Computers & Industrial Engineering 124 (2018) 139–156

Contents lists available at ScienceDirect

Computers & Industrial Engineering

journal homepage: www.elsevier.com/locate/caie

Scaling feature selection method for enhancing the classification performance of Support Vector Machines in text mining

S. Manochandar, M. Punniyamoorthy*

Department of Management Studies, National Institute of Technology, Tiruchirappalli, India

ARTICLE INFO

Keywords:
Glasgow
Opinion mining
Support vector machine
Term weighting
Tf-idf

ABSTRACT

The classification of opinion based on customer reviews is a complex process owing to high dimensionality. In this study, our objective is to select the minimum number of features to effectively classify reviews. The tf-idf and Glasgow methods are commonly for feature selection in opinion mining. We propose two modifications to the traditional tf-idf and Glasgow expressions using graphical representations to reduce the size of the feature set. The accuracy of the proposed expressions is established through the support vector machine technique. In addition, a new framework is devised to measure the effectiveness of the term weighting expressions adopted for feature selection. Finally, the strength of the expressions is established through evaluation criteria and effectiveness, and this strength is tested statistically. Based on our experimental results, our modified tf-idf and Glasgow methods performed better than the traditional term weighting expressions for the extraction of the minimum number of prominent features required for classification, thus enhancing the performance of the Support Vector Machine.

1. Introduction

Opinion mining, which is also known by other terms, such as sentiment analysis and subjective analysis, is a subfield of Text Mining (Claypo and Jaiyen, 2014). In recent times, owing to the proliferation of forums, blogs, e-commerce portals, websites, news reports, and additional web resources, where people tend to post their opinion, opinion mining has attracted significant interest from researchers (Pang and Lee, 2004; Wu, Wang, Li, & Long, 2014); in particular, online reviews by customers about products and services on the abovementioned sites have had a significant impact on the traditional decision making process (Pepin, Kuntz, Blanchard, Guillet, & Suignard, 2017; Bag, Tiwari, & Chan., 2017). Therefore, opinion mining helps managers better understand and explore the opinions of their customers regarding products or services (Pepin et al., 2017). In fact, individuals and companies are always interested in others' opinions for various reasons. For example, when an individual wants to purchase a new product, he/she might first obtain opinions from buyers who have previously purchased the product, and then, based on those reviews, he/she will take a decision whether to purchase the product. In the same vein, companies selling products and services also pay attention to the consumer reviews about their offerings. These reviews possess critical information on customer concerns and their experience, which can be useful for conceptual design, personalization, product recommendation, better customer understanding, and customer acquisition (Zhan, 2009). Furthermore, in another study, it was observed that, compared with the traditional customer survey methods, web-based reviews (i.e., opinion mining) have the advantages of easy accessibility and low costs (Wu et al., 2014).

In particular, opinion mining is the process of recognizing whether the opinion among the target group about a given topic expressed in a document is positive or negative. Opinion mining is typically used to classify reviews as either positive or negative reviews based on word counts (Allen, Sui, & Parker, 2017). The negative reviews (i.e., customer complaints) convey important information about the reason for the dissatisfaction of customers with products/services (Lee, Wang, & Trappey, 2015). Consequently, opinion mining helps in customer retention (Lee et al., 2015; Bag, Tiwari, & Chan, 2017). Not only is it difficult to read all the opinions across different forums or blogs, but also it is challenging for companies to manage the high number of reviews to evaluate their products or services. Therefore, it is necessary to develop methods that can improve the accuracy of evaluation based on such unstructured reviews (Rushdi Saleh, Martin-Valdivia, Montejo-Raez, & Urena- Lopez, 2011).

Abbreviations: VSM, vector space model; FS, feature selection; SVM, support vector machine; tf, term frequency; idf, inverse document frequency; EC, evaluation criteria; TW, term weight

* Corresponding author.
E-mail address: punniya@nitt.edu (M. Punniyamoorthy).

https://doi.org/10.1016/j.cie.2018.07.008
Received 5 June 2017; Received in revised form 29 March 2018; Accepted 7 July 2018
Available online 09 July 2018
0360-8352/

Many potential sentiment analysis applications are not feasible because of the voluminous number of features. Text data is typically represented using the VSM. After the pre-processing stage, each document/review is transformed into a feature vector, while each word is considered a dimension in the feature space. In general, documents are a sequence of words, which can be represented by "*d*" (Allen, Xiong, & Afful-Dadzie, 2016); because a document contains several distinct words, there are several text vector space dimensions, which are sparse as well. The high-dimensional and sparse features lead to considerable noise that negatively affects classification (Zhao, Zhang, & Wan, 2013). Further, in opinion mining, documents are identified by sets of terms or keywords that are collectively used to represent their contents (Zhang, Yoshida, & Tang, 2011). FS is a process in which a subset of the features available from the data is selected for application in a learning algorithm. The accuracy of a learning algorithm is strongly dependent on the features that are used to train the system.

The FS method differs across different domains (Rushdi Saleh et al., 2011). For example, the FS for a corpus of movie reviews is different from that of electronic product reviews. Moreover, training in the case of machine learning techniques using high-dimensional features is more difficult task; therefore, the selection of prominent features using an effective strategy is important for opinion mining. Tf-idf is one of the earliest weighting schemas that is used for FS; it reflects the importance of a term in a document collection (Salton, & Buckley, 1988; Zhang et al., 2011; Zhao et al., 2013). Tf-idf is based on the theory of language modeling, in which the terms in a given document (*d*) can be divided into two categories: words with and without eliteness (Zhang et al., 2011).

In the case of idf is that low-frequency terms are more informative compared to high-frequency terms, especially when the stopwords are included as features. Further, the number of features is selected in a subjective manner using the tf-idf and Glasgow expressions; in some of the cases, machine learning techniques are used to select the features using prior information available for the class labels. Considering the abovementioned shortcoming, the stopwords are removed. After removal of these stopwords, high-frequency terms are more informative than the low-frequency terms. Thus, accordingly traditional tf-idf and Glasgow expressions are modified, with a new graphical technique to select the prominent features. A new framework is also employed to evaluate the effectiveness of the proposed expressions. The accuracy of the proposed expressions is established through SVM classification. The evaluation criteria and effectiveness are analyzed to evaluate the robustness of the proposed expressions. Experiments were performed using benchmark and primary datasets; then, the obtained results were compared with those obtained using the traditional methods.

The remainder of the paper is organized as follows. Related literature is discussed in Section 2; the proposed methods are presented in Section 3. The experimental results and findings are presented in Section 4; Finally, Section 5 concludes the paper.

2. Literature review

This section discusses the previous works related to FS in two parts: (i) those related to the usage of tf-idf and other FS techniques independently, and (ii) those related to the usage of hybridization techniques.

2.1. Literature related to the independent use of tf-idf and other FS techniques

tf is the frequency of a term in a document. Because of different document lengths, typically, the frequency of term occurrence is high in a long document than in a short one. Therefore, in order to normalize this term, the term frequency is often divided by the document length, i.e.,

$$tf = \frac{fr_{td}}{length_d} \tag{1}$$

Further, the idf, which was proposed by Jones (1972), measures the importance of a term in the corpus, where corpus refers to a collection of documents (Allen et al., 2016); this idf provides lesser weightage for high frequency terms and vice versa. The idf is given by the following formula:

$$idf = log\left(\frac{N}{n_j}\right) \tag{2}$$

Tf-idf represents the product of tf and idf. Initially, this term weighting method was proposed for information retrieval (Salton, & Buckley, 1988; Yoon, Seo, Coh, Song, & Lee, 2017); however, later it was used for the FS process. The Glasgow weighting scheme was proposed by Sanderson, & Ruthven (1996) to avoid long documents with a considerable number of irrelevant words than in the small documents (Sabbah, Selamat, Selamat, Ibrahim, & Fujita, 2015; Thangairulappan et al., 2016). The Glasgow formula is as follows:

$$Glasgow = \frac{log\,(fr_{td} + 1)}{log\,(length_d)} * log\left(\frac{N}{n_j} + 1\right) \tag{3}$$

A relevancy weight was employed by Wiener, Pedersen, and Weigend (1995) for FS; thereafter, this relevancy weight is multiplied by the idf weight for each term. Further, a new FS metric called correlation coefficient was introduced by Ng, Goh, and Low (1997). In addition, a comparative study on FS was conducted by Yang & Pedersen (1997), in which five methods were compared, including the document frequency, information gain, mutual information, χ^2, and term strength methods. A new FS method based on χ^2 statistics was proposed by Galavotti, Sebastiani, and Simi (2000). Furthermore, an iterative FS method based on Term Contribution in which only the document frequency is considered for the FS was proposed by Liu, Liu, Chen, and Ma (2003). Lan et al. (2006) proposed a tf-relevance frequency term weighting method for FS. The methods are validated through unsupervised and supervised learning. The author compared the results with those of the χ^2, odd ratio, and information gain methods. Furthermore, a method involving the combination of tf with extended document frequency for FS was proposed by Xu, Wang, Li, Jing, and (2008); this method was evaluated through k-nearest-neighbor and Naïve Bayes classifier. Liu, Loh, and Sun (2009) proposed the tf feature value and probability-based term weighting schemes directly utilizes two critical information ratios (relevance indicators); the performance of their schemes was evaluated through SVMs and naive Bayes classifiers on two benchmark datasets: MCV1 and Reuters-21578. Eirinaki, Pisal, and Singh (2012) proposed the high adjective count algorithm for FS; this method is used to extract the features that contribute to categorization and can be used to extract important aspects from reviews. The results of the high adjective count method were compared with the traditional tf and tf-idf methods. Naderalvojoud et al. (2014) insisted that one sided and two-sided FS metrics (tffs) are based on the term weighting expressions; it is noteworthy that one sided metrics incorporate only the positive features, whereas the two-sided metrics incorporate both the relevant and non-relevant terms for a category. Based on the experimental results, the authors recommended an SVM for the one-sided FS metrics. Yi, Yang, and Wan (2016) suggested FS based on category discrimination. Category discrimination is the product of inter-category dispersion and intra-category information entropy. The SVM method was adopted to validate this FS method; the micro and macro-*F*1 values were 84.12% and 79.81% for feature sizes of 14,000 and 16,000, respectively. Sabbah et al. (2017) proposed the modified tf and idf schemes for feature extraction, referred to as mTF and mIDF, respectively. In the mTF, the total token count is introduced in a manner similar to that of mIDF, and the difference between the total number of documents in the collection and the number of

documents, in the term "*t*" is used in the mIDF. The effectiveness of this method was tested on three benchmark datasets, including Recurters-21578, Classic4, and WebKB. The three data SVMs provided better results for all the three datasets; the micro-averaged *F*-measures for Recurters-21578, Classic4, and WebKB were as follows: 97%, 75%, and 94% for mTFmIDF, mTFIDF, and mTFmIDF, respectively.

2.2. Literature related to the use of hybridization techniques

Yu, & Jiang (2004) combined the tf-idf and mutual information methods for the FS process. Luo, Wu, and Yang (2006) used the χ^2 and tf methods together for FS to overcome the deficiency of the χ^2 method, which ignores the word frequency of the feature word. Bharti and Singh (2014) developed the three-stage unsupervised dimension reduction method; in their method, mean absolute difference, mean-median, and absolute cosine are used for FS and principal component analysis is performed for feature extraction. Zong, Wu, Chu, and Sculli (2015) proposed the discriminative and semantic similarity FS method; the validation of their method was conducted using an SVM and the results obtained for Reuters-21578 are 92.20% and 96.82% for *F*-macro and *F*-micro, respectively. In addition, they were 77.17% and 78.22% for the 20 Newsgroup, respectively. Bharti and Singh (2015) proposed the modified union approach for FS; in their study, term variance, document frequency, and principal component analysis were adopted for FS. Sabbah et al. (2015) proposed the hybridized term weighting method for web contents classification using SVM. The hybridized method combined feature sets generated by the term weighting schemes such as tf, document frequency, tf-idf, Glasgow, and Entropy schemes into one feature set for classification. The classification techniques such as SVM, K-nearest neighbor, decision tree, naive Bayes, and ELM were adopted and it was shown that SVM outperforms all the other classifiers. A maximum accuracy of 93.60% was achieved by the SVM method for 1094 features. Furthermore, Ghareb, Abu Bakar, and Hamdan, (2016) investigated the six hybridization FS methods, including class discriminating measure GSS (Galavotti, Sebastiani, and Simi (2000)), *F*-measure of training text features, odd ratio, information gain, and tf-idf with enhanced genetic algorithm methods. The SVM and Naive Bayes classification techniques were adopted to test the performance of this hybrid FS. The maximum macro average *F* measure value for tf-idf was attained with Ensemble Genetic Algorithm. Agnihotri, Verma, and Tripathi (2017) proposed a Variable Global Feature Selection Scheme for FS based on the distribution of features in each category. This method is embedded with other FS methods such as Global Feature Selection Scheme (Mutual Information, Information Gain, Gini Index, Distinguishing Feature Selector, and Gain Ratio), Improved Global Feature Selection Scheme and odds ratio. SVM and SOFTMAX regression is used to evaluate the FS methods. Ordinal-based and frequency-based integration of different feature subsets is introduced by Yousefpour, Ibrahim, and Hamed, 2017. Four machine learning techniques (SVM, Naive Bayes, Maximum Entropy and Linear Discriminant Function) are adopted to validate the FS methods. The summary of these works is shown in Table 1.

Based on these previous works in literature, it is clear that tf-idf is a seed expression for FS, whether it is used independently or in combination with other related TW expressions.

When the tf-idf and Glasgow methods were analyzed, it was deduced that the tf provides the importance of a term in the document; however, idf provides less weightage to terms with high occurrence. Therefore, this is a constraint of traditional tf-idf and Glasgow methods. Some of the FS techniques such as odds ratio, GSS, Information Gain, class discriminating measure, Gini Index, tf combined extended document frequency, Discriminative integrated Sematic Similarity, and Category Discrimination methods (Ghareb et al., 2016; Yang & Pedersen, 1997; Yi et al., 2016; Wiener, Pedersen, and Weigend, 1995; Xu et al., 2008; Zong et al. 2015; Agnihotri et al., 2017) require a prior knowledge about the class labels. In hybridization techniques, searching for a best feature set is more complicated (Yousefpour, Ibrahim, and Hamed, 2017), while integration of FS with feature integration leads to higher computational complexity. To overcome these shortcomings, we propose modified expressions with graphical representations to select prominent features.

