Turkish Journal of Electrical Engineering & Computer Sciences

http://journals.tubitak.gov.tr/elektrik/

Research Article

Turk J Elec Eng & Comp Sci
(2018) 26: 2570 – 2582

doi:10.3906/elk-1704-178

Extended correlated principal component analysis with SVM-PUK in opinion mining

Kollimarla Anusha DEVI, Deepak Chowdary EDARA*, Venkatrama Phani Kumar SISTLA, Venkata Krishna Kishore KOLLI
Department of CSE, VFSTR University, Guntur, India

Received: 15.04.2017 • **Accepted/Published Online:** 18.06.2018 • **Final Version:** 28.09.2018

Abstract: With the rapid growth of microblogs and online sites, an inordinate number of product reviews are available on the Internet. They not only help in analyzing, but also assist in making informed decisions about product quality. In the proposed work, an extended correlated principal component analysis (ECPCA) is used for dimensionality reduction. A comparative analysis is conducted on movie reviews (DB-1) and Twitter datasets (DB-2 and DB-3) in opinion mining extraction. The performance of naïve Bayes, CHIRP, and support vector machine (SVM) with kernel methods such as radial basis function (RBF), polynomial, and Pearson (PUK) are compared and analyzed on the three datasets. The experimental results using ECPCA for selecting relevant features and SVM-PUK as a classifier exhibit better performance on movie reviews and Twitter datasets. The performance of the proposed approach is 99.69%, 99.4%, and 99.54% on the DB-1, DB-2, and DB-3 datasets, respectively, and comparatively outperforms the existing methods.

Key words: Opinion mining, latent Dirichlet allocation, principal component analysis, dimensionality reduction, support vector machine

1. Introduction

With the evolution of Web 2.0, a huge amount of product reviews is circulating on the Web. People are interested in gathering online reviews before making the decision to purchase a product. From these reviews, customers can gather high-quality judgments regarding product information and accordingly regulate their purchase actions [1]. Concurrently, manufacturers can get prompt feedback about their product and enhance the quality of the product. An increasing number of studies are concentrating on expressing the relationship between customer satisfaction and purchase rate of products. Opinion mining [2] came into existence to compute overall demand for a product in the market. It is a subdisciplinary area of data mining and refers to extracting, allocating, and assessing opinions being expressed in multiple online stores, comments from social media sites, and information generated by customers.

Nowadays, research is mostly carried out across online movie reviews [3] and Twitter data [4], but there are many challenges in modelling and extracting those review opinions. It is a highly tedious process to find out opinions from unstructured data that have several emotions and diplomatic opinions about a product. In general, when a potential buyer reads such kinds of diplomatic reviews, he/she cannot estimate the product's value in terms of positive and negative aspects. To overcome such situations, words that reflect opinion polarity are detected, and then the overall polarity of those diplomatic statements is computed. In certain conditions, online

*Correspondence: phani_cse@vignanuniversity.org

reviews may contain some harsh statements about a product. In this work, analysis of such harsh statements will be conducted by extracting word tokens for classifying opinions. Usage of obscure factors yields clear and definite opinions about products so as to address the limitations of the two-class text classification problem. The other sections of the paper are as follows. A detailed study of the literature and significant observations in relevance to the proposed objective is stated in section 2, the synopsis of the problem is described in section 3, methods and materials are stated in section 4, and the experimental results and discussions are presented in section 5.

2. Related work

In this section, a detailed survey of various approaches used for the extraction of user opinions, topic modeling, feature extraction, and selection is presented. The machine learning field has the provision of several approaches to deal with the issues in opinion mining.

2.1. Survey on opinion mining analysis

Researchers [5,6] have developed a direct approach called an opinion search engine to retrieve product reviews. This approach gives more priority to sentiment implication, which is based on opinion words. This further reveals the features of classified reviews and semantic orientations. These kinds of techniques deal with issues like data sparsity, which may lead to noisy and unstructured data [7]. In [8], the author proposed a framework for detecting polarity words based on word sense disambiguation using WordNet. When all the words are authorized, the polarity of the sentence is detected with the help of the rule-based classifier. Another author [9] characterized the polarity of movie reviews (IMDb) by means of three machine learning algorithms: naïve Bayes, SVM, and maximum entropy. He also conducted an in-depth study of more than 300 scholarly articles describing the common problems faced in two major tasks, namely sentiment analysis and opinion mining with polarity. Omar et al. [10] carried out a study on reducing the features in the dataset by selecting only the appropriate features for further classification. The main aim of their study was to reduce the workload of the classifier using a feature selection strategy. Classification with a feature selection strategy demonstrates incomparable accuracy. Several techniques were widely used to extract opinions and sentiments from various blogs and are proposed in [11]. In [12], probabilistic latent semantic analysis and latent Dirichlet allocation were considered in analyzing topic distributions, which were based on variables ranging from topic words to documents. A topic model describes latent topics from each document, and this process is defined by the allocation of top words in a document. Hu and Liu [13] proposed an association mining approach to extract product features. The main aim of their methodology was related to the common words used by people when they commented on different product features. In [14], the authors introduced the first deep learning approach toward extracting aspects from the collected opinion corpus. They used a 7-layer deep convoluted neural network for tagging each word in the opinion sentences by developing a set of linguistic patterns. The experiment was carried out by extracting the aspect terms from the opinions with the proposed method and then the integration of linguistic patterns was performed with a deep learning classifier. In [15], opinions in the form of quotes were collected from different newspapers such as The Hindu, Deccan Chronicle, and The Times of India, and a comparative study was conducted on these three dailies based on a particular issue. The methodology labeled the newspapers as positive, negative, or neutral by exploring the objectivity and subjectivity of quotes. A supervised learning technique, i.e. SVM, was used for the classification of opinions. Sahu and Ahuja [16] explored a process of deriving opinions and extracting emotions about a topic from unstructured, structured, or

