emotion space model for classifying opinions in stock message board," *Expert Systems with Applications*, vol. 44, pp. 138–146, 2016.

[14] C. Selvi, C. Ahuja, and E. Sivasankar, "A comparative study of feature selection and machine learning methods for sentiment classification on movie data set," in *Intelligent Computing and Applications*, D. Mandal, R. Kar, S. Das, and B. K. Panigrahi, Eds., pp. 367–379, Springer, India, 2015.

[15] I. Habernal, T. Ptáček, and J. Steinberger, "Supervised sentiment analysis in Czech social media," *Information Processing & Management*, vol. 50, no. 5, pp. 693–707, 2014.

[16] B. Pang, L. Lee, and S. Vaithyanathan, "Thumbs up? sentiment classification using machine learning techniques," in *Proceedings of the Conference on Empirical Methods in Natural Language Processing (ACL '02)*, pp. 79–86, Stroudsburg, Pa, USA, 2002.

[17] Z.-J. Zha, J. Yu, J. Tang, M. Wang, and T.-S. Chua, "Product aspect ranking and its applications," *IEEE Transactions on Knowledge and Data Engineering*, vol. 26, no. 5, pp. 1211–1224, 2014.

[18] M. Ghosh and G. Sanyal, "Preprocessing and feature selection approach for efficient sentiment analysis on product reviews," in *Proceedings of the 5th International Conference on Frontiers in Intelligent Computing: Theory and Applications*, S. C. Satapathy, Ed., AISC 515, Springer, India, 2016.

[19] V. Hatzivassiloglou and K. R. McKeown, "Predicting the semantic orientation of adjectives," in *Proceedings of the 8th conference on European chapter of the Association for Computational Linguistics*, pp. 174–181, Madrid, Spain, July 1997.

[20] B. Pang and L. Lee, "A sentimental education: sentiment analysis using subjectivity summarization based on minimum cuts," in *Proceedings of the 42nd Annual Meeting on Association for Computational Linguistics (ACL'04)*, 278, 271 pages, Association for Computational Linguistics, Barcelona, Spain, July 2004.

[21] Y. Yang and J. Pedersen, "A comparative study on feature selection in text categorization," in *Proceedings of the 14th International Conference on Machine Learning (ICML'97)*, 1997.

[22] W. Shang, H. Huang, H. Zhu, Y. Lin, Y. Qu, and Z. Wang, "A novel feature selection algorithm for text categorization," *Expert Systems with Applications*, vol. 33, no. 1, pp. 1–5, 2007.

[23] H. Park, S. Kwon, and H.-C. Kwon, "Complete gini-index text (git) feature-selection algorithm for text classification," in *Proceedings of the 2nd International Conference on Software Engineering and Data Mining (SEDM '10)*, pp. 366–371, China, June 2010.

[24] A. M. Kibriya, E. Frank, B. Pfahringer, and G. Holmes, "Multinomial naïve bayes for text categorization revised," in *AI 2004: Advances in Artificial Intelligence*, vol. 3339 of *Lecture Notes in Computer Science*, pp. 488–499, Springer, 2004.

[25] C. W. Hsu, C. C. Chang, and C. J. Lin, *A practical guide to support vector classification*, Simon Fraser University, 8888 University Drive, Burnaby BC, Canada, 2005.

[26] T. Joachims, "Text categorization with support vector machines: learning with many relevant features," in *Proceedings of the European Conference on Machine Learning (ECML'98)*, vol. 1398, pp. 137–142, Springer, 1998.

[27] Y. Lihua, D. Qi, and G. Yanjun, "Study on KNN text categorization algorithm," *Micro Computer Information*, vol. 21, no. 2006, pp. 269–271, 2006.

[28] K. Nigam, J. Lafferty, and A. McCallum, "Using maximum entropy for text classification," in *IJCAI-99 Workshop on Machine Learning for Information Filtering*, vol. 1, pp. 61–67, 1999.

[29] B. Agarwal and N. Mittal, "Sentiment classification using rough set based hybrid feature selection," in *Proceedings of*

the 4th Workshop on Computational Approaches to Subjectivity, Sentiment and Social Media Analysis, pp. 115–119, Atlanta, 2013.

[30] P. Kalaivani and K. L. Shunmuganathan, "Sentiment classification of movie reviews by supervised machine learning approaches," *Indian Journal of Computer Science and Engineering (IJCSE)*, vol. 4, pp. 286–292, 2013.

[31] A. Tripathy, A. Agrawal, and S. K. Rath, "Classification of sentiment reviews using n-gram machine learning approach," *Expert Systems with Applications*, vol. 57, pp. 117–126, 2016.

[32] T. Al-Moslmi, S. Gaber, A. Al-Shabi, M. Albared, and N. Omar, "Feature selection methods effects on machine learning approaches in malay sentiment analysis," in *Proceedings of the 1st ICRIL International Conference on Innovation in Science and Technology (lICIST '15)*, pp. 444–447, 2015.