3. Proposed method

The primary objective of this study is to extract a minimum number of prominent features that effectively discriminate positive and negative reviews. The detailed description of the proposed methods is discussed below. For easy understanding, the description of symbols and notations used in this paper hereafter are listed in Table 2.

3.1. Modified tf-idf and Glasgow methods

Considering document frequency, a basic assumption is that low frequency features are considered non-informative for text categorization, and thus, removal of those low frequency features leads to dimensionality reduction. In information retrieval, it is typically assumed that high frequency features are non-informative; however, these assumptions are not feasible for text categorization. Furthermore, in the case of text categorization, high positive correlations between the document frequency and information gain show that high frequency features are often informative. In addition, it enhances the performance of the classification algorithm. (Yang & Pedersen, 1997; Xu et al., 2008).

Based on the abovementioned assumptions, we suggest modifications to the idf term. Further, tf is high for a frequently used term (*f*), but idf provides minimal weightage to that term; this is not useful in extracting prominent features. After the removal of stopwords, only the TW remains. Considering this, the high value *tf* should have high weightage and low value *tf* should have low weightage.

In general, idf is inversely proportional to the frequency of features. We modify this relationship to render it a directly proportional one by taking the inverse of idf (iidf) weighed with the tf as follows:

$$tf-iidf = tf * \frac{1}{log\left(\frac{N}{n_j}\right)} \tag{4}$$

Tf-df is the refined formulation of the tf-idf to extract the prominent features from the abundance features. Here, the idf is converted by eliminating the logarithm, and the total term occurrence in the documents is divided by total number of documents (i.e., the probability of occurrence of "f_j"). To maximize the TW value, the following formula is used.

$$tf-df = tf * \frac{n_j}{N} \tag{5}$$

The explanation for the probability of features "f_j" is given below.

Let $D = \{d_1, d_2, \ldots d_N\}$, be the number of documents and $F = \{f_1, f_2, \ldots f_J\}$ be the unigram features that are associated through the binary matrix $B = [b_{ij}]$, where $i = 1, 2, \ldots, N$ and $j = 1, 2, \ldots, J$, where,

$$B = \begin{array}{c} \\ d_1 \\ \vdots \\ d_N \end{array} \begin{array}{c} \begin{array}{ccc} f_1 & \ldots & f_J \end{array} \\ \begin{bmatrix} b_{11} & \cdots & b_{1J} \\ \vdots & \ddots & \vdots \\ b_{N1} & \cdots & b_{NJ} \end{bmatrix} \end{array} \tag{6}$$

$$b_{ij} = \begin{cases} 1 & if\ f_j \in d_i \\ 0 & otherwise \end{cases} \tag{7}$$

In terms of the probability of feature (f_j) occurrence in the documents, the probability of the *j*th feature in the training data (*N*) is the calculated as follows (Robertson, 2004),

Table 1
Summary of literature review.

	FS methods	Classification technique	EC	References
tf-idf and other techniques	Relevancy weight	Neural Network	Precision, and Recall	Wiener et al. (1995)
	Correlation coefficient	Perceptron learning	Precision, Recall, and F-measure	Ng et al. (1997)
	Document frequency, Information Gain, Mutual Information, χ^2 and term strength	k-Nearest neighbor Linear Least Square Fit Mapping	Precision, and Recall	Yang and Pedersen (1997)
	χ^2 Stastics	k-Nearest neighbor	Micro-averaged F Measure	Galavotti et al. (2000)
	Term contribution	k-means clustering	Entropy, and precision	Liu et al., (2003)
	Tf-relevance frequency	SVM k-nearest neighbor	Micro F-measure	Lan et al. (2006)
	tf combined with an extended document Frequency	k-nearest-neighbor and Naïve Bayes classifier	Accuracy	Xu et al. (2008)
	TFFV and probability based Weighting	SVM Naive Bayes	Macro-averaged F measure	Liu et al. (2009)
	High Adjective Count	Maximum Opinion Score	Precision	Eirinaki et al. (2012)
	Two sided FS metrics	SVM Naive Bayes Decision Tree k-Nearest Neighbor	F-Measure	Naderalvojoud et al. (2014)
	Tf*Category Discrimination	SVM	Micro and Macro F-Measure	Yi et al. (2016)
	mtf.midf	SVM k-Nearest Neighbor Naive Bayes Extreme Learning Machine	Micro and Macro F Measure	Sabbah et al. (2017)
Hybridization techniques	Tf and χ^2	Naive Bayes	Accuracy	Yu and Jiang (2004)
	Tf and Mutual information	Bayes algorithm	Accuracy	Luo et al. (2006)
	Mean Absolute Deviation, Mean-Median, Absolute Cosine and Principal Component Analysis	k-means clustering	Micro and macro F measure	Bharti and Singh (2014)
	Discriminative and sematic similarity	SVM	Micro and macro F measure	Zong et al. (2015)
	Term variance, document frequency and Principal Component Analysis	k-means clustering	Micro average f measure	Bharti and Singh (2014)
	tf, document frequency, tf-idf, Glasgow and Entropy	SVM k-Nearest Neighbor Decision Tree Naive Bayes Extreme Learning Machine	Micro and Macro Averaged F	Sabbah et al. (2015)
	Ensemble Genetic Algorithm with Combinations of class discriminating measure, GSS, odds ratio, information gain, tf-idf and F-measure	SVM Naïve Bayes	Micro Average F Measure	Ghareb et al. (2016)
	Global feature selection embedded with Mutual Information, information gain, Gini index, Distinguishing Feature Selector and Gain Ration	SVM Softmax regression	Micro and Macro F Measure	Agnihotri et al. (2017)
	Ordinal-based and frequency based integrated uni-gram and parts of speech	SVM Naïve Bayes Maximum Entropy Linear Discriminant Function	Accuracy, Precision, Recall, and F-Measure	Yousefpour et al. (2017)

$$p(f_j) = p(f_j \ occurs \ in \ d_i) = \frac{1}{N}\sum_{i=1}^{N} b_{ij} \approx \frac{n_j}{N} \tag{8}$$

$$n_j = \sum_{i=1}^{N} b_{ij} \tag{9}$$

Similarly, the Glasgow is modified for the idf. The modified Glasgow expressions are as follows:

$$Glasgow{-}I = \frac{log(fr_{td} + 1)}{log(length_d)} * \frac{1}{log\left(\frac{N}{n_j} + 1\right)} \tag{10}$$

$$Glasgow{-}II = \frac{log(fr_{td} + 1)}{log(length_d)} * \left(\frac{n_j}{N}\right) \tag{11}$$

Fig. 1 shows the pictorial representation of the term weighting FS framework presented in this study.

3.2. Phase-I

The unstructured reviews are collected from websites, blogs, forums, and other sources. In this phase, tokenization is performed to divide a stream of text into phrases, words, symbols, or other meaningful terms. After tokenization, filtering is performed to remove stopwords, numbers, and punctuation marks, which do not provide any useful information to classify reviews (Wu et al., 2014; Pepin et al., 2017). The final process in this phase involves reducing the derived words to their base word stem (Claypo and Jaiyen, 2014; Bharti & Singh, 2015); this is a commonly used procedure for text modeling using natural language processing (Allen, Sui, & Parker, 2017).

3.3. Phase-II

In this phase, a proposed graphical technique is used to select the

Table 2
Descriptions of symbols and notations used in this paper.

Symbols and notations	Description
D = {d1, d2, … d_N}	Documents/Review vector
F = {f_1,f_2,…f_J}	Uni-gram feature/term vector
B = [b_{ij}]	Binary occurrence matrix
P(f_j)	Probability of the jth features
N	Total No. of documents/Reviews in the training data
δ	Threshold value
n_j	Frequency of jth features in the training data
TWM = [TW_{NJ}]	Term weighting matrix
φ	Null/Zero Vector
DTM = [b_{kl}]	Document term matrix
PIM = [b_{kl}]	Pre-processed input matrix
ρ	Effectiveness measure
N_f	Total number of features extracted from the reviews
r_f	Number of features removed based on the term weighting methods
r_d	Number of documents removed after extracting the prominent features by Phase-II process
SE(i)	Strength of the expressions (i∈ [1, 2… m], m no. of expressions)
'γ' and 'λ'	Relative weightage values ∈ [0,1]
'α' and 'β'	Weightage values ∈ [0.1, 0.9]
CPM_{EC}(i)	Combined preference measure for evaluation criteria (i∈ [1, 2 … m], m no. of expressions)
RW_ρ(i)	Relative weightage of effectiveness measure (i∈ [1, 2 … m], m no. of expressions)

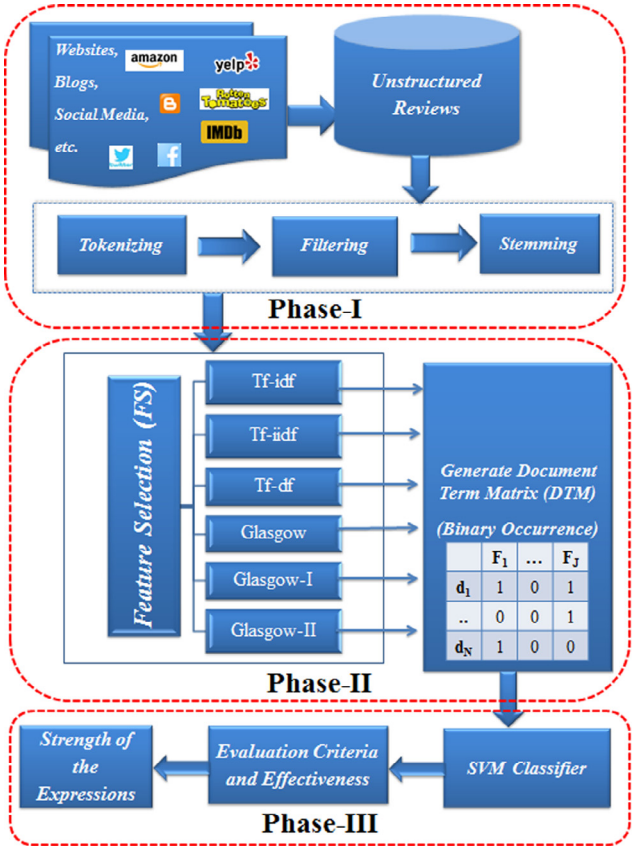

Fig. 1. Term weighting FS framework.

number of features and a cumulative curve for each term weighting expression is plotted as shown below. This procedure is detailed in the following figure.

The different term weighting expressions used to select the number of features is included in this process. A graphical representation comprising the plotting of the cumulative relative frequency of TW against the features is shown in Fig. 3. The number of features is determined using the slope.

When the nature of these curves is analyzed, initially, a steep slope is observed; then, an approximately constant slope is gradually reached. The numbers of features is identified for each method using the point just before the start of the constant slope. The next step involves extracting the prominent features through term weighting. A document term matrix is constructed from these extracted features. This document term matrix comprises reviews and terms and if a term is present in the review, then, it is indicated in the matrix as 1, otherwise 0 (see Fig. 2).

The algorithm for the construction of this pre-processed input matrix for classification is given below:

Algorithm: Pre-processed Input Matrix *(PIM)*

Require: $TWM = \begin{bmatrix} TW_{11} & . & . & TW_{1J} \\ . & TW_{ij} & . & . \\ . & . & . & . \\ TW_{N1} & . & . & TW_{NJ} \end{bmatrix}$, δ = *Threshold*

// where i={1…N} and j={1…J) //

For all f_j in *TWM* **do**

$W_j = \sum_{i=1}^{n} TW_{ij}$ *// W_j is the overall TW of the each f_j //*

End For

Sort (W_j) *// sort W_j in descending order //*

Compute crf_j

Remove $f_j < \delta$

Construct $DTM = \begin{bmatrix} b_{11} & . & . & b_{1m} \\ . & b_{kl} & . & . \\ . & . & . & . \\ b_{N1} & . & . & b_{Nm} \end{bmatrix}$

// where DTM is the N X m matrix ($m < J$), b_{kl} are binary occurrence of the each f_j in the d_i //

If $d_i = \varphi$ then *// where φ is the zero vector //*

Eliminate d_i

Else $\{d_i \neq \varphi\}$

i = 1:N

End if

Construct $PIM = \begin{bmatrix} b_{11} & . & . & b_{1m} \\ . & b_{kl} & . & . \\ . & . & . & . \\ b_{z1} & . & . & b_{zm} \end{bmatrix}$ *//where PIM is z X m matrix, $z \leq N$ //*

TWs are computed for the entire "f_j" present in the "d_i". Then, the TWs are sorted in descending order and their cumulative relative frequencies are computed. The threshold (δ) is set and the values which are less than δ are removed. Then, the document term matrix is constructed. The obtained null documents (φ), containing zero vectors, are identified and eliminated from this matrix; this algorithm is applied for both the traditional and proposed methods.

In general, a high recall indicates that most of the relevant features are retrieved by the FS method, while high precision indicates that the FS returned substantially more relevant features than the irrelevant ones. In Phase-I, stemming and removal of stopwords, numbers, and punctuations is performed to improve recall and precision. Based on the document frequency assumptions, high-frequency features tend to be informative or more relevant features for text categorization. In the proposed method, a higher weight is assigned for high-frequency features leading to extraction of more relevant features, consequently resulting in higher precision. Negation terms such as "no", "not", "never", "nor", "nothing" and "nobody" are treated as features, even if they are removed because they are included in the stopword list, they can indicate a different opinion in the review.

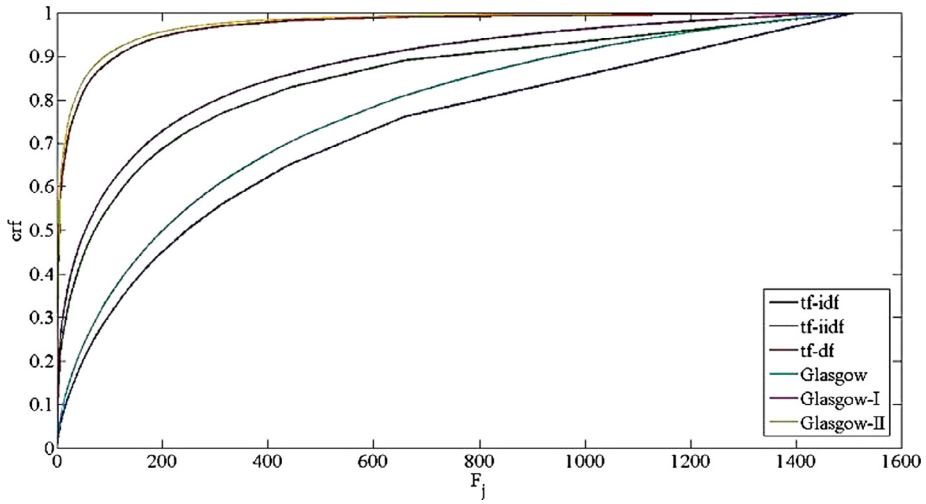

Fig. 3. Graphical representation to select the No. of features.