semistructured textual data. He stressed the importance of the preprocessing step, which includes stemming, stop-word removal, and POS tagging to extract useful features in opinion mining. Porter's stemming algorithm was used for the elimination of redundancies held in the text by trimming the words and obtaining the root word. Claypo and Jaiyen [17] proposed unsupervised learning with K-Means and MRF for feature selection to extract opinions from reviews of a Thai restaurant. This study improved the business of the restaurant by making use of customers' feedback on services and products offered. In the preprocessing phase, review sentences were transformed into words by eliminating the stop words, which do not offer any specific meaning. Bouazizi and Ohtsuki [18] described a sentimental approach with 13 features to evaluate tweet polarity. They explored sentiment, punctuation, syntactic, and pattern features to detect sarcastic tweets for automatic classification. Chen et al. [19] applied SVM for the classification of opinions on Twitter and an online movie review corpus. Li et al. [20] explored a methodology, named customer voice sensor (CVS), to analyze and generate opinions from call center conversations. Their study analyzed the issues faced by customers with regard to the product and whether those issues were resolved. The proposed model consisted of two parties: a state-of-the-art SVM technique was used to detect the attitudes of the parties and the topic is extracted from a conversation, which includes call type and intention of caller. The proposed method was evaluated on the mobile reviews corpus, and the results have proven its effectiveness compared with other state-of-the-art techniques. Zhang et al. [21] discussed trust models based on fame that are extensively used in e-commerce sites like eBay and Amazon. It is clearly evident that the buyers who bought items and provided high ratings can also express their opinions in a negative way. All such ratings are summated to define the seller's reputation, which is taken into account while deciding to buy a product. Zhang proposed a system called "CommTrust" to compute the score of reputation from the feedback provided by the user. They made use of techniques such as natural language processing, topic modeling, and opinion mining for reducing the dimension ratings and weights.

3. Problem synopsis

In order to resolve the issues described in the above literature, the proposed model serves as a novel approach for identifying trust in an honest manner. In this paper, various machine learning algorithms were used to classify the reviewed data, and the opinion polarity was estimated as either positive or negative. While estimating opinion polarity along with feature extraction, several obstacles were encountered. To overcome these obstacles, a statistical model known as LDA topic modelling was incorporated in the present paper. The features extracted were further analyzed for discarding irrelevant features or terms by employing the extended correlated principal component analysis (ECPCA) feature reduction approach.

Organization of the proposed work is:

Input: A reviewed dataset R with a set of r training instances.

Output: An anticipated model.

1. Initially, a review corpus was prepared by performing preprocessing.
2. Measures such as term frequency and inverse document frequency (TF-IDF) were computed on a dataset to generate opinion polarity.
3. A probabilistic LDA was used for converging the text data to form topics.
4. The features were extracted based on probability computed on distributed topics.

5. The irrelevant features were eliminated by applying the ECPCA method for reducing features on distributed dimensions.
6. The efficiency of the proposed methodology was measured by comparing it to various machine learning algorithms with the help of evaluation metrics.

4. Methods and materials

With reference to the literature mentioned in section 2, opinion mining took place by way of analyzing the movie reviews dataset (DB-1) [22] and Twitter dataset (DB-2 and DB-3 (SEM Eval 2016, Task-4, Subtask-B)) [23,24]. Movie reviews are a good example of overall feedback given by a user regarding a particular movie. From the reviews, a user can make a good decision about the movie before watching it. It is even useful to the team of moviemakers to know about the flaws and positives present in the film. Microblogging service providers such as Twitter also act as a good source to mine the opinions of users. Due to its active usage, many companies are showing a keen interest in knowing about their services and extracting product opinions from users. In this section, we define the proposed framework in Figure 1, which is a combination of two important activities: opinion analysis and construction of useful knowledge.

