Received 29 August 2017
Revised 21 November 2017
Accepted 26 November 2017

Analysing user sentiment of Indian movie reviews

A probabilistic committee selection model

Shrawan Kumar Trivedi
Department of IT and Systems, Indian Institute of Management Sirmaur, Sirmaur, India, and

Shubhamoy Dey
Department of Information Systems, Indian Institute of Management Indore, Indore, India

Abstract

Purpose – To be sustainable and competitive in the current business environment, it is useful to understand users' sentiment towards products and services. This critical task can be achieved via natural language processing and machine learning classifiers. This paper aims to propose a novel probabilistic committee selection classifier (PCC) to analyse and classify the sentiment polarities of movie reviews.

Design/methodology/approach – An Indian movie review corpus is assembled for this study. Another publicly available movie review polarity corpus is also involved with regard to validating the results. The greedy stepwise search method is used to extract the features/words of the reviews. The performance of the proposed classifier is measured using different metrics, such as F-measure, false positive rate, receiver operating characteristic (ROC) curve and training time. Further, the proposed classifier is compared with other popular machine-learning classifiers, such as Bayesian, Naïve Bayes, Decision Tree (J48), Support Vector Machine and Random Forest.

Findings – The results of this study show that the proposed classifier is good at predicting the positive or negative polarity of movie reviews. Its performance accuracy and the value of the ROC curve of the PCC is found to be the most suitable of all other classifiers tested in this study. This classifier is also found to be efficient at identifying positive sentiments of reviews, where it gives low false positive rates for both the Indian Movie Review and Review Polarity corpora used in this study. The training time of the proposed classifier is found to be slightly higher than that of Bayesian, Naïve Bayes and J48.

Research limitations/implications – Only movie review sentiments written in English are considered. In addition, the proposed committee selection classifier is prepared only using the committee of probabilistic classifiers; however, other classifier committees can also be built, tested and compared with the present experiment scenario.

Practical implications – In this paper, a novel probabilistic approach is proposed and used for classifying movie reviews, and is found to be highly effective in comparison with other state-of-the-art classifiers. This classifier may be tested for different applications and may provide new insights for developers and researchers.

Social implications – The proposed PCC may be used to classify different product reviews, and hence may be beneficial to organizations to justify users' reviews about specific products or services. By using authentic positive and negative sentiments of users, the credibility of the specific product, service or event may be enhanced. PCC may also be applied to other applications, such as spam detection, blog mining, news mining and various other data-mining applications.

Originality/value – The constructed PCC is novel and was tested on Indian movie review data.

Keywords Sentiment analysis, Indian movie reviews, Machine learning classifiers, Greedy stepwise search method, Probabilistic committee selection

Paper type Technical paper

The Electronic Library
Vol. 36 No. 4, 2018
pp. 590-606

0264-0473
DOI 10.1108/EL-08-2017-0182

1. Introduction

With the rise of Web 2.0, the ways in which people share and express their thoughts have changed significantly. For instance, if a person wants to buy something, such as a phone, or wants to watch a movie, for example, he/she would typically go online and read user reviews to find a phone/movie that matches his/her expectations. This possibility has arisen because people have started sharing their feelings online. Web 2.0 has created a medium for sharing opinions, feelings and thoughts. This freedom of expression has made users more open to expressing their opinions, and for analysts has facilitated analyses of hidden patterns to predict customer attitude towards a particular product. This method is termed sentiment analysis or opinion mining (Mostafa, 2013; Ye *et al.*, 2009). Sentiment analysis entails a process of discovering opinions, emotions, feelings or attitudes from a piece of text, which is generally written by a user. The technique is universally applicable, and generally applies to assessment of customer reviews in the market domain including movie reviews.

The basic aim of sentiment analysis is to classify the polarity of text or documents, whether this is positive, negative or neutral. It also helps in classifying whether the text or phrases are subjective or objective. Classification of subjectivity and objectivity is a difficult task, however, because subjectivity depends on context and objectivity contains subjective data. Sentiment classification is a domain-specific problem (Aue and Gamon, 2005; Moraes *et al.*, 2013). In natural language processing, it is considered a special case of text classification. Text mining, natural language processing and computational linguistics methods are often used for such analysis. There are several challenges associated with sentiment analysis; for example, negative sentiments can be expressed by users without using any negative words, usually as ironic sentences or through sarcasm. Identifying sentiments behind such text is extremely difficult.

English is generally considered the most appropriate language for sentiment analysis because of its universal applicability and its wide reach in terms of usage. Machine learning (ML) classifiers are popular in such studies. Under the ML approach, data are converted into a feature vector and then used to train the ML classifier to infer a combination of specific features yielding a specific class (Pang and Lee, 2008); a model is then created to predict the class.

In this particular work, a novel committee selection method is proposed in which probabilistic classifiers (Bayesian and Naïve Bayes [NB]) are used to build a committee of classifiers. The proposed probabilistic committee classifier (PCC) is then compared with other popular ML classifiers, such as Bayesian (Ye *et al.*, 2009), Decision Tree (J48) (Wan and Gao, 2015), Random Forest (Liu and Chen, 2015) and Support Vector Machine (SVM) (Moraes *et al.*, 2013). All classifiers are tested with the help of Indian Movie Review and Movie Review Polarity corpora. The Indian Movie Review corpus was created specifically for this research. For the training of the classifiers, the greedy stepwise search method is used.