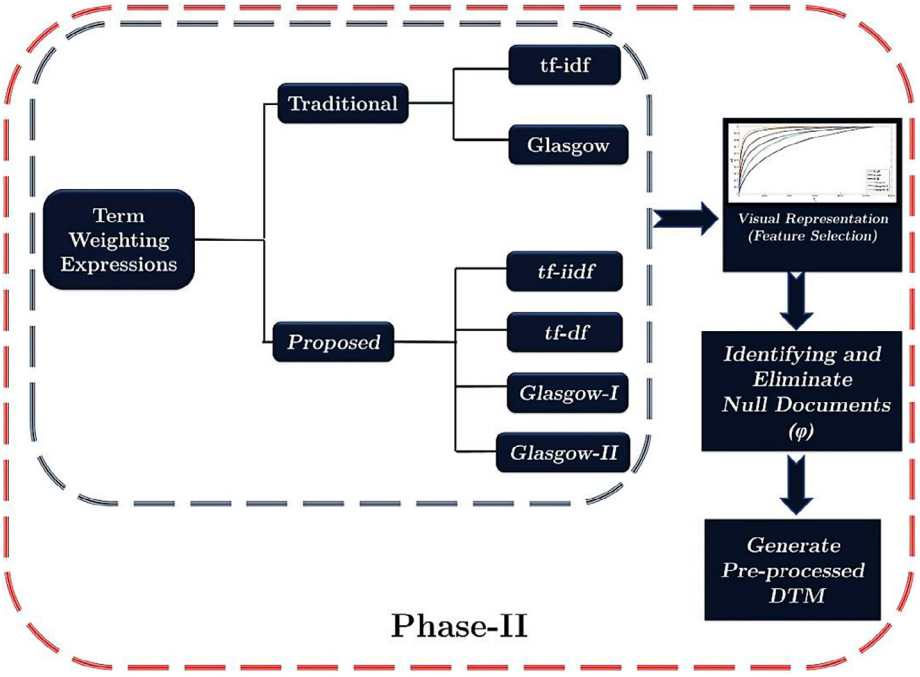

Fig. 2. Flowchart of Phase II.

Table 3
Different kernel functions and default parameter values.

Kernel	Formula	Default parameters value
Linear	$K(x_i, x_j) = (x_i{}'x_j + u)$	u = 1
Polynomial	$K(x_i, x_j) = (\omega x_i{}'x_j + u)^h$	ω = 1, u = 1, h = 2
RBF	$K(x_i, x_j) = \exp\left(-\frac{\|x_i - x_j\|^2}{2\sigma^2}\right)$	σ = 1
Sigmoidal	$K(x_i, x_j) = \tanh(\omega\, x_i{}'x_j + u)$	ω = 1, u = 1

Table 4
Confusion matrix format.

		Predicted	
Actual		+1	−1
	+1	TP	FN
	−1	FP	TN

Table 5
List of EC adopted to evaluate the SVM.

Evaluation criteria	Formula
Accuracy (A)	$\frac{TP+TN}{TP+TN+FP+FN}$
Precision (P)	$\frac{TP}{TP+FP}$
Recall (R)	$\frac{TP}{TP+FN}$
F-measure (F)	$\frac{2*P*R}{P+R}$
True Negative Rate (TNR)	$\frac{TN}{FP+TN}$
False Positive Rate (FPR)	$\frac{FP}{FP+TN}$
False Negative Rate (FNR)	$\frac{FN}{FN+TP}$
False Discovery Rate (FDR)	$\frac{FP}{FP+TP}$
Negative Predictive Value (NPV)	$\frac{TN}{TN+FN}$

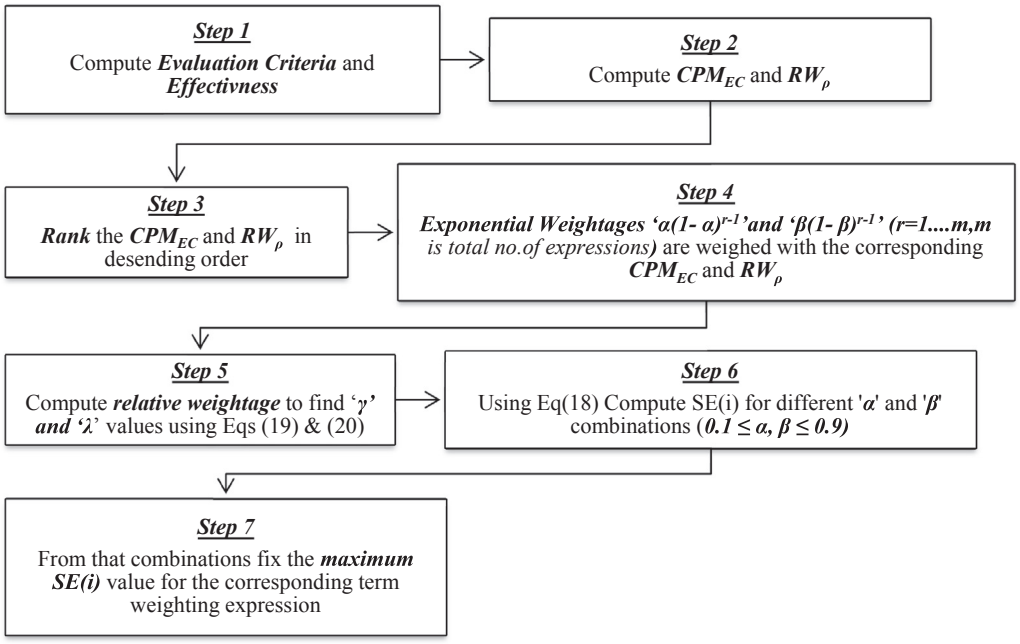

Fig. 4. Flow chart for computing strength of the expression SE(i).

Table 6
Summary of the mobile reviews.

Brand	No. of reviews
Apple iphone 6	130
Apple iphone 7	68
LG G4	22
LG Nexus	450
BlackBerry passport	107
Sony xperia Z3	90
Samsung Galaxy note 5	55
Motorola	78
Total	1000

Table 7
Summary of the Movie reviews.

Movie name	No. of reviews
Conjuring 2	515
Batmen vs Superman	250
X-men Acolypse	1040
Finding Dory	255
Fast and the Furious 8	295

Table 8
Summary of the SemEval reviews.

Domain	No. of reviews
Laptop	3841
Restaurant	3845
Total	7686

3.4. Phase-III

In this section, the SVM methods are discussed that are used to determine the accuracy of the term weighting expressions.

3.4.1. SVM

The supervised machine learning approaches are popular among individuals who analyze documents to predict their overall sentiment. Among these, SVM is one of the most popular algorithm based on the statistical learning theory (Cortes & Vapnik, 1995; Xanthopoulos & Razzaghi, 2014; Kartal, Oztekin, Gunasekaran, & Cebi, 2016) and a kernel function is used to classify data, when the data points are non-linearly separable. SVM is used in many applications because of its robust advantages, including good generalization performance, efficiency in terms of computation time and speed, and robustness in higher dimensions (Lee, Cho, & Asfour, 2011). In addition, the SVM method is known to be an effective and efficient classification technique for text processing (Rushdi Saleh et al., 2011; Zaghloul, 2009; Yang & Liu, 1999; Pang & Lee, 2005; Xanthopoulos & Razzaghi, 2014).

The objective of this phase is to classify each review as a positive or a negative expression about an object. Pang & Lee (2004) compared the different supervised learning algorithms such as Naive Bayes, Maximum Entropy, and SVM methods. Their results show that the SVM algorithm performed better than the others for high feature space. Tan & Zhang (2008) conducted an empirical study on four FS methods and five learning algorithms including the centroid classifier, K-nearest neighbor, winnow classifier, Naive Bayes, and SVM methods. Even in this case, the SVM algorithm exhibited better performance among the mentioned learning algorithms. The effectiveness of the SVM algorithm has also been shown in the case of opinion mining, where it has outperformed other machine learning techniques as well (O'Keefe & Koprinska, 2009; Rushdi Saleh et al., 2011; Zheng, Li, Chen, & Huang, 2006). The SVM algorithm is described as follows.

Let (x_i, y_i), where x_i belongs to R^n be the training data, and y_i is the class label. For a linearly separable space, the separation hyperplane defined by the parameters w and b is computed by solving the convex optimization problem (primal objective) as follows (Xanthopoulos & Razzaghi, 2014; Kartal et al., 2016):

$$\text{Min}\frac{1}{2}||w||^2 \tag{12a}$$

$$s.\ t.\ y_i(w^T x_i + b) \geqslant 1 \quad i = 1, ...n \tag{12b}$$

The decision surface is a hyperplane which can be represented as

$$w^T x_i + b = 0 \tag{13}$$

where 'x_i' are arbitrary objects to be classified, the vector 'w' and constant 'b' are learned from a training set of linearly separable objects.

(a) Mobile

	tf-idf	tf-üdf	tf-df	Glasgow	Glasgow-I	Glasgow-II
tf-idf	1	1**	1**	0.76**	0.81**	0.86**
tf-üdf	1	1	1	0.76	0.81**	0.86**
tf-df	1	1	1	0.76	0.81**	0.86**
Glasgow	0.76	0.76	0.76	1	0.994**	0.97**
Glasgow-I	0.81	0.81	0.81	0.994	1	0.995**
Glasgow-II	0.86	0.86	0.86	0.97	0.995	1

(b) MR1

	tf-idf	tf-üdf	tf-df	Glasgow	Glasgow-I	Glasgow-II
tf-idf	1	0.99**	0.95**	0.74**	0.79**	0.82**
tf-üdf	0.99	1	1**	0.71**	0.81**	0.85**
tf-df	0.95	1	1	0.76**	0.81**	0.79**
Glasgow	0.74	0.71	0.76	1	0.98**	0.94**
Glasgow-I	0.79	0.81	0.85	0.98	1	0.96**
Glasgow-II	0.82	0.86	0.79	0.94	0.96	1

(c) Amazon

	tf-idf	tf-üdf	tf-df	Glasgow	Glasgow-I	Glasgow-II
tf-idf	1	1**	1**	0.794**	0.837**	0.874**
tf-üdf	1	1	1	0.795	0.838**	0.877**
tf-df	1	1	1	0.794**	0.837**	0.875**
Glasgow	0.794	0.795	0.794	1	0.995**	0.974**
Glasgow-I	0.837	0.838	0.837	0.995	1	0.992**
Glasgow-II	0.874	0.877	0.875	0.974	0.992	1

(d) Yelp

	tf-idf	tf-üdf	tf-df	Glasgow	Glasgow-I	Glasgow-II
tf-idf	1	1**	1**	0.74**	0.77**	0.84**
tf-üdf	1	1	1**	0.75**	0.88**	0.82**
tf-df	1	1	1	0.73**	0.82**	0.83**
Glasgow	0.74	0.75	0.73	1	0.99**	0.92**
Glasgow-I	0.77	0.88	0.82	0.99	1	0.95**
Glasgow-II	0.84	0.82	0.83	0.92	0.95	1

(e) IMDb

	tf-idf	tf-üdf	tf-df	Glasgow	Glasgow-I	Glasgow-II
tf-idf	1	0.96**	0.99**	0.79**	0.77**	0.84**
tf-üdf	0.96	1	0.97**	0.76**	0.81**	0.78**
tf-df	0.99	0.97	1	0.76	0.81**	0.86**
Glasgow	0.79	0.76	0.76	1	0.99**	0.97**
Glasgow-I	0.77	0.81	0.81	0.99	1	0.96**
Glasgow-II	0.84	0.78	0.86	0.97	0.96	1

(f) MR2

	tf-idf	tf-üdf	tf-df	Glasgow	Glasgow-I	Glasgow-II
tf-idf	1	1**	1**	0.982**	0.983**	0.987**
tf-üdf	1	1	1**	0.982**	0.983**	0.987**
tf-df	1	1	1	0.982**	0.983**	0.987**
Glasgow	0.982	0.982	0.982	1	1.00**	1.00**
Glasgow-I	0.983	0.983	0.983	1.00	1	1.00**
Glasgow-II	0.987	0.987	0.987	1.00	1.00	1

(g) SemEval

	tf-idf	tf-üdf	tf-df	Glasgow	Glasgow-I	Glasgow-II
tf-idf	1	0.97**	0.96**	0.88**	0.93**	0.97**
tf-üdf	0.97	1	0.95**	0.88**	0.93**	0.97**
tf-df	0.96	0.95	1	0.88**	0.93**	0.97**
Glasgow	0.88	0.88	0.88	1	0.93**	0.96**
Glasgow-I	0.93	0.93	0.93	0.93	1	0.98**
Glasgow-II	0.97	0.97	0.97	0.96	0.98	1

(h) CR

	tf-idf	tf-üdf	tf-df	Glasgow	Glasgow-I	Glasgow-II
tf-idf	1	0.98**	0.99**	0.88**	0.92**	0.97**
tf-üdf	0.98	1	0.98**	0.86**	0.91**	0.95**
tf-df	0.99	0.98	1	0.88**	0.92**	0.97**
Glasgow	0.88	0.86	0.88	1	0.99**	0.96**
Glasgow-I	0.92	0.91	0.92	0.99	1	0.98**
Glasgow-II	0.97	0.95	0.97	0.96	0.98	1

(i) Pros & Cons

	tf-idf	tf-üdf	tf-df	Glasgow	Glasgow-I	Glasgow-II
tf-idf	1	0.90**	0.96**	0.68**	0.77**	0.78**
tf-üdf	0.90	1	0.96**	0.93**	0.89**	0.68**
tf-df	0.96	0.96	1	0.95**	0.80**	0.55**
Glasgow	0.68	0.93	0.95	1	0.86**	0.83**
Glasgow-I	0.77	0.89	0.80	0.86	1	0.91**
Glasgow-II	0.78	0.68	0.55	0.83	0.91	1

(j) Subj

	tf-idf	tf-üdf	tf-df	Glasgow	Glasgow-I	Glasgow-II
tf-idf	1	0.98**	0.95**	0.89**	0.93**	0.98**
tf-üdf	0.98	1	0.95**	0.89**	0.93**	0.98**
tf-df	0.95	0.95	1	0.89**	0.93**	0.98**
Glasgow	0.89	0.89	0.89	1	0.99**	0.96**
Glasgow-I	0.93	0.93	0.93	0.99	1	0.98**
Glasgow-II	0.98	0.98	0.98	0.96	0.98	1

Fig. 5. Comparison of rank correlation matrix between traditional and proposed method for all datasets.