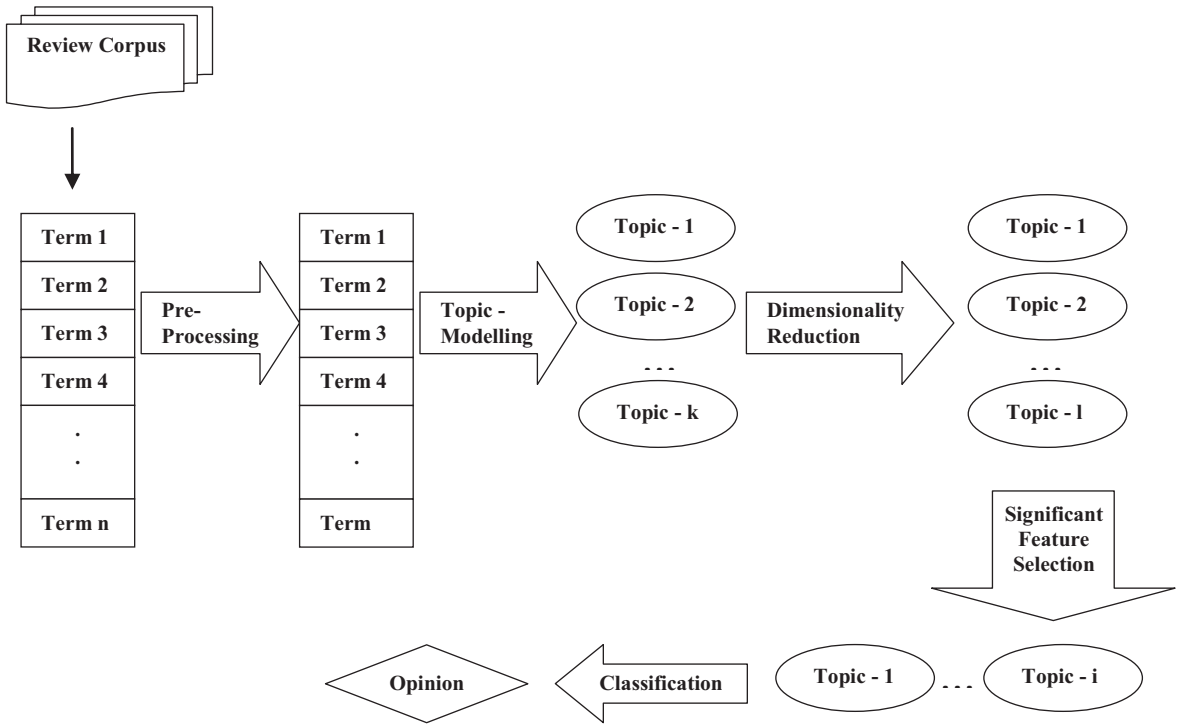

Figure 1. Architecture of the proposed framework.

4.1. Architectural flow of the proposed framework

The proposed framework starts with the data collected from various web sources. Preprocessing of data is a key task that includes certain functions such as conversion of words to lower case, stop-word removal, and punctuation removal. The basic advantage of the preprocessing phase is to diminish the size of data. Generally, a text corpus is unpolished in nature, and it is highly necessary to preprocess the data before transferring it to

further stages of analysis. An example of a movie review with various preprocessing techniques applied to it is represented in Table 1. After preprocessing, opinion words are extracted by word tokenization to define the polarity of the review by computing measures such as TF-IDF. These measures evaluate how often a word is in the document. They increase with the number of times the word appears in a document. With regard to the TF-IDF factor, the superiority of words in a review is analyzed, and the polarity of the review is predicted, i.e. positive or negative. The key terms of the data are then extracted using LDA topic modeling from the preprocessed corpus and modeled into a number of dimensions. These modeled topics are considered with probabilities as the number of dimensions based on the term score. The proposed dimensionality reduction technique finds out the significant features from a large number of dimensions based on the term probability. The dimensions with low term score are eliminated by calculating their mean, median, and variance. Finally, the important features are generated and classified on the basis of reduced feature dimensions.

Table 1. An example of data reduction with preprocessing techniques on the corpus.

Technique	Preprocessed Review
Actual review	The Rock is destined to be the 21st centurys new conan and that he's going to make a splash even greater than Arnold Schwarzenegger jean claud van damme or Stevensegal.
Removal of special characters	the rock is destined to be the stcenturys new conan and that hes going to make a splash even greater than arnold schwarzenegger jeanclaud van damme or stevensegal
Removal of stop-words	rock destined centurys new conan make splash greater arnoldschwarzenegger-jeanclaud van dammestevensegal
Removal of whitespaces	rock destined centurys new conan make splash greater arnold schwarzenegger jeanclaud van dammestevensegal
Word tokenization	"rock" "destin" "centuri" "new" "conan" "make" "splash" "greater" "arnold" "schwarzenegg""jeanclaud" "van" "damm""steven" "segal"