The remainder of the paper is organized as follows. Section 2 deals with related work on the sentiment analysis field. Section 3 details the corpora testing, where preparation of the dataset and all experimental design methods are discussed and the proposed PCC and other classifiers used to compare the proposed classifier are described. Section 4 outlines the results and analysis, Section 5 presents a discussion and Section 6 concludes the paper.

2. Related work

Sentiment analysis, also known as opinion mining, is carried out using text mining techniques in which sentiments of users are tracked and analysed. A plethora of research has been conducted in this area to capture the sentiments of users' opinions about products,

services, organizations, individuals and events, using various tools and techniques. ML methods are popular and well known in the sentiment- and opinion-mining domain. Some ML approaches, such as Bayesian classifier, NB, SVM, J48, Random Forest (RF), etc., have been used in experiments and have a prominent place in the literature.

Pang *et al.* (2002) built a model using NB, Maximum Entropy (ME) and SVM and tested these on a data set constructed from the IMDB website with 700 positive and 700 negative movie reviews. The accuracy of these models was found to be 77-82.9 per cent. In a study conducted by Dave *et al.* (2003), a product review data set from the Amazon website was constructed. Different ML classifiers (i.e. NB, SVM and ME) were considered for testing and up to 88.9 per cent accuracy found. Research conducted by Pang and Lee (2004) used the ML models NB and SVM to test a data set of 1,000 positive and 1,000 negative movie reviews. The accuracy of these models was found to be 86.4-87.2 per cent.

In research by Gamon (2004), an SVM model was trained and tested on another customer review data set. The model accuracy was found to be 69.5-77.5 per cent. Research by Kennedy and Inkpen (2006) used an SVM classifier tested on a dataset of 1,000 positive and 1,000 negative movie reviews constructed from the IMDB website. In this case, the accuracy was 80-85.9 per cent. Another model, created by Boiy *et al.* (2007), tested SVM, Multinomial NB, and ME on a dataset of 1,000 positive and 1,000 negative reviews, and validated the model using a dataset of 550 positive and 222 negative car reviews. This model's accuracy was 90.25 per cent.

Abbasi *et al.* (2008) found that in multi-language (English and Arabic) sentiment analysis, the positive sentiment text in both languages was shorter than the negative sentiment text, while using stylistic features increased the performances of the analysis. In a study by Ye *et al.* (2009), NB, SVM and a character-based N-gram model were tested on a data set of 591 negative and 600 positive travel blogs from travel.yahoo.com. In this case, the accuracy was 80.71-85.14 per cent. Paltoglou and Thelwall (2010) tested an SVM model on a data set of 1,000 positive and 1,000 negative movie reviews, and a multi-domain sentiment data set (MSMD) of 8,000 reviews. This research yielded an accuracy of 96.90 per cent for the movie reviews and 96.40 per cent for the MSMD.

Xia *et al.* (2011) created a model using an NB, ME, SVM meta-classifier combination and collected a dataset of 1,000 positive and 1,000 negative reviews of products (via Amazon) and movies (via IMDB), finding an accuracy of 88.65 per cent. Kang *et al.* (2012) proposed a new senti-lexicon-based approach to analyse the sentiment of restaurant reviews. In this research, an improved NB algorithm was proposed with unigram and bigram features, and was found to be effective. Kontopoulos *et al.* (2013) proposed the deployment of an original ontology-based technique for sentiment analysis of Twitter posts. The results were suitable for analysing post opinions of a specific topic. Kang and Park (2014) suggested a framework to evaluate customer satisfaction with mobile services using customer reviews by combining VIKOR and sentiment analysis techniques. The results were promising, as the method saved time in accurately determining customers' satisfaction.

A study by Singh *et al.* (2013) used a lexicon-based approach that works with SentiWordNet, to identify features related to sentiments using noun, verb and adverb. Fersini *et al.* (2014) developed a sentiment classifier by proposing an ensemble-based Bayesian network classifier to improve the training of the model. Mesnil *et al.* (2014) developed ensemble-based discriminative techniques for sentiment analysis and released this software as open access (https://github.com/mesnilgr/iclr15). Khadjeh Nassirtoussi *et al.* (2015) proposed a Heuristic-Hypernyms Feature-Selection approach by creating a means to identify words with the same parent-word, to be regarded as one entity. The work improved performance accuracy by up to 83.33 per cent. Tripathy *et al.* (2016) used four different ML

algorithms and unigram, bigram and trigram models to classify the sentiment of movie reviews. The performance of the classifiers was evaluated via various metrics, such as precision, recall, F-measure and accuracy.

A study by Nagamma *et al.* (2015) identified the relationship between the success of a movie at the box-office and the user's online movie reviews. This research incorporated a clustering approach with the term frequency – inverse document frequency technique, and showed improved performance accuracy. Aspect-based sentiment analysis is popular in the opinion-mining domain, and is primarily based on heuristic patterns to extract aspect sentiments (Htay and Lynn, 2013