The Lagrangian dual problem for linearly separable case is based on the Karush-Kuhn-Tucker conditions (Xanthopoulos & Razzaghi, 2014; Kartal et al., 2016)

$$\max {}_{\mu}L(\mu) = \sum_{i=1}^{n} \mu_i - \frac{1}{2}\sum_{i-1}^{n}\sum_{j=1}^{n} \mu_i\mu_j y_i y_j x_i^T x_j \tag{14a}$$

$$s.\ t. \sum_{i=1}^{n} \mu_i y_i = 0 \quad i = 1, ...n \tag{14b}$$

where "μ_i" represents the Lagrangian multiplier that is used to find the support vectors. In practice, data are mostly non-linearly separable. To overcome this shortcoming, the dual problem for the non-linearly separable case is considered. A penalty parameter (C) is introduced to account for training error to separate the hyperplane in an easier manner (Kartal et al., 2016).

$$\max_{\mu} L(\mu) = \sum_{i=1}^{n} \mu_i - \frac{1}{2}\sum_{i-1}^{n}\sum_{j=1}^{n} \mu_i\mu_j y_i y_j \phi(x)^T \phi(x) \tag{15a}$$

$$s.\ t. \sum_{i=1}^{n} \mu_i y_i = 0i = 1, ...n \tag{15b}$$

$$0 \leqslant \mu_i \leqslant C \tag{15c}$$

where "$\Phi(x)$" represents the mapping of the input vector into a higher dimension; this process makes group classification easier (Vapnik, 1998; Kartal et al., 2016). The kernel function is used to measure the similarity between the two data points x_i and x_j. However, this leads to all the data objects being assigned the same importance in the training process. (Xanthopoulos & Razzaghi, 2014)

$$K(x_i, x_j) = \Phi(x_i)^T \Phi(x_j) \tag{16}$$

In general, a linear kernel is adopted for the robust advantage that training an SVM with a linear kernel is faster than with other kernels because it requires fewer parameters for optimization (Hsu, Chang, & Li, 2016). Some popular kernel functions such as polynomial, radial basics function (RBF), and sigmoidal kernels (Kartal et al., 2016) are adopted. These kernel functions are listed in Table 3.

3.4.2. EC

The information regarding the actual and predicted classifications using the classification techniques is extracted from the confusion matrix. The data in the matrix are commonly used to evaluate the performance of the classification techniques. The confusion matrix for a two-class classifier is shown in Table 4 (Powers, 2011). The EC adopted to evaluate the SVM is listed in Table 5.

In the confusion matrix, +1 and −1 denote the frequency of positive and negative reviews, respectively; further, TP-True Positive is the count of actual positive reviews correctly classified as positive

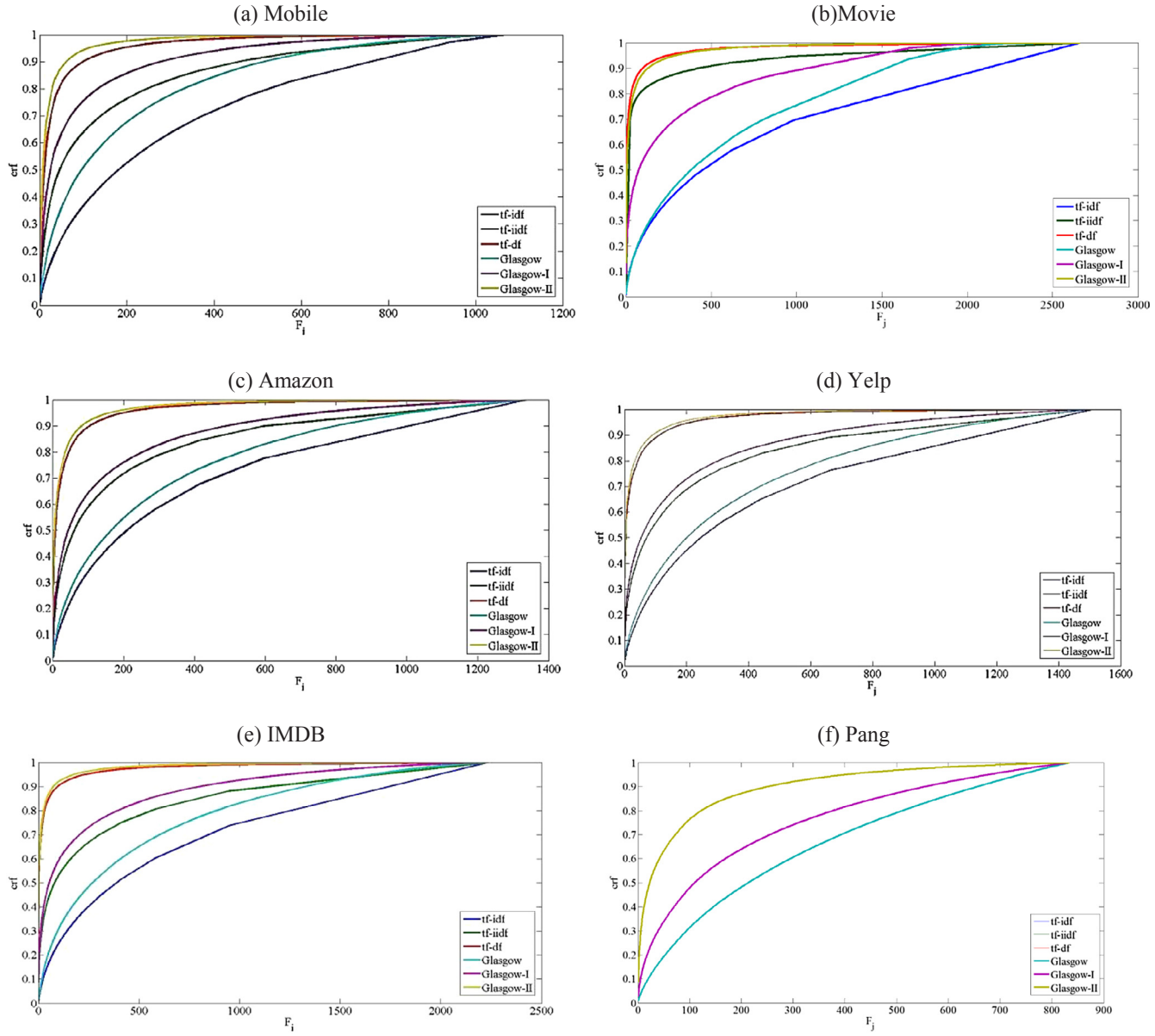

Fig. 6. Graphical representation for all datasets.

reviews, FP-False Positive is the count of actual negative reviews incorrectly classified as positive reviews, FN-False Negative is the count of actual positive reviews incorrectly classified as negative reviews, and TN-True Negative is the count of actual negative reviews is correctly classified as negative reviews.

3.5. Effectiveness measure (ρ)

Here, a new framework is presented to compute the effectiveness of the term weighting expressions through the removal of irrelevant features and irrelevant documents. The required formula is given as

$$\rho = \frac{1}{\log\left[\frac{N_f}{r_f} * \frac{N}{r_d}\right]} \tag{17}$$

where N_f is the total number of features extracted from the reviews, r_f is the number of features removed based on the term weighting methods, N is total number of documents, and r_d is the number of documents removed after extracting the prominent features in the Phase-II process. As is clear, a method will achieve better effectiveness if the logarithm value is smaller. Consequently, the inverse value should lead to a more effective method.

3.6. Strength of the expressions (SE(i))

As discussed above, the listed EC are utilized to check the accuracy or effectiveness of the algorithm. Li et al. (1995) proposed the consolidated structural strength formula for clustering techniques. For a classification technique, the equation is refined to obtain the strength of the expressions (SE(i)). The formula is given below:

$$SE(i) = \gamma * (CPM_{EC}(i)) + \lambda * (RW_{\rho}(i)) \tag{18}$$

where "γ" and "λ" are weightages. To compute $SE(i)$, first, "CPM_{EC}" and "RW_{ρ}" of the expressions are ranked individually. Higher weightage is given for higher ranking of "CPM_{EC}" and "RW_{ρ}", whereas lesser weightage is given for lower ranking. The weightings are decreased exponentially as the ranking decreases. For each instance, it is ensured that the summation of "γ" and "λ" is equal to 1, for which Eqs. (19) and (20) are adopted. This process is repeated for all the combinations wherein "γ" and "λ" vary between 0 and 1.

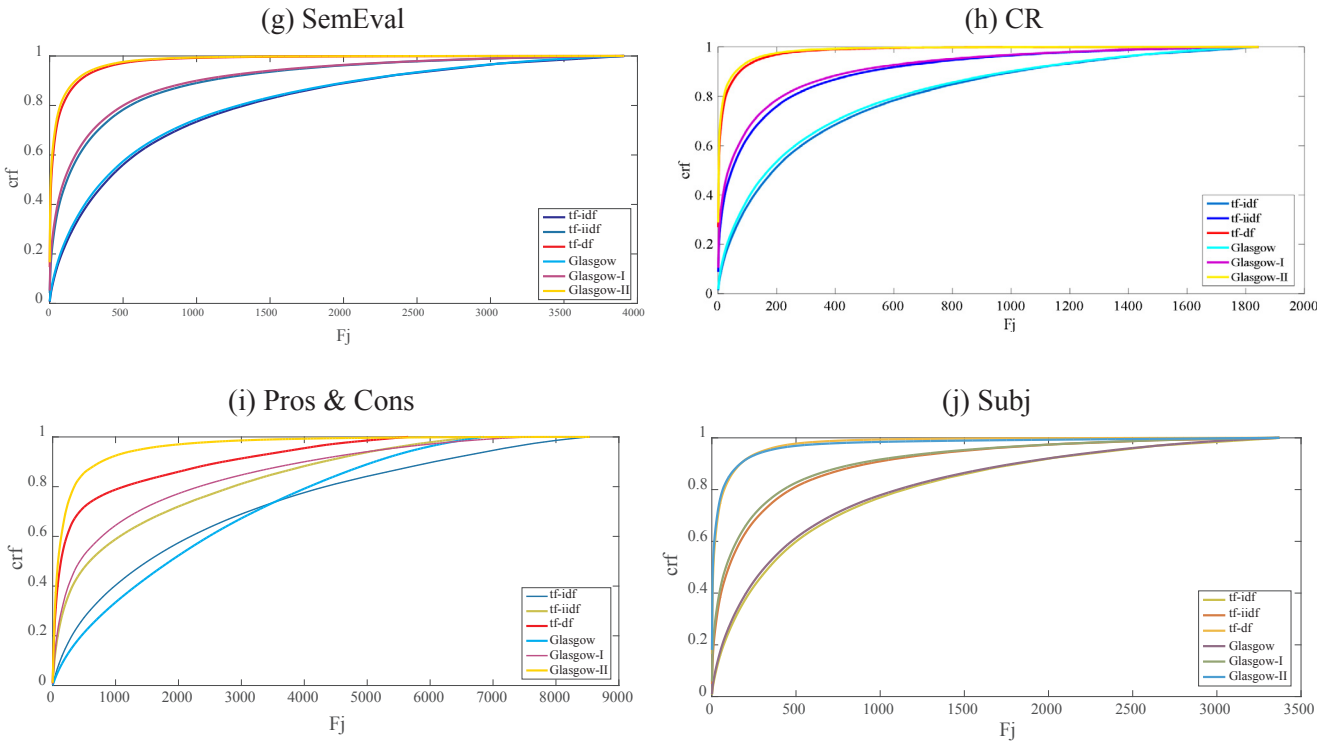

Fig. 6. (*continued*)

Table 9
Sample movie reviews with opinion.

Documents	Review	Opinion
d_1	Perfect Horror Movie	Positive
d_2	Not that Scary … Disappointment	Negative
d_3	It was good. Very good	Positive
d_4	Horrible, do you people even had eyes & any sense. Very worst. Disappointing	Negative
d_5	Completely disappointment	Positive
d_6	Great movie…Highly recommended for any horror fan	Positive
d_7	Awesome! Very Scary, great Character development & great Story. Highly Recommended	Positive
d_8	This Movie was crap, Horrible movie	Negative
d_9	Good love so much	Positive
d_{10}	Boring. Disappointed with this one!	Negative

Table 10
Uni-gram keywords extracted after Phase-I pre-processing.

Documents	Uni-gram keywords
d_1	{perfect; horror; movie}
d_2	{not; scary; disappoint}
d_3	{good; very}
d_4	{horrible; people; eyes; sense; very; worst; disappoint}
d_5	{complete; disappoint}
d_6	{great; movie; high; recommend; fan; horror}
d_7	{awesome; very; scary; great; character; develop; story; high; recommend}
d_8	{crap; horrible; movie}
d_9	{good; love}
d_{10}	{bore; disappoint}

$$\gamma = \frac{\alpha(1-\alpha)^{r-1}}{[\alpha(1-\alpha)^{r-1}] + [\beta(1-\beta)^{r-1}]} \tag{19}$$

$$\lambda = \frac{\beta(1-\beta)^{r-1}}{[\alpha(1-\alpha)^{r-1}] + [\beta(1-\beta)^{r-1}]} \tag{20}$$

Because there are more than one EC, the Combined Preference Measure for Evaluation Criteria is given as

$$CPM_{EC}(i) = \left[\frac{A_i}{\sum_{i=1}^{m} A_i}\right] + \left[\frac{F_i}{\sum_{i=1}^{m} F_i}\right] + \left[\frac{TNR_i}{\sum_{i=1}^{m} TNR_i}\right] + \left[\frac{NPV_i}{\sum_{i=1}^{m} NPV_i}\right] + \left[E_i \sum_{i=1}^{m} \frac{1}{E_i}\right]^{-1} + \left[FPR_i \sum_{i=1}^{m} \frac{1}{FPR_i}\right]^{-1} + \left[FNR_i \sum_{i=1}^{m} \frac{1}{FNR_i}\right]^{-1} + \left[FDR_i \sum_{i=1}^{m} \frac{1}{FDR_i}\right]^{-1} \tag{21}$$

Similarly, for effectiveness, the Relative Weightage (RW_ρ) is computed as

$$RW_\rho(i) = \frac{\rho_i}{\sum_{i=1}^{m} \rho_i} \tag{22}$$

3.6.1. Selection of weightage values

To determine the "γ" and "λ" values in the $SE(i)$, the $SE(i)$ is computed for the different combinations of "α" and "β" weightage values that vary between 0.1 and 0.9. Then, the maximum $SE(i)$ value is selected from the combinations.