4.2. An approach for constructing essential knowledge using LDA

To realize the proposed objective, a probabilistic model based on LDA was employed in our approach. LDA uses Gibb's sampling method for mapping the text corpus into lower dimensional space by implementing topic modeling. A word is the basic component of continuous data and w is a vector of words. In the vocabulary, the mth word is denoted as $w_m = \begin{cases} 1 \\ 0 \ \& m \neq n \end{cases}$

A document d is a collection of n words, where d = $\{w_1, w_2, w_3 \ldots w_n\}$; the text corpus consists of m documents, where D = $\{d_1, d_2, d_3 \ldots d_m\}$. $\propto_k$ is the prior weight of topic k in the document, and β_k is the prior weight of the word in a topic k. W is the total no. of words in the document. $W = \sum_{i=1}^{m} W_i$, and W_i describes number of words in the ith document. LDA is a generative probabilistic model and it transforms the document D as random mixtures of latent topics.

$$Choose \ \ \theta_i \sim \ Dir\,(\alpha)\,, \text{where}\, i = \{1,\ 2,\ 3 \ldots m\} \tag{1}$$

$$\varphi_i \sim \ Dir\,(\beta)\,, \ where\, i = \{1,\ 2,\ 3 \ldots m\} \tag{2}$$

For each word in the document, choose a topic $t_i \sim$ Multinomial(θ_i) and a word in topic with $w_i \sim$ Multinomial(φ_i). To compute the distributions θ and φ, expectation maximization and Gibbs sampling were used. In the proposed work, topic-wise probability distributions were calculated using Gibbs sampling as shown in Eq. (3).

$$p\left(\theta/_{\varphi}\right) = p\left(z_i = k/_{w_i = v, z_{-i}, w_{-i}}\right) \propto \varphi_{wt}\theta_{dt}, \tag{3}$$

where $t_i = k$, i.e. k is the topic of the ith term in a document.$w_i = v$, which symbolizes that the required term is the vth term in the text corpus vocabulary, and z_{-i} denotes all the topic assignments other than the ith term. The document-topic distribution θ and topic-word distribution φ are estimated using Eqs. (4) and (5). N_{dt} is the number of times topic t has occurred in a document d and T is the total no. of topics. N_{wt} is the number of times the term is assigned to topic t. W is the size of the vocabulary or total no. of words.

$$\theta_{dt} = \frac{N_{dt} + \alpha}{\sum_{t} N_{dt} + T\alpha} \tag{4}$$

$$\varphi_{wt} = \frac{N_{wt} + \beta}{\sum_{t} N_{dt} + W\beta} \tag{5}$$

Topic modeling is used to discover topics from text corpus of opinion mining and distribute them into latent topics. It improves the efficiency factor to predict the tasks related to NLP. During the process of knowledge construction, there may be some terms in reviews that may not be a stop-word. Examples of such terms are "a", "and", "the", "do", "don't", "there", "there's", "hers", "etc." The presence of such words in the corpus does not carry any useful information and so they have to be removed before converging them to form topics. In LDA, it is necessary to predefine the number of topics and terms to model. Suppose $T = 50$ defines that the number of topics to be modeled is 50. Incorporation of topic modeling leads to extraction of the key features from the distributed topics based on the probabilities associated with each term in the modeled topics.

4.3. Extended correlated principal component analysis (ECPCA)

To reduce the dimensionality of the feature set, the ECPCA method was employed in the proposed model.

Algorithmic steps of ECPCA:

i) Let Y be a data matrix with n documents and m topics:

ii) Evaluate covariance matrix,

$$C_{i,j} = \frac{1}{M-1} \sum_{t=1}^{M} Y_{t,i} Y_{t,j} \tag{6}$$

$C_{i,i}$ (diagonal) defines the variance of a variable i. $C_{i,j}$ (off-diagonal) defines the covariance existing between i and j.

iii) Compute eigenvectors and eigenvalues of covariance matrix $C_{i,i}$ and then δ is the threshold to choose λ_k significant eigenvalues, which is also used to reduce the dimensionality of the data.

$$\max(\lambda_k) > \delta \tag{7}$$

iv) Evaluate a standardized transformation matrix T with domain features

v) $T = \{t_1, t_2, t_3 \ldots t_k\}$ is the set of all k-original topics, and $ST = \{st_1, st_2, st_3 \ldots st_k\}$ is the k-shadow topics of T. The feature pattern of an original topic is randomly rearranged to form a shadow topic.

vi) Topics of both T and ST are integrated and they are straggled in correlation.

vii) Significance score is computed for the topics of elongated sequence of T and ST at each iteration by random forest method and those scores are normalized with the Z-score method.

$$T_z = \frac{w - \bar{w}}{\sigma} \tag{8}$$

viii) $RT = \{rt_1, rt_2, rt_3 \ldots rt_k\}$ is a set of relevant topics, where a topic is said to be relevant only if its score is greater than the maximum Z-score values of ST. The set of all topics having less significance than maximum Z-score values of ST are treated as irrelevant and are permanently eliminated from T.

$$T = \begin{cases} RT\ T_z \geq max?(ST_z) \\ IRTT_z < max?(ST_z) \end{cases} \tag{9}$$

ix) The above steps are iteratively applied until the significance is computed for T.

x) Finally, eliminate all the shadow topics.