The flow chart for the computation of the $SE(i)$ is shown in the Fig. 4.

3.7. Statistical significance "t" test for SE(i)

The statistical significance test for the best strength values for all the datasets is carried out to check the significance of the $SE(i)$. Two correlated sample t-tests are adopted and the hypotheses are framed as follows: Null hypothesis (H_0): There is no significant difference between the means values (H_0: $SA(i) = SA(j)$, where $i \neq j$). Alternate hypothesis (H_1): There is a significance difference between the mean values (H_1: $SA(i) \neq SA(j)$).

Traditional expressions						Proposed Expressions											
tf-idf			Glasgow			tf-iidf			tf-df			Glasgow-I			Glasgow-II		
Features	TW	crf	Features	TW	crf	Features	TW	crf	Features	TW	crf	Features	TW	crf	Features	TW	crf
good	3.48	0.13	good	6.68	0.11	disappoint	1.12	0.23	disappoint	0.59	0.27	disappoint	1.65	0.20	disappoint	1.19	0.27
disappoint	1.95	0.21	disappoint	5.40	0.20	good	0.65	0.37	good	0.30	0.41	good	1.00	0.31	good	0.52	0.38
complete	1.66	0.27	movie	3.58	0.26	movie	0.50	0.48	movie	0.26	0.53	movie	0.80	0.41	movie	0.51	0.50
love	1.66	0.33	very	3.54	0.32	very	0.43	0.57	very	0.23	0.63	very	0.79	0.50	very	0.50	0.61
bore	1.66	0.40	complete	3.46	0.38	horrible	0.21	0.61	horrible	0.10	0.68	horrible	0.38	0.55	horrible	0.20	0.65
movie	1.51	0.45	love	3.46	0.43	scary	0.19	0.65	scary	0.09	0.72	scary	0.37	0.59	scary	0.19	0.70
very	1.31	0.50	bore	3.46	0.49	great	0.18	0.69	great	0.08	0.76	great	0.36	0.64	great	0.19	0.74
perfect	1.11	0.55	horrible	2.55	0.53	complete	0.15	0.72	high	0.06	0.79	complete	0.29	0.67	high	0.15	0.77
horror	1.11	0.59	scary	2.45	0.57	love	0.15	0.75	recommend	0.06	0.81	love	0.29	0.70	recommend	0.15	0.81
crap	1.11	0.63	great	2.41	0.61	bore	0.15	0.78	complete	0.05	0.84	bore	0.29	0.74	complete	0.10	0.83
not	1.11	0.67	perfect	2.18	0.65	high	0.13	0.81	love	0.05	0.86	high	0.29	0.77	love	0.10	0.85
horrible	1.11	0.71	horror	2.18	0.69	recommend	0.13	0.84	bore	0.05	0.88	recommend	0.29	0.81	bore	0.10	0.87
scary	1.03	0.75	crap	2.18	0.72	perfect	0.10	0.86	perfect	0.03	0.90	perfect	0.18	0.83	perfect	0.06	0.89
great	0.98	0.79	not	2.18	0.76	horror	0.10	0.88	horror	0.03	0.91	horror	0.18	0.85	horror	0.06	0.90
high	0.72	0.82	high	1.93	0.79	crap	0.10	0.90	crap	0.03	0.93	crap	0.18	0.87	crap	0.06	0.92
recommend	0.72	0.85	recommend	1.93	0.82	not	0.10	0.92	not	0.03	0.94	not	0.18	0.89	not	0.06	0.93
fan	0.66	0.87	fan	1.49	0.85	fan	0.06	0.94	fan	0.02	0.95	fan	0.12	0.91	fan	0.04	0.94
people	0.48	0.89	people	1.23	0.87	people	0.04	0.95	people	0.01	0.96	people	0.10	0.92	people	0.04	0.95
eve	0.48	0.91	eve	1.23	0.89	eve	0.04	0.95	eve	0.01	0.97	worst	0.10	0.94	eve	0.04	0.96
sense	0.48	0.93	sense	1.23	0.91	sense	0.04	0.96	sense	0.01	0.97	sense	0.10	0.96	sense	0.04	0.96
worst	0.47	0.94	worst	1.23	0.93	worst	0.04	0.97	worst	0.01	0.98	eve	0.10	0.93	worst	0.04	0.97
awesome	0.37	0.96	awesome	1.09	0.95	awesome	0.03	0.98	awesome	0.01	0.98	awesome	0.09	0.97	awesome	0.03	0.98
character	0.37	0.97	character	1.09	0.96	character	0.03	0.99	character	0.01	0.99	character	0.09	0.98	character	0.03	0.99
develop	0.37	0.99	develop	1.09	0.98	develop	0.03	0.99	develop	0.01	0.99	develop	0.09	0.99	develop	0.03	0.99
story	0.37	1.00	story	1.09	1.00	story	0.03	1.00	story	0.01	1.00	story	0.09	1.00	story	0.03	1.00
	26.26			60.35			4.761			2.175			8.43			4.46	

Features above δ are highlighted

Fig. 7. FS based on crf for sample MR1 reviews.

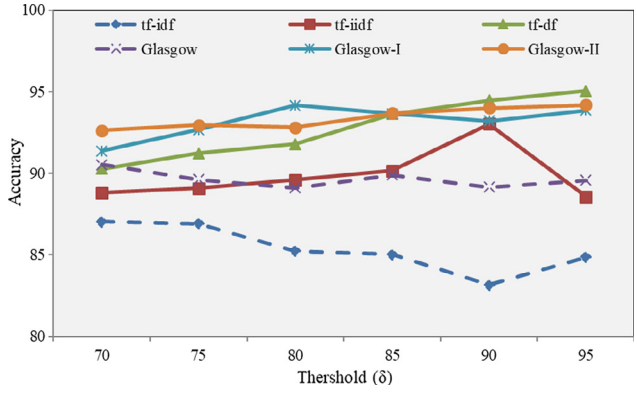

Fig. 8. Comparison of different "δ" with accuracy for MR1 dataset.

4. Experiments

4.1. Datasets

The experiments are conducted for both imbalanced and balanced classes. These include Mobile and Movie (MR1) datasets that are primary datasets, and Amazon, Yelp, IMDb, MR2, CR, SemEval, Pros & Cons and Subj that are secondary datasets. The scraping approach is adopted to extract the primary reviews, which is targeted at specific websites for specific data. The "rvest" package in R is used for scraping reviews from websites.

4.1.1. Mobile

The reviews for this dataset are related to smartphones of different brands and are obtained from the Amazon website (https://www.amazon.com/). In total, 1000 reviews are collected, consisting of 566 and 434 positive and negative reviews, respectively. The label descriptions are as follows: one and two stars are classified as negative reviews, whereas three, four, and five stars are considered as positive reviews. The distribution of reviews for each mobile brand is presented in Table 6.

4.1.2. Movie (MR1)

This dataset consists of 2355 reviews that are collected form the IMDb (http://www.imdb.com/) and Rotten Tomatoes websites (https://www.rottentomatoes.com/). Four films are taken into consideration. The label descriptions are as follows: For IMDb, ratings 1–3 are treated as negative reviews, whereas 7–10 are treated as positive reviews. For Rotten Tomatoes, a score of 1 or 2 is treated as a negative review, whereas a score of 4 or 5 is treated as a positive review. Neutral reviews are not considered. The summary of these reviews is listed in Table 7:

4.1.3. Amazon, Yelp, and IMDb

This dataset is taken from the study "From Group to Individual Labels using Deep Features" (Kotzias, 2015). It consists of three review datasets consists of (i) Product Reviews from Amazon.com, (ii) Restaurant Reviews from Yelp.com, and (iii) Movie Reviews from IMDb.com. This dataset consists of equal strength class labels with 50% in the positive class and 50% in the negative class.

4.1.4. Movie review (MR2)

This dataset consists of 10,662 snippets with 50% positive snippets and 50% negative reviews collected by Pang, & Lee (2005).

4.1.5. SemEval

Pontiki et al. (2016) developed this dataset; it consists of reviews about Laptop and Restaurant products. Training and Testing data ratio is 80:20. The characteristics of this dataset are listed in Table 8.

4.1.6. Customer review of electronic products (CR)

This dataset consists of five electronic product reviews including Canon, Nikon, DVD, MP3, and Cellphone customer reviews (Hu, & Liu, 2004).

4.1.7. Pros & cons

This dataset consists of reviews related to competing products with their positive and negative reviews. In total 30,324 reviews are used for evaluation with equal distribution of positive and negative reviews. (Liu et al., 2005)

4.1.8. Subj

This dataset consists of 5000 subjective and 5000 objective processed sentences (Pang & Lee, 2004).

Table 11
Comparison of slope values, N_f, reduction rate percentage, and r_d.

Datasets		tf-idf	tf-iidf	tf-df	Glasgow	Glasgow-I	Glasgow-II
Mobile	Slope	0.185	0.378	0.45	0.2	0.4	0.518
	N_f	375	185	190	351	178	167
	Reduction rate percentage	64.85	82.66	82.19	67.10	83.32	84.35
	r_d	12	48	51	13	54	49
MR1	Slope	0.069	0.210	0.311	0.0861	0.187	0.318
	N_f	1008	428	305	813	401	298
	Reduction rate percentage	62.12	83.92	88.54	69.45	84.93	88.80
	r_d	36	51	54	40	39	52
Amazon	Slope	0.15	0.29	0.39	0.15	0.3	0.42
	N_f	459	243	202	451	235	191
	Reduction rate percentage	65.67	81.82	84.89	66.27	82.42	85.71
	r_d	20	61	59	22	60	65
Yelp	Slope	0.128	0.245	0.36	0.135	0.25	0.43
	N_f	539	284	218	510	277	180
	Reduction rate percentage	64.35	81.22	85.58	66.27	81.68	88.10
	r_d	27	49	47	29	49	51
IMDb	slope	0.082	0.203	0.285	0.086	0.206	0.299
	N_f	846	303	233	807	297	220
	Reduction rate percentage	62.13	86.44	89.57	63.88	86.71	90.15
	r_d	14	21	25	14	22	49
MR2	Slope	0.165	0.161	0.216	0.165	0.18	0.216
	N_f	451	377	401	450	458	401
	Reduction rate percentage	45.99	54.85	51.98	46.11	45.15	51.98
	r_d	581	582	576	581	581	582
SemEval	Slope	0.081	0.178	0.263	0.083	0.196	0.298
	N_f	869	423	359	830	384	319
	Reduction rate percentage	77.77	89.18	90.82	78.77	90.18	91.84
	r_d	298	362	368	301	372	381
CR	Slope	0.165	0.379	0.625	0.175	0.460	0.669
	N_f	426	198	157	400	163	142
	Reduction rate percentage	76.90	89.26	91.49	78.31	91.16	92.30
	r_d	6	10	13	6	11	14
Pros & Cons	Slope	0.022	0.033	0.046	0.022	0.042	0.066
	N_f	3107	2303	1844	3236	1775	1434
	reduction rate percentage	63.60	73.02	78.39	62.09	79.20	83.20
	r_d	692	878	988	652	891	1041
Subj	Slope	0.081	0.266	0.309	0.085	0.231	0.289
	N_f	921	282	307	882	324	329
	Reduction rate percentage	72.69	91.64	90.89	73.85	90.39	90.25
	r_d	221	287	286	225	285	284

The dataset is preprocessed using Phase-I and II processes described in Section 3. The document term matrix is constructed using the binary occurrence of the terms in the reviews.

4.2. Results and discussion

As a part of Phase-II, the validation process is performed to evaluate whether the proposed methods are in agreement with the results of the traditional methods. This validation is based on rank correlation.

Fig. 5 shows the rank correlation matrix for all datasets. In this figure, when the correlation between the traditional and proposed methods is boldface, the ranks are in the same order, i.e., a positive relationship exists between the results of the traditional and proposed methods for all the datasets. The correlation coefficients are statistically tested and found significant.

As can be seen from Fig. 6, for the tf-df and Glasgow-II methods, a steep slope is observed initially, which then becomes constant at a certain point. A similar curve is obtained for all the datasets. The proposed method yields this steep curve structure for nearly 200 features. Phase-I and II are illustrated using sample movie reviews (MR1), where N = 10 with opinion labeled as positive and negative (see Table 9).

In Phase-I, the pre-processing tokenizing, filtering, and stemming methods are conducted for the uni-gram keywords of d_i listed in Table 10.

In Phase-II, the FS is carried out based on the cumulative relative frequency value, for illustration purpose we consider δ is 95 (see Fig. 7).

Our proposed method leads to the extraction of more relevant features than irrelevant features, resulting in higher precision. However, a stemming process is performed to improve the recall. Similarly, for all the other datasets, Phases-I and II are similarly applied.

The total number of features extracted after Phase-I from the Mobile and MR1 is 1067 and 2661 features, respectively. Further, for the same phase, 1337, 1512, 2234, 835, 3910, 1844, 8536, and 3373 features are extracted for the Amazon, Yelp, IMDb, MR2, SemEval, CR, Pros & Cons, and Subj datasets, respectively.

The SVM method is adopted to test the accuracy of these FS methods. To perform the SVM method, the R studio with the “kernlab” library is adopted. To test our proposed methods, 10-fold cross validation is applied for each TW method. The “δ” ranges from 70 to 95 with incremental intervals of 5. For each “δ” value, a corresponding feature set is extracted, and its accuracy is computed through SVM. For each TW, the maximum accuracy is noted and the corresponding feature set is selected for further analysis.

As an illustration, the comparative chart for the different “δ” with accuracy through SVM (linear kernel) for the MR1 dataset is shown in Fig. 8.

The horizontal axis represents “δ” values, while the vertical axis shows accuracy values. For each TW, the accuracy obtained through the

Table 12
Comparison of EC (%) for SVM.

	(a) Mobile						(b) MR1					
EC	tf-idf	tf-iidf	tf-df	Glasgow	Glasgow-I	Glasgow-II	tf-idf	tf-iidf	tf-df	Glasgow	Glasgow-I	Glasgow-II
A	86.07	87.66	91.68	86.52	91.29	**93.50**	87.02	93.03	**95.06**	90.53	94.16	94.18
E	13.93	12.34	8.32	13.48	8.71	**6.50**	14.48	6.97	**4.94**	9.47	5.84	5.82
P	75.40	77.33	**90.14**	75.38	88.18	90.12	92.10	94.47	**95.30**	91.64	94.55	95.08
R	92.89	**96.28**	93.74	95.25	95.09	96.17	83.62	92.92	94.67	90.19	**93.88**	93.20
F	75.54	77.45	91.41	75.98	88.89	**92.65**	88.00	93.65	**94.98**	90.88	94.15	94.13
TNR	85.92	87.33	91.15	86.08	90.67	**92.30**	93.35	93.07	**95.44**	91.02	94.51	95.16
FPR	14.08	12.67	8.85	13.92	9.33	**7.70**	6.65	6.93	**4.56**	8.58	5.49	4.84
FNR	5.60	3.72	6.26	4.75	4.91	**3.83**	16.38	7.08	**5.33**	8.61	6.12	6.80
FDR	24.60	22.67	9.86	24.62	11.82	**9.88**	7.00	5.53	**4.70**	7.56	5.45	4.92
NPV	93.79	94.83	90.89	94.40	92.40	**95.22**	75.21	90.50	**94.83**	89.00	92.35	93.32

	(c) Amazon						(d) Yelp					
EC	tf-idf	tfiidf	tf-df	Glasgow	Glasgow-I	Glasgow-II	tf-idf	tf-iidf	tf-df	Glasgow	Glasgow-I	Glasgow-II
A	87.36	92.07	91.54	89.08	92.14	**93.37**	87.73	90.28	90.53	88.24	**91.54**	89.98
E	12.64	7.93	8.46	10.92	7.86	**6.63**	12.27	9.72	9.47	11.76	**8.46**	10.02
P	86.63	92.65	91.48	89.91	93.70	**93.97**	88.35	90.80	**91.41**	86.82	89.97	87.80
R	87.96	92.53	92.07	89.96	91.73	**93.56**	87.91	90.42	89.92	90.51	**93.65**	93.50
F	87.29	92.54	91.72	89.66	92.57	**93.67**	88.10	90.58	90.60	88.56	**91.52**	90.24
TNR	86.76	91.52	91.17	88.90	93.06	**93.53**	87.60	90.20	**91.18**	86.20	90.05	87.18
FPR	13.24	8.48	8.83	11.10	6.94	**6.47**	12.40	9.80	**8.82**	13.80	9.95	12.82
FNR	12.04	7.47	7.93	10.04	8.27	**6.44**	12.09	9.58	10.08	9.49	**6.35**	6.50
FDR	13.37	7.35	8.52	10.09	6.30	**6.03**	11.65	9.20	**8.59**	13.18	10.03	12.20
NPV	88.07	91.44	91.50	88.59	90.57	**92.76**	87.11	89.80	89.67	89.69	**93.69**	93.12

	(e) IMDb						(f) MR2					
EC	tf-idf	tf-iidf	tf-df	Glasgow	Glasgow-I	Glasgow-II	tf-idf	tf-iidf	tf-df	Glasgow	Glasgow-I	Glasgow-II
A	83.03	85.16	**88.12**	83.53	86.51	88.10	86.97	87.40	91.88	88.16	91.02	**92.72**
E	16.97	14.84	**11.88**	16.47	13.49	11.90	13.03	12.60	8.12	11.84	8.98	**7.28**
P	89.66	**95.82**	91.57	91.05	92.24	94.02	87.72	89.58	91.68	88.55	92.38	**92.72**
R	80.45	83.02	87.47	81.22	85.23	**86.74**	85.92	87.44	91.76	87.49	91.23	**92.37**
F	86.27	88.10	89.02	86.40	88.23	**89.87**	86.81	88.46	91.71	88.01	91.79	**92.54**
TNR	93.85	**94.17**	91.21	89.21	90.38	92.07	87.99	87.13	92.03	88.84	90.53	**93.00**
FPR	6.15	**5.83**	8.79	10.79	9.52	7.93	12.01	12.87	7.97	11.16	9.47	**7.00**
FNR	19.55	16.98	**12.53**	18.78	14.47	13.26	14.08	12.56	8.24	12.51	8.77	**7.63**
FDR	6.44	**4.18**	8.43	6.95	7.56	5.98	12.28	10.42	8.32	11.45	7.62	**7.28**
NPV	69.60	73.12	**83.21**	72.10	79.52	80.91	86.26	84.21	92.05	87.76	88.98	**92.64**

	(g) SemEval						(h) CR					
EC	tf-idf	tf-iidf	tf-df	Glasgow	Glasgow-I	Glasgow-II	tf-idf	tf-iidf	tf-df	Glasgow	Glasgow-I	Glasgow-II
A	82.32	84.15	85.25	83.39	85.13	**87.66**	83.65	**89.39**	87.35	84.81	87.61	86.69
E	17.68	15.85	14.75	16.41	14.87	**12.14**	16.35	**10.61**	12.65	15.19	12.39	13.31
P	80.62	81.60	89.57	86.47	82.16	**86.06**	82.02	**88.27**	86.27	83.31	85.31	85.00
R	81.24	83.53	83.47	83.99	84.96	**88.76**	83.51	**89.02**	87.02	84.61	88.56	87.02
F	80.35	82.28	89.02	83.70	83.26	**87.23**	82.63	**88.58**	86.58	83.82	86.80	85.91
TNR	83.89	84.65	91.21	88.22	85.22	**86.29**	83.95	**89.77**	87.69	85.15	86.86	86.48
FPR	16.11	15.35	8.79	11.78	14.78	**13.71**	16.05	**10.23**	12.31	14.85	13.14	13.52
FNR	18.76	16.47	12.53	16.01	15.04	**11.24**	16.49	**10.98**	12.98	15.39	11.44	12.98
FDR	19.38	18.40	8.43	13.53	17.84	**13.94**	17.98	**11.73**	13.73	16.69	14.69	15.00
NPV	83.78	85.97	83.21	82.47	87.41	**86.54**	84.90	**90.04**	87.95	85.96	89.54	87.98

	(i) Pros & Cons						(j) Subj					
EC	tf-idf	tf-iidf	tf-df	Glasgow	Glasgow-I	Glasgow-II	tf-idf	tf-iidf	tf-df	Glasgow	Glasgow-I	Glasgow-II
A	88.38	90.18	**93.24**	88.44	90.34	91.02	83.65	88.2	92.37	86.49	88.69	**95.14**
E	11.62	9.82	**6.76**	11.56	9.66	8.98	17.85	11.8	7.63	15.01	12.81	**4.86**
P	85.52	88.22	**91.70**	86.57	89.13	88.01	88.74	86.74	88.33	87.12	88.55	**93.55**
R	87.94	90.76	92.18	88.54	**93.17**	90.95	79.70	86.72	93.73	82.76	85.56	**94.75**
F	86.53	89.34	**91.86**	87.52	89.71	89.27	84.35	84.95	90.66	87.33	89.47	**94.14**
TNR	87.41	88.93	**94.17**	87.65	86.45	89.79	90.50	89.85	90.48	93.07	94.48	**94.8**
FPR	12.59	11.07	**5.83**	12.35	13.55	10.21	9.50	10.15	9.52	6.93	5.52	**5.2**
FNR	12.06	9.24	**7.82**	11.46	7.93	9.05	20.30	13.28	6.27	17.24	14.44	**5.25**
FDR	14.48	11.78	**8.30**	13.43	12.17	11.99	10.36	13.26	11.67	7.48	6.05	**6.45**
NPV	90.77	92.14	91.17	89.81	90.94	**93.46**	71.84	91.18	95.91	74.64	77.91	**96.17**

The best results are boldfaced.

Fig. 9. Comparison of accuracy, precision and recall for all datasets.

SVM is denoted on the line plot. From the figure, it is clear that the proposed expressions lead to higher accuracy in the case of most of the "δ" values.

Furthermore, the slope values, N_f, reduction rate percentages $((1-\frac{N_f\,afterphase-II}{N_f\,afterphase-I})X100)$, and r_d values for the maximum accuracy case are listed in Table 11.

The linear, polynomial, RBF, and sigmoidal kernels are used for the evaluation process from which the maximum accuracy value is extracted corresponding to the evaluation criteria results; these values are listed in Table 12. For most of the datasets such as Mobile, MR1, IMDb, CR, Amazon, Yelp, and Subj, the linear and polynomial kernels outperformed the RBF and sigmoidal kernels. However, for MR2, SemEval, and Pros & Cons, the RBF kernel provides better EC than the linear, polynomial, and sigmoidal kernels. Our proposed method outperforms the traditional methods while using either the linear, polynomial, or RBF kernels for all the test datasets. However, this is not true in the case of sigmoidal kernel. Nevertheless, in all cases, the linear, polynomial and RBF kernels achieve better results; these results are shown in Table 12.

As an example, the accuracy, precision, and recall values are visually represented in Fig. 9.

Fig. 9 shows the number of features extracted for each term weighting expressions against accuracy as a bar chart; the line in the figure represents the precision and recall values for each term weighting. Our results show that the accuracy values are high for the proposed methods even with a smaller number of features; in addition, the precision and recall for the proposed methods are better than the traditional methods.

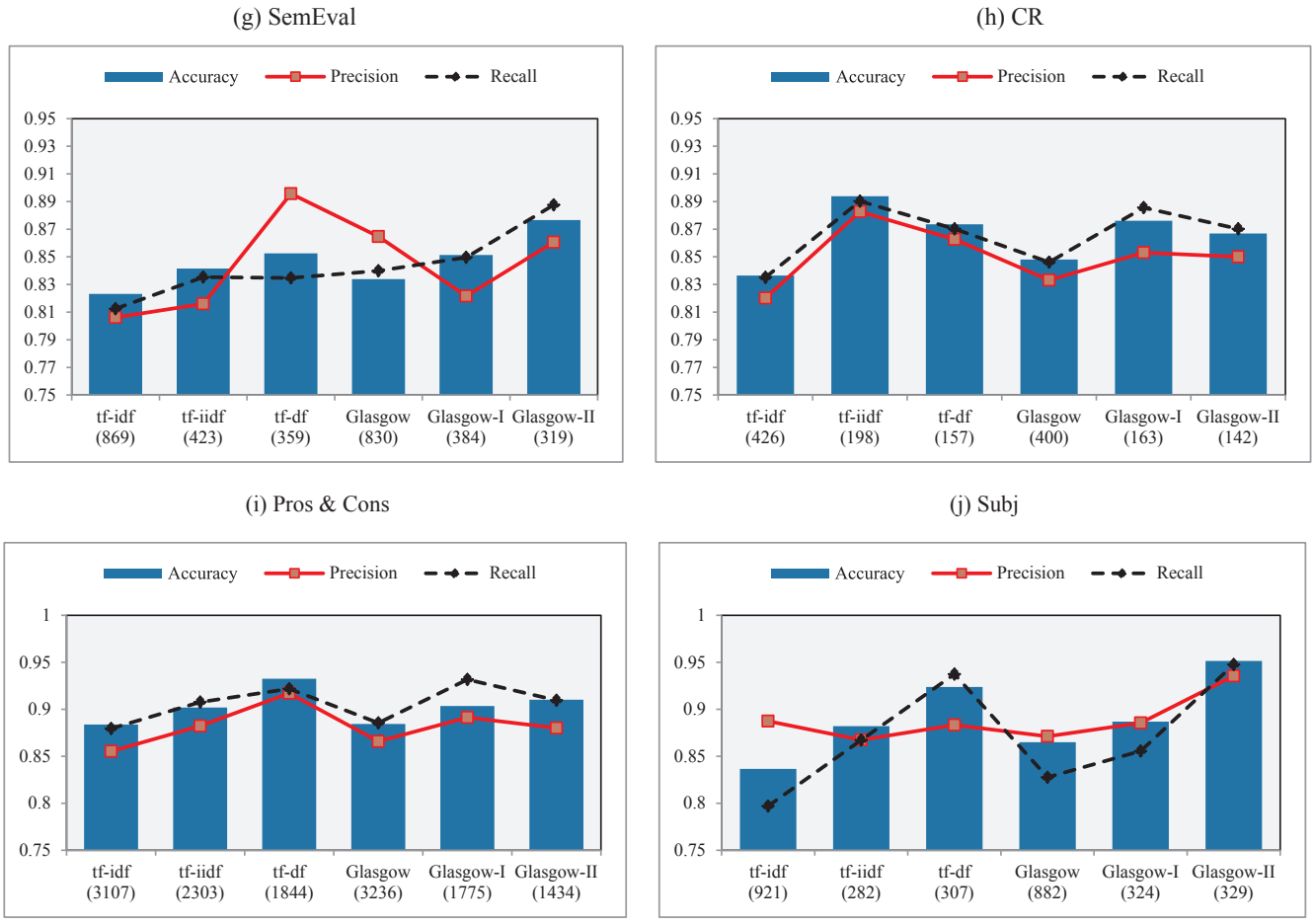

Fig. 9. (*continued*)

Table 13
Comparison of "ρ" for traditional and proposed methods.

Datasets	tf-idf	tf-iidf	tf-df	Glasgow	Glasgow-I	Glasgow-II
Mobile	0.474	0.714	0.726	0.486	**0.742**	0.723
MR1	0.494	0.575	**0.591**	0.519	0.540	0.586
Amazon	0.531	0.768	0.769	0.545	0.766	**0.797**
Yelp	0.568	0.714	0.717	0.583	0.715	**0.742**
IMDb	0.485	0.574	0.606	0.488	0.582	**0.738**
MR2	0.625	**0.656**	0.644	0.625	0.622	0.646
SemEval	0.658	0.726	0.734	0.662	0.735	**0.745**
CR	0.546	0.647	0.703	0.548	0.669	**0.722**
Pros & Cons	0.544	0.597	0.628	0.533	0.612	**0.648**
Subj	0.557	0.633	0.631	0.562	**0.629**	0.628

The best results are boldfaced.

Table 14
Comparison of CPM_{EC} for traditional and proposed methods.