In the proposed work, the ECPCA technique is employed, which aims at high dimensionality reduction for extracting essential knowledge from the corpus.

4.4. Opinion classification using the SVM-PUK algorithm

The focus of the feature reduction mechanism is to eliminate irrelevant features, reduce computational cost, and enhance classification efficiency. It has been proven that utilization of the proposed framework made the classification task easier as it reduced the extracted features and selected the relevant ones. In this work, the corpus analyzed consisted of Twitter and movie reviews dataset.

4.4.1. Emphasis on SVM technique with Pearson (PUK) kernel

SVM is a category of supervised learning used to inspect the data for classification and regression. SVM acts as a tool for insolvency evaluation if there is nonregularity in data by providing solutions for two-class problems. It performs linear classification by representing the points in space with a gap demonstrating that the points are of discriminating categories. Nonlinear classification is also achieved with SVM using RBF, polynomial, and Pearson kernels. The Pearson kernel offers an optimized classification due to its robust nature and is applicable to real world datasets. It has a vigorous mapping feature to produce the best performance among all the kernels of SVM. As such, PUK is also named a universal kernel. The general form of PUK is given as Eq. (10)

$$f(v) = \frac{l}{\left[1 + \left(\left(2\left(v - v_0\right)\sqrt{2^{1/\rho} - 1}\right) \Big/ \gamma\right)^2\right]^{\rho}} \tag{10}$$

where v is an independent variable and l is the peak height of the curve at the center x_0. The variables γ and ρ indicate the width and the tail of the curve. A kernel function is said to be valid only if its corresponding kernel matrix is positive, semide?nite, and symmetric.

5. Experimental results and discussion

To evaluate the performance of the proposed method, further analysis was conducted on DB-1 [22], DB-2 [23], and DB-3 [24], which consist of 10,662, 15,219, and 10,551 positive and negative reviews, respectively.

5.1. Analysis on impact of feature selection techniques on DB-1, DB-2, and DB-3

Before preprocessing, the data with total numbers of reviews present in DB-1 were 10,662 with 124,079 terms. In the preprocessing phase, the duplicate reviews, which do not provide any knowledge, were eliminated, and the total reviews were reduced to 10,066 with 19,091 terms. After preprocessing, the total terms were reduced to 735 by eliminating all the stop-words and sparse terms and then considered as the top terms for DB-1. The total numbers of reviews before preprocessing in DB-2 were 15,219 with 179,088 terms. They were reduced to 13,007 reviews with 18,192 after preprocessing. For DB-2, after removing stop-words and sparse terms, the terms were reduced to 345 top terms. Similarly, the total number of reviews before preprocessing in DB-3 were 10,551 with 136,402 terms. Finally, they were reduced to 10,207 reviews with 14,343 terms. After preprocessing they were reduced to 614 terms by removing stop-words and sparse terms. The reduced top terms in each dataset are shown in Figure 2; 50 topics were considered as final and also provided as input to LDA. Each topic in LDA is associated to each review in the corpus, with the top terms in the review based on term score and then considered as a dimension.

For DB-1, naïve Bayes and SVM classifiers achieved an accuracy of 53.04% and 50.67% on 50 dimensions without any dimensionality reduction technique. For improvement in accuracy, dimensionality reduction was applied with PCA and then condensed into 30 dimensions based on the term probability importance. Consequently, accuracy rates of 97.01% and 99.54% were obtained with the help of naïve Bayes and SVM classifiers, respectively. Similarly, the same procedure was performed on DB-2, and 47.02% and 53.46% accuracies were obtained without dimensionality reduction. The dimensions were reduced to 30 with PCA with an accuracy rate of 91.33% and 99.2%. Further, performance evaluation was carried out on DB-3, and obtained accuracies as 57.91% and 62.73% and then dimensions were reduced to 27 with PCA with an accuracy rate of 95.31% and 99.24%, respectively. The proposed correlated feature selection algorithm was applied to improve the classification accuracy by reducing more dimensions to 25 on DB-1 with an accuracy rate of 51.82% and 50.43% and 33 dimensions on DB-2 with an accuracy rate of 49.41% and 52.29% on DB-3 with an accuracy rate of 57.46% and 73.52% with naïve Bayes and SVM classifiers, respectively. Later, feature selection is done with linear discriminant analysis and extracted two dimensions; naïve Bayes and SVM classifier achieved 98.56% and 99.43% accuracy on DB-1, 86.82% and 99.18% on DB-2, and 96.94% and 99.4% on DB-3, respectively. Finally, to improve the classification accuracy, an innovative approach of ECPCA was incorporated and showed improvements by reducing more dimensions to 15 with an improved accuracy rate of 98.18% and 99.60% on DB-1, 93.66% and 99.38% on DB-2 with 25 dimensions, and to 21 dimensions with an accuracy of 95.33% and 99.54% using naïve Bayes and SVM models.