Dataset	tf-idf	tf-iidf	tf-df	Glasgow	Glasgow-I	Glasgow-II
Mobile	1.12	1.24	1.45	1.15	1.42	**1.63**
MR1	1.059	1.323	**1.570**	1.159	1.431	1.457
Amazon	1.079	1.4	1.336	1.2	1.446	**1.561**
Yelp	1.205	1.370	1.404	1.214	**1.464**	1.344
IMDb	1.289	**1.444**	1.360	1.195	1.298	1.413
MR2	1.15	1.18	1.47	1.21	1.42	**1.57**
SemEval	0.99	1.04	**1.28**	1.10	1.07	1.18
CR	1.03	**1.28**	1.17	1.06	1.18	1.14
Pros & Cons	1.19	1.32	**1.64**	1.21	1.31	1.35
Subj	1.055	1.2	1.419	1.2	1.351	**1.814**

The best results are boldfaced.

Table 15
Rank of CPM_{EC} and RW_{ρ} values for different TWE.

Term weighting	CPM_{EC}	RW_{ρ}	Rank for CPM_{EC}	Rank for RW_{ρ}
tf-idf	1.059	0.150	6	6
tf-iidf	1.323	0.174	4	3
tf-df	1.570	0.179	1	1
Glasgow	1.159	0.157	5	5
Glasgow-I	1.431	0.163	3	4
Glasgow-II	1.457	0.177	2	2

Table 16
Distribution of exponential weightages.

Rank	Exponential weightage for CPM_{EC}	Exponential weightage for RW_{ρ}
1	α	β
2	$\alpha(1-\alpha)$	$\beta(1-\beta)$
3	$\alpha(1-\alpha)^2$	$\beta(1-\beta)^2$
4	$\alpha(1-\alpha)^3$	$\beta(1-\beta)^3$
5	$\alpha(1-\alpha)^4$	$\beta(1-\beta)^4$
6	$\alpha(1-\alpha)^5$	$\beta(1-\beta)^5$

As previously discussed in Section 3.6, the effectiveness of the proposed and traditional methods is calculated and compared; this comparison is shown in Table 13.

The values listed in Table 13 clearly show that our proposed methods show better effectiveness measures than traditional methods. Similarly, Glasgow-II shows better effectiveness measures for Amazon, Yelp, IMDb, SemEval, CR, and Pros & Cons Datasets.

Table 14 shows the comparison of CPM_{EC} for the proposed and

Table 17
Example of relative weightage with α = 0.2 and β = 0.1.

Term weighting expressions	γ	λ
tf-idf	0.526	0.474
tf-iidf	0.558	0.442
tf-df	0.690	0.310
Glasgow	0.555	0.445
Glasgow-I	0.637	0.363
Glasgow-II	0.615	0.385

Table 18
Weightage value for the maximum SE(i).

Term weighting expressions	γ	λ
tf-idf	0.9998	0.0002
tf-iidf	0.8901	0.1099
tf-df	0.1000	0.9000
Glasgow	0.9986	0.0014
Glasgow-I	0.9890	0.0110
Glasgow-II	0.5000	0.5000

traditional methods. Thus, with all evaluation criteria, the proposed methods, tf-iidf, tf-df, Glasgow-I, and Glasgow-II show better results than the traditional methods.

The flowchart in Fig. 5 is explained below using CPM_{EC} and RW_{ρ}. The MR1 dataset is considered in the illustration.

The steps involved in this calculation are as follows. The CPM_{EC} and RW_{ρ} values are ranked in descending order; they are listed in Table 15.

The highest values of CPM_{EC} and RW_{ρ} are given the highest weightage of "α" and "β", respectively. The 'α' and 'β' are varied from 0.1 to 0.9. Then, the smaller rank is weighed with exponential weightage $\alpha(1-\alpha)^{r-1}$ for CPM_{EC} and $\beta(1-\beta)^{r-1}$ for RW_{ρ} for r = 1….m, where m is total number of expressions. The distribution of exponential weightage is shown in Table 16.

The term weighting expressions are weighted correspondingly to the rank of CPM_{EC} and RW_{ρ}. Then, the relative weightage "γ" and "λ" for the MR1 dataset is computed as follows (see Table 17).

The "γ" and "λ" are weighted to the corresponding ranking of CPM_{EC} and RW_{ρ} for each of the term weighting expressions. Then, the SE(i) is computed for different "α" and "β" combinations. From those combinations, the maximum value of the SE(i) is selected as the best SE (i) (see Fig. 10).

For the maximum SE(i), corresponding "γ" and "λ" values for all term weighting expressions are listed in Table 18.

Table 19
Comparison of SE(i) values for traditional and proposed methods.

Datasets	tf-idf	tf-iidf	tf-df	Glasgow	Glasgow-I	Glasgow-II
Mobile	1.12	1.20	1.24	1.15	1.42	1.56
MR1	1.06	1.48	**1.62**	1.16	1.42	1.13
Amazon	1.08	1.50	1.35	1.19	1.43	**1.55**
Yelp	1.21	**1.46**	1.41	1.21	1.36	1.34
IMDb	1.24	1.23	**1.34**	1.2	1.26	1.3
MR2	1.10	1.16	1.16	1.12	1.33	**1.46**
SemEval	0.99	1.05	**1.31**	1.10	1.06	1.14
CR	1.02	**1.36**	1.13	1.03	1.14	1.13
Pros & Cons	1.18	1.39	**1.64**	1.19	1.21	1.23
Subj	1.05	1.28	1.58	1.17	1.23	**1.79**

The best results are boldfaced.

(a) tf-idf

α \ β	0.1	0.2	0.3	0.4	0.5	0.6	0.7	0.8	0.9
0.1	0.60	0.628	0.569	0.464	0.34	0.24	0.18	0.15	0.15
0.2	0.592	0.617	0.554	0.57	0.32	0.23	0.17	0.15	0.15
0.3	0.656	0.681	0.617	0.50	0.36	0.25	0.18	0.15	0.15
0.4	0.768	0.791	0.733	0.60	0.45	0.30	0.20	0.16	0.15
0.5	0.87	0.88	0.84	0.75	0.60	0.41	0.24	0.16	0.15
0.6	0.97	1.021	0.96	0.91	0.80	0.60	0.35	0.19	0.15
0.7	1.034	1.036	1.029	1.01	0.970	0.862	0.605	0.269	0.155
0.8	1.100	1.100	1.099	1.05	1.044	1.023	0.940	0.605	0.891
0.9	1.059	1.059	1.059	1.059	1.059	**1.06**	1.054	1.028	0.605

(b) tf-iidf

α \ β	0.1	0.2	0.3	0.4	0.5	0.6	0.7	0.8	0.9
0.1	0.75	0.79	0.74	0.62	0.34	0.28	0.13	0.04	0.02
0.2	0.70	0.75	0.69	0.57	0.42	0.25	0.12	0.06	0.03
0.3	0.76	0.81	0.75	0.63	0.47	0.29	0.15	0.07	0.05
0.4	0.88	0.93	0.87	0.75	0.57	0.37	0.19	0.09	0.06
0.5	1.04	1.08	1.03	0.92	0.75	0.52	0.28	0.12	0.07
0.6	1.22	1.24	1.21	1.12	0.98	0.75	0.44	0.18	0.08
0.7	1.37	1.37	1.35	1.30	1.22	1.06	0.75	0.32	0.10
0.8	1.45	1.44	1.43	1.41	1.38	1.32	1.17	0.75	0.17
0.9	**1.48**	1.46	1.45	1.44	1.43	1.42	1.40	1.33	0.75

(c) tf-df

α \ β	0.1	0.2	0.3	0.4	0.5	0.6	0.7	0.8	0.9
0.1	0.88	0.90	0.80	0.63	0.43	0.27	0.19	0.17	0.18
0.2	0.86	0.88	0.78	0.61	0.42	0.27	0.20	0.18	0.18
0.3	0.97	0.99	0.89	0.71	0.50	0.31	0.21	0.18	0.18
0.4	1.15	1.17	1.08	0.90	0.65	0.40	0.24	0.18	0.18
0.5	1.35	1.37	1.30	1.15	0.90	0.58	0.31	0.20	0.18
0.6	1.52	1.52	1.49	1.41	1.24	0.91	0.49	0.23	0.18
0.7	1.61	1.60	1.59	1.56	1.51	1.34	0.92	0.37	0.18
0.8	1.62	1.61	1.60	1.61	1.61	1.59	1.48	0.93	0.23
0.9	1.59	1.58	1.58	1.59	1.59	1.60	1.62	**1.62**	0.95

(d) Glasgow

α \ β	0.1	0.2	0.3	0.4	0.5	0.6	0.7	0.8	0.9
0.1	0.66	0.71	0.68	0.60	0.48	0.35	0.24	0.18	0.16
0.2	0.60	0.66	0.63	0.55	0.43	0.32	0.22	0.17	0.16
0.3	0.63	0.69	0.66	0.58	0.46	0.33	0.23	0.17	0.16
0.4	0.72	0.77	0.74	0.66	0.53	0.39	0.26	0.18	0.16
0.5	0.84	0.88	0.86	0.78	0.66	0.49	0.31	0.20	0.16
0.6	0.97	1.00	0.98	0.93	0.83	0.66	0.43	0.23	0.16
0.7	1.08	1.09	1.09	1.06	1.01	0.89	0.66	0.34	0.17
0.8	1.14	1.14	1.14	1.13	1.12	1.08	0.97	0.66	0.22
0.9	1.16	**1.16**	1.16	1.16	1.16	1.15	1.14	1.40	0.66

(e) Glasgow-I

α \ β	0.1	0.2	0.3	0.4	0.5	0.6	0.7	0.8	0.9
0.1	0.831	0.971	1.011	1.005	0.964	0.884	0.751	0.550	0.303
0.2	0.723	0.868	0.911	0.904	0.958	0.777	0.646	0.465	0.266
0.3	0.722	0.866	0.909	0.903	0.859	0.775	0.645	0.464	0.265
0.4	0.777	0.920	0.962	0.956	0.913	0.831	0.698	0.506	0.283
0.5	0.879	1.015	1.053	1.047	1.008	0.931	0.800	0.593	0.323
0.6	1.023	1.138	1.168	1.164	1.133	1.069	0.951	0.740	0.404
0.7	1.191	1.268	1.287	1.284	1.265	1.222	1.138	0.960	0.572
0.8	1.338	1.371	1.378	1.377	1.369	1.352	1.314	1.220	0.904
0.9	1.417	1.422	1.073	**1.423**	1.422	1.419	1.413	1.396	1.316

(f) Glasgow-II

α \ β	0.1	0.2	0.3	0.4	0.5	0.6	0.7	0.8	0.9
0.1	0.82	1.00	1.07	1.11	**1.13**	1.11	1.07	1.00	0.82
0.2	0.64	0.82	0.90	0.95	0.96	0.95	0.90	0.82	0.64
0.3	0.56	0.73	0.82	0.86	0.87	0.86	0.82	0.73	0.56
0.4	0.53	0.69	0.77	0.82	0.83	0.82	0.77	0.69	0.53
0.5	0.52	0.68	0.76	0.80	0.82	0.80	0.76	0.68	0.52
0.6	0.53	0.69	0.77	0.82	0.83	0.82	0.77	0.69	0.53
0.7	0.56	0.73	0.82	0.86	0.87	0.86	0.82	0.73	0.56
0.8	0.64	0.82	0.90	0.95	0.96	0.95	0.90	0.82	0.64
0.9	0.82	1.00	1.07	1.11	1.13	1.11	1.07	1.00	0.82

Highlighted cells denotes maximum SE(i)

Fig. 10. SE(i) values for MR1 dataset with different combinations of 'α' and 'β'.

Table 20
Comparison of *t*-test $_{(9,0.05)}$ with table value *1.833*.

	tf-idf	tf-iidf	tf-df	Glasgow	Glasgow-I	Glasgow-II
tf-idf		**4.19****	**4.61****	2.59	**4.54****	**3.50****
tf-iidf			1.19	**3.59****	0.57	0.59
tf-df				**4.57****	1.38	.17
Glasgow					**3.90****	**3.19****
Glasgow-I						1.16
Glasgow-II						

** Accept Alternate Hypothesis (H_1).

Table 21
Comparison of proposed method with other approaches A: Accuracy, P: Precision, R: Recall, F: F-Measure.

FS Methods	MR2	CR	SemEval	Pros & Cons	subj
Proposed Methods	***92.72 (A)***	***89.39 (A)*** 88.27(P) ***89.02(R)*** ***88.58 (F)***	***86.06 (P)*** ***88.76 (R)*** ***87.23 (F)***	***93.24 (A)***	95.14 (A)
Ensemble TextHMMs	82.39 (A)	87.0 (A)	–	92.5 (A)	***98.1 (A)***
CNN-non-static	81.5 (A)	79.8 (A)	–	–	93.2 (A)
CNN-rand	76.1 (A)	85.0 (A)	–	–	89.6 (A)
CNN-Multi	81.1 (A)	81.8 (A)	–	–	93.2 (A)
MV-RNN	79.0 (A)	–	–	–	–
RAE	77.7 (A)	–	–	–	–
NBSVM	79.4 (A)	80.2 (A)	–	–	93.2 (A)
MNB	79.0 (A)	–	–	–	93.6 (A)
Tree – CRF	77.3 (A)	81.4 (A)	–	–	–
Opinion observer	–	–	–	90.2 (A)	–
Deep CNN + LP	–	***90.2 (P)*** 86.2 R) 88.1 (F)	82.2 (P) 87.5 (R) 84.7 (F)	–	–
Deep CNN	–	85.6 (P) 81.2 (R) 83.3 (F)	74.6 (P) 85.1 (R) 79.3 (F)	–	–
Ordinal-based Uni-gram	–	–	–	–	90.70 (A)
frequency-based Uni-gram	–	–	–	–	90.76 (A)
Ordinal-based Parts of Speech	–	–	–	–	92.35 (A)
frequency-based Parts of Speech	–	–	–	–	92.92 (A)

The boldfaced and italic are best results.

Similarly, for all the datasets, the computed SE(i) values are listed Table 19.

The *t*-test is performed to test the statistical significance of the SE(i). The confidence level is fixed as 95 percent and degrees of freedom are set to 9.

For all comparisons, the proposed methods show better results than the traditional methods (see Table 20).

4.3. Comparison using other approaches

This section we compare our proposed method with other FS approaches. Kang, Ahn, and Lee (2017) proposed an ensemble of text-based Hidden Markov models (TextHMMs) for text classification. Nakagawa et al. (2010) proposed a dependency tree method using conditional random fields. Further, Matrix Vector Recursive Neural Network methods are proposed by Socher, Huval, Manning, and Ng (2012). Kim (2014) proposed the Convolutional Neural Network using pre-trained vectors (CNN-non-static), random initialization (CNN-rand), and a multichannel architecture (CNN-multi). Socher, Pennington, Huang, and Ng, & Manning (2011) proposed the use of recursive autoencoders with pre-trained word vectors, naive Bayes with uni-bigrams, and multinomial naive Bayes. Yousefpour, Ibrahim, and Hamed, 2017 proposed ordinal-based and frequency-based integration of different feature subsets. Our proposed methods with above FS approaches for same EC are tabulated. The results are listed in Table 21.