The analysis of opinion mining validated using 10-Fold-CV, which consisted of all reviews labeled as either positive or negative, concludes that the proposed ECPCA clearly demonstrates the need for dimensionality reduction and it outperforms the PCA and LDA approaches. Figure 3 illustrate the dimensionality reduction achieved with the proposed approach.

Figure 2. (a) Positive Cloud on DB-1; (b) Negative Cloud on DB-1; (c) Positive Cloud on DB-2; (d) Negative Cloud on DB-2; (e) Positive Cloud on DB-3; (f) Negative Cloud on DB-3.

5.2. Performance evaluation and comparative analysis of the proposed approach with other classification methods

Several models such as naïve Bayes, CHIRP, KNN, and SVM with kernel functions such as RBF, polynomial, and Pearson exist for classifying opinions. The performance of each classifier with their confusion matrix values on DB-1 are tabulated in Table 2. The results of DB-1 with the KNN classifier, CHIRP classifier, and naïve Bayes classifier were 95.47%, 98.39% and 98.18%, respectively. The accuracies obtained when experimenting with

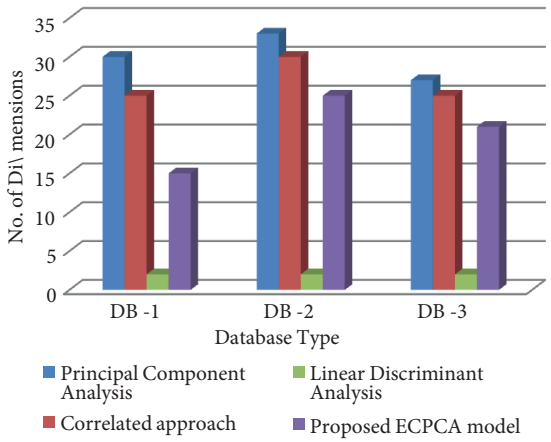

Figure 3. No. of features extracted with various dimensionality reduction methods.

Table 2. Comparison of classifiers with DB-1.

Classifier	Precision	Recall	F-Measure	Accuracy
KNN	0.953	0.957	0.955	95.47%
CHIRP	0.982	0.984	0.983	98.39%
NB	0.981	0.983	0.982	98.18%
SVM Poly	0.996	0.996	0.996	99.60%
SVM-RBF	0.995	0.996	0.996	99.60%
SVM-PUK	0.997	0.997	0.997	99.69%

SVM kernel functions such as polynomial and RBF were 99.60% and 99.60%, respectively. SVM-PUK yielded an accuracy of 99.69%. In Table 3, the performance results obtained with regard to DB-2 are clearly visualized with their confusion matrices. The results of DB-2 with the KNN classifier, CHIRP classifier, and naïve Bayes classifier were 96.10%, 98.32%, and 93.66%, respectively. The accuracies obtained when experimenting with SVM kernel functions such as polynomial and RBF were 99.38% and 99.23%, respectively. The results of DB-3 with the KNN classifier, CHIRP classifier, and naïve Bayes classifier were 95.05%, 98.11%, and 95.33%, respectively, as shown in Table 4. The accuracies obtained when experimenting with SVM kernel functions such as polynomial and RBF were 99.54% and 99.41%, respectively. The proposed SVM-PUK achieved a better accuracy rate of 99.54% when compared with other existing methods. Figure 4 presents the performance of the proposed method on DB-1, DB-2, and DB-3 with various classification models using ROC curves.

Table 3. Comparison of classifiers with DB-2.

Classifier	Precision	Recall	F-Measure	Accuracy
KNN	0.961	0.961	0.961	96.10%
CHIRP	0.984	0.982	0.983	98.32%
NB	0.936	0.938	0.937	93.66%
SVM Poly	0.994	0.994	0.994	99.38%
SVM-RBF	0.992	0.992	0.992	99.23%
SVM-PUK	0.994	0.994	0.994	99.43%

Table 4. Comparison of classifiers with DB-3.

Classifier	Precision	Recall	F-Measure	Accuracy
KNN	0.966	0.967	0.966	95.05%
CHIRP	0.986	0.988	0.987	98.11%
NB	0.953	0.953	0.953	95.33%
SVM Poly	0.995	0.995	0.995	99.54%
SVM-RBF	0.987	0.991	0.989	99.41%
SVM-PUK	0.995	0.995	0.995	99.54%

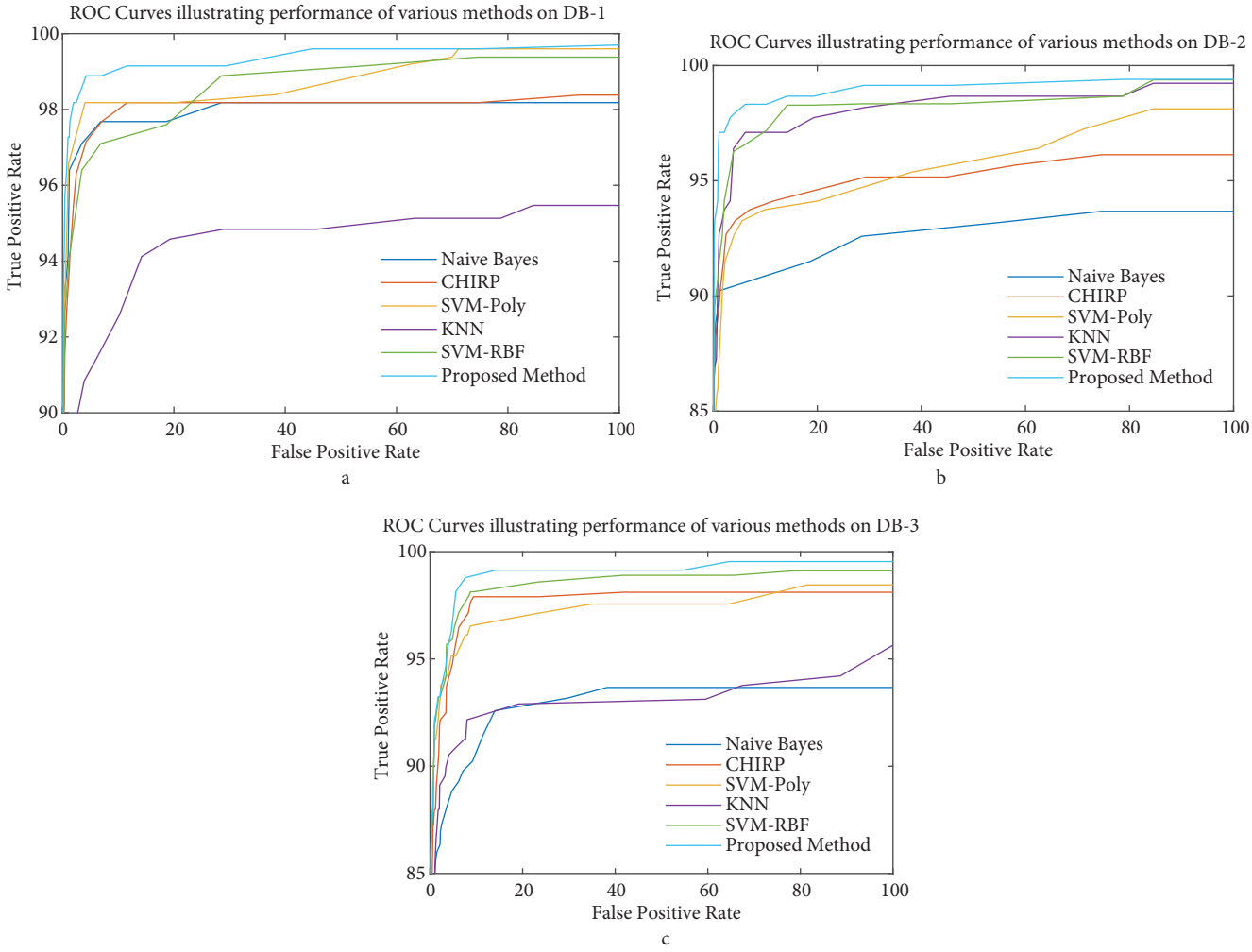

Figure 4. ROC curves illustrating performance comparison of proposed method with various methods.

6. Conclusion

The proposed work is concerned with conducting opinion mining analysis and constructing essential knowledge from it. The feature extraction process is achieved by a probabilistic model, i.e. LDA topic modeling. We proposed ECPCA as a novel approach for efficient feature selection aiming for a high dimensionality feature

reduction along with SVM-PUK as a classification technique. Due to the incorporation of the proposed framework, classification performance in opinion mining analysis is enhanced. In the proposed work, a comparative experimental study was carried out on movie reviews and Twitter datasets. Classifiers such as CHIRP, naïve Bayes, and SVM with kernel methods (RBF, polynomial, and Pearson) were applied to the datasets. The results signify that feature selection by employing the proposed framework along with SVM-PUK outperformed the other techniques.

References

[1] Lin C, He Y, Everson R, Ruger S. Weakly supervised joint sentiment-topic detection from text. IEEE T Knowl Data En 2012; 24: 1134-1145.