For most of the datasets, our proposed term weighting expressions provides better EC than the other FS approaches.

5. Conclusions

This study addresses the problem of FS and provides solutions in terms of new expressions. In doing so, the seed expressions tf-idf term weighting is refined. The suggested cumulative curve is used to determine the number of features. SVMs with different kernels are used to evaluate the performance of the proposed feature selection expressions. Primary and secondary datasets are used for analysis to validate the proposed new expressions. The evaluation criteria results are estimated using SVMs for different term weighting expressions, which are then tabulated. The proposed term weighting expressions provides higher accuracy, precision, and recall values for all the datasets. A new effectiveness measure is proposed and tested on different term weighting expressions, which perform substantially better than the traditional methods. The overall strength of the expressions (SE(i)) is computed through the combinations of evaluation criteria and effectiveness. Both the evaluation criteria and effectiveness are exponentially weighted based on the corresponding rank positions of the different term weighting expressions. The significance test is conducted on the SE(i) values; the results shows that the proposed methods perform better. The proposed expressions outperformed other methods in all cases involving the extraction of the lesser feature set to enhance the performance of the classification technique, proving that the proposed term weighting expressions are the better than the existing methods for FS. However, the proposed FS method is only based on the high-frequency features; thus, there is a possibility to that some low-frequency critical terms might be excluded, which is a limitation of our expressions. This shortcoming could be addressed as a future work. Furthermore, in the pursuit of developing and selecting better methods for FS, as future work, the existing and proposed methods may be integrated, to develop better expressions for FS.

References

Agnihotri, D., Verma, K., & Tripathi, P. (2017). Variable global feature selection scheme for automatic classification of text documents. *Expert Systems with Applications, 81*, 268–281.

Allen, T., Sui, Z., & Parker, L. N. (2017). Timely decision analysis enabled by efficient social media modeling. *Decision Analysis,* 1–11.

Allen, T., Xiong, H., & Afful-Dadzie, A. (2016). A directed topic model applied to call center improvement (and Computational Natural Language Learning pp. 1201–1211. Association for Computational Linguistics.) *Applied Stochastic Models in Business and Industry, 32*(1), 57–73.

Bag, S., Tiwari, M, K., & Chan, F. T. S. (2017). Predicting the consumer's purchase intention of durable goods: An attribute-level analysis, Journal of Business Research.

Bharti, K. K., & Singh, P. K. (2014). A three-stage unsupervised dimension reduction method for text clustering. *Journal of Computational Science, 5*(2), 156–169.

Bharti, K. K., & Singh, P. K. (2015). Hybrid dimension reduction by integrating feature selection with feature extraction method for text clustering. *Expert Systems with Applications, 42*, 3105–3114.

Claypo, N., & Jaiyen, S. (2014). Opinion mining for Thai restaurant reviews using neural networks and mRMR feature selection. In Computer Science and Engineering Conference (ICSEC), 2014 International (pp. 394–397). IEEE.

Cortes, C., & Vapnik, V. (1995). Support vector networks. *Machine Learning, 20*, 273–297.

Eirinaki, M., Pisal, S., & Singh, J. (2012). Feature-based opinion mining and ranking. *Journal of Computer and System Sciences, 78*(4), 1175–1184.

Galavotti, L., Sebastiani. F., & Simi, M. (2000). Experiments on the use of feature selection and negative evidence in automated text categorization. In J. L. Borbinha and T. Baker, editors, Proceedings of ECDL-00, 4th European Conference on Research and Advanced Technology for Digital Libraries, pages 59–68, Lisbon, PT, Springer Verlag, Heidelberg, DE. Published in the"Lecture Notes in Computer Science" Series, pp. 19–23.

Ghareb, A. S., Abu Bakar, A., & Hamdan, A. R. (2016). Hybrid feature selection based on enhanced genetic algorithm for text categorization. *Expert Systems with Applications, 49*, 31–47.

Hsu, C. W., Chang, C. C., & Li, C. J. (2016). A Practical Guide to Support Vector Classification.

Hu, M., & Liu, B. (2004, July). Mining opinion features in customer reviews. In Proceedings of nineteenth national conference on artificial intelligence (AAAI-2004), San Jose, USA.

Jones, K. S. (1972). A statistical interpretation of term specificity and its application in retrieval. *Journal of Documentation, 28*, 11–21.

Kang, M., Ahn, J., & Lee, K. (2017). Opinion mining using ensemble text hidden Markov models for text classification. *Expert Systems with Applications, 94*, 218–227.

Kartal, H., Oztekin, A., Gunasekaran, A., & Cebi, F. (2016). An integrated decision analytics framework of machine learning with multi-criteria decision making for multi-attribute inventory classification. *Computers & Industrial Engineering, 101*, 599–613.

Kim. Y. (2014). Convolutional neural networks for Sentence classification. arXiv preprint arXiv:1408.5882.

Kotzias, D., Denil, M., De Freitas, N., & Smyth, P. (2015). From group to individual labels using deep features. In Proceeding KDD '15 Proceedings of the 21th ACM SIGKDD international conference on knowledge discovery and data mining, pp. 597–606.

Lan, M., Tan, C. L., & Low, H. B. (2006, July). Proposing a new term weighting scheme for text categorization. In AAAI (Vol. 6, pp. 763–768).

Lee, K., Cho, S., & Asfour, S. (2011). Web-based algorithm for cylindricity evaluation using support vector machine learning. *Computers & Industrial Engineering, 60*, 228–235.

Lee, C., Wang, Y., & Trappey, A. (2015). Ontology based reasoning for the intelligent handling of customer complaints. *Computers & Industrial Engineering, 84*, 144–155.

Li, R. P., & Mukaidono, M. (1995, March). A maximum-entropy approach to fuzzy clustering. In Fuzzy systems, 1995. International joint conference of the fourth IEEE international conference on fuzzy systems and the second international fuzzy engineering symposium, proceedings of 1995 IEEE int. (Vol. 4, pp. 2227–2232). IEEE.

Liu, T., Liu, S., Chen, Z., & Ma, W. Y. (2003). An evaluation on feature selection for text clustering. In Proceedings of the twentieth International Conference on Machine Learning (ICML-2003), Washington DC.

Liu, B., Hu, M., & Cheng, J. (2005, May). Opinion observer: analyzing and comparing opinions on the web. In Proceedings of the 14th international conference on world wide web (pp. 342–351). ACM.

Liu, Y., Loh, H. T., & Sun, Aixin (2009). Imbalanced text classification: A term weighting approach. *Experts Systems with Applications, 36*, 690–701.

Luo, H. F., Wu, G., & Yang, J. S. (2006). Way of text classification based on Bayes. *Computer Engineering and Design, 24*, 039.

Naderalvojoud, B., Bozkir, A. S., & Sezer, E. A. (2014). Investigation of term weighting schemes in classification of imbalanced texts. In European conference data mining 2014 and international conferences intelligent systems and agents 2014 and theory and practice in modern computing.

Nakagawa, T., Inui, K., & Kurohashi, S. (2010, June). Dependency tree-based sentiment classification using CRFs with hidden variables. In Human language technologies: The 2010 annual conference of the North American chapter of the Association for Computational Linguistics (pp. 786–794). Association for Computational Linguistics.

Ng, H. W., Goh, W. B., & Low, K. L. (1997). Feature selection perceptron learning and a usability case study for text categorization. In Proceeding SIGIR '97 proceedings of the 20th annual international ACM SIGIR conference on research and development in information retrieval (pp. 67–73).

O'Keefe, T., & Koprinska, I. (2009, December). Feature selection and weighting methods in sentiment analysis. In Proceedings of the 14th Australasian document computing symposium, Sydney (pp. 67–74).

Pang, B., & Lee, L. (2004, July). A sentimental education: Sentiment analysis using subjectivity summarization based on minimum cuts. In Proceedings of the 42nd annual meeting on Association for Computational Linguistics (p. 271). Association for Computational Linguistics.

Pang, B., & Lee, L. (2005, June). Seeing stars: Exploiting class relationships for sentiment categorization with respect to rating scales. In Proceedings of the 43rd annual meeting on Association for Computational Linguistics (pp. 115–124). Association for Computational Linguistics.

Pepin, L., Kuntz, P., Blanchard, J., Guillet, F., & Suignard, P. (2017). Visual analytics for exploring topic long-term evolution and detecting weak signal in company targeted tweets. *Computers & Industrial Engineering, 112*, 450–458.

Pontiki, M., Galanis, D., Papageorgiou, H., Androutsopoulos, I., Manandhar, S., AL-Smadi, M., & Hoste, V. (2016). SemEval-2016 task 5: Aspect based sentiment analysis. In ProWorkshop on Semantic Evaluation (SemEval-2016) (pp. 19–30). Association for Computational Linguistics.

Powers, D. M. W. (2011). Evaluation: From precision, recall and F-measure to ROC, informedness, markedness & correlation. *Journal of Machine Learning Technologies. 2*(1), 37–63.

Robertson, S. (2004). Understanding inverse document frequency: On theoretical arguments for IDF. *Journal of Documentation. 60*(5), 503–520.

Rushdi Saleh, M., Martin-Valdivia, M. T., Montejo-Raez, A., & Urena-Lopez, L. A. (2011). Experiments with SVM to classify opinions in different domains. *Expert System with Application, 38*(12), 14799–14804.

Sabbah, T., Selamat, A., Selamat, M. H., Ibrahim, R., Al-Anzi, Fawaz S., Viedma, E. H., ... Fujita, H. (2017). Modified frequency-based term weighting schemes for text classification. *Applied Soft Computing, 58*, 193–206.

Sabbah, T., Selamat, A., Selamat, M. H., Ibrahim, R., & Fujita, H. (2015). Hybridized term weighting method for Web contents classification using SVM. *Neuro Computing, 173*, 1908–1926.

Salton, G., & Buckley, C. (1988). Term-weighting approaches in automatic text retrieval. *Information Processing and Management, 24*(5), 513–523.

Sanderson, M., & Ruthven, I. (1996, November). Report on the Glasgow IR group (glair4) submission. In Proceedings of the Fifth Text Retrieval Conference (TREC-5) (pp. 517–520).

Socher, R., Pennington, J., Huang, E. H., Ng., A. Y., & Manning, C. D. (2011). Semi-supervised recursive autoencoders for predicting sentiment distribution. In Proceedings of the conference on empirical methods in natural language processing (pp. 151–161). Association for Computational Linguistics.

Socher, R., Huval, B., Manning, C. D., & Ng, A. Y. (2012). *Empirical Methods in Natural.* Language Processing.

Tan, S., & Zhang, J. (2008). An empirical study of sentiment analysis for Chinese documents. *Expert Systems with Applications, 34*(4), 2622–2629.

Thangairulappan, K., & Kanagavel, A. D. (2016). Improved term weighting technique for Automatic Web page classification. *Journal of Intelligent Learning Systems and Applications, 8*(04), 63.

Vapnik, V. (1998). *Statistical learning theory*. New York: Wiley.

Wiener, E., Pedersen, J. O., & Weigend, A. S. (1995, April). A neural network approach to topic spotting. In: Proceedings of SDAIR-95, 4th annual symposium on document analysis and information retrieval (Vol. 317, p. 332).

Wu, M., Wang, L., Li, M., & Long, H. (2014). An approach of product usability evaluation based on Web mining in feature fatigue analysis. *Computers & Industrial Engineering, 75*, 230–238.

Xanthopoulos, P., & Razzaghi, T. (2014). A weighted support vector machine method for control chart pattern recognition. *Computers & Industrial Engineering, 70*, 134–149.

Xu Y., Wang B., Li J., & Jing H. (2008) An extended document frequency metric for feature selection in text categorization. In H. Li, T. Liu, W.Y. Ma, T. Sakai, K.F. Wong, G. Zhou (eds) Information retrieval technology. AIRS 2008. Lecture notes in computer science, vol 4993. Springer, Berlin, Heidelberg.

Yang, Y., & Liu, X. (1999, August). A re-examination of text categorization methods. In Proceedings of the 22nd annual international ACM SIGIR conference on research and development in information retrieval (pp. 42–49). ACM.

Yang, Y., & Pedersen, J. O. (1997, July). A comparative study on feature selection in text categorization. In ICML (Vol. 97, pp. 412–420).

Yi, J., Yang, G., & Wan, J. (2016). Category discrimination based feature selection algorithm in Chinese text classification. *Journal of Information Science and Engineering, 32*, 1145–1159.

Yoon, J., Seo, W., Coh, B. Y., Song, I., & Lee, J. M. (2017). Identifying product opportunities using collaborative filtering- based patent analysis. *Computers & Industrial Engineering, 107*, 376–387.

Yousefpour, A., Ibrahim, R., & Hamed, H. N. A. (2017). Ordinal based and frequency-based integration of feature selection methods for sentiment analysis. *Expert Systems with Applications. 75*, 80–93.

Yu, F., & Jiang, Y. F. (2004). A feature selection method for NB-based classifier. *Acta Scientiarum Naturalium Universitatis Sunyatseni. 43*(5), 118–120.

Zaghloul, W., Lee, S. M., & Trimi, S. (2009). Text classification: Neural networks vs. support vector machines. *Industrial Management & Data Systems. 109*(5), 708–717.

Zhan, J., Loh, H. T., & Liu, Y. (2009). Gather customer concerns from online product reviews – A text summarization approach. *Expert Systems with Applications, 36*(2), 2107–2115.

Zhang, W., Yoshida, T., & Tang, X. (2011). A comparative study of TF* IDF, LSI and multi-words for text classification. *Expert Systems with Applications, 38*(3), 2758–2765.

Zhao, J., Zhang, K., & Wan, J. (2013). Research of feature selection for text clustering based on cloud model. *Journal of Software, 8*(12), 3246–3252.

Zheng, R., Li, J., Chen, H., & Huang, Z. (2006). A framework for authorship identification of online messages: Writing style features and classification techniques. *Journal of the Association for Information Science and Technology, 57*(3), 378–393.

Zong, W., Wu, F., Chu, L. K., & Sculli, D. (2015). A discriminative and semantic feature selection method for text categorization. *International Journal of Production Economics, 165*, 215–222.