[2] Cambria E, Schuller B, Xia Y, Havasi C. New avenues in opinion mining and sentiment analysis. IEEE Intell Syst 2013; 28: 15-21.

[3] Zhu J, Wang H, Zhu M, Tsou B K, Ma M. Aspect-based opinion polling from customer reviews. IEEE T Affective Computing 2011; 2: 37-49.

[4] Jain AP, Katkar VD. Sentiments analysis of Twitter data using data mining. In: 2015 International Conference on Information Processing (ICIP); 16 Dec 2015, USA: IEEE, pp. 807-810.

[5] Blei DM, Ng AY, Jordan MI. Latent dirichlet allocation. J Mach Learn Res 2003; 3: 993-1022.

[6] Eirinaki M, Pisal S, Singh J. Feature-based opinion mining and ranking. J Comput Syst Sci 2012; 78: 1175-1184.

[7] Go A, Bhayani R, Huang L. Twitter sentiment classification using distant supervision. CS224N Project Report, Stanford, 2009.

[8] Ortega R, Fonseca A, Montoyo A. SSA-UO: unsupervised Twitter sentiment analysis. In: Second Joint Conference on Lexical and Computational Semantics; 13–14 June 2013; Atlanta, Georgia, USA: ACL. pp. 501-507.

[9] Pang B, Lee L. Opinion mining and sentiment analysis. In: Foundations and Trends. Information Retrieval. USA: ACM, 2008.

[10] Zhang W, Yu C, Meng W. Opinion retrieval from blogs. In: Sixteenth Conference on Information and Knowledge Management; 6–10 Nov 2007; Lisbon, Portugal: ACM. pp. 831-840.

[11] Ikram M, Butt N, Afzal M. Open source software adoption evaluation through feature level sentiment analysis using Twitter data. Turk J Elec Eng & Comp Sci 2016; 24: 4481-4496.

[12] Hofmann T. Probabilistic latent semantic indexing. In: 22nd annual International SIGIR Conference on Research and Development in Information Retrieval; 15–19 Aug 1999; Berkeley, CA, USA: ACM. pp. 50-57.

[13] Hu M, Liu B. Mining opinion features in customer reviews. In: AAAI; 25–29 July 2004; San Jose, CA, USA: IEEE. pp. 755-760.

[14] Poria S, Cambria E, Gelbukh A. Aspect extraction for opinion mining with a deep convolutional neural network. Knowl-Based Syst 2016; 108: 42-49.

[15] Ahmed S, Danti A. A novel approach for Sentimental Analysis and Opinion Mining based on SentiWordNet using web data. In: International Conference on Trends in Automation, Communications and Computing Technology; 21–22 December 2015; Bangalore, India: IEEE. pp. 1-5.

[16] Sahu T, Ahuja S. Sentiment analysis of movie reviews: a study on feature selection & classification algorithms. In: International Conference on Microelectronics, Computing & Communications; 23–25 January 2016; Durgapur, India: IEEE, pp.1-6.

[17] Claypo N, Jaiyen S. Opinion mining for Thai restaurant reviews using K-Means clustering and MRF feature selection. In: 7th International Conference on Knowledge and Smart Technology; 28–31 January 2015; Chonburi, Thailand: IEEE. pp. 105-108.

[18] Bouazizi M, Ohtsuki T. Opinion mining in Twitter: how to make use of sarcasm to enhance sentiment analysis. In: 2015 International Conference on Advances in Social Networks Analysis and Mining; 25–28 August 2015; Paris, France: IEEE. pp. 1594-1597.

[19] Chen C, Chen Z, Wu C. An Unsupervised approach for person name bipolarization using principal component analysis. IEEE T Knowl Data En 2012; 24: 1963-1976.

[20] Li P, Yan Y, Wang C, Ren Z, Cong P, Wang H, Feng J. Customer voice sensor: a comprehensive opinion mining system for call center conversation. In: 2016 IEEE International Conference on Cloud Computing and Big Data Analysis; 5–7 July 2016; Chengdu, China: IEEE; pp. 324-329.

[21] Zhang X, Cui L, Wang Y. CommTrust: Computing Multi-Dimensional Trust by Mining E-Commerce Feedback Comments. IEEE T Knowl Data En 2014; 26: 1631-1643.

[22] Pang B, Lee L. Seeing stars: Exploiting class relationships for sentiment categorization with respect to rating scales. In: 43rd annual meeting on association for computational linguistics; 25–30 June 2005; Michigan, USA: ACM. pp. 115-124.

[23] Mukherjee S, Pushpak B. Sentiment analysis in Twitter with lightweight discourse analysis. In: *Coling*; 8–15 December 2012; Bombay, India: ACL. pp. 1847-1864.

[24] Esuli A, Sebastiani F. Sentiment quantification. IEEE Intell Syst 2010; 25: 72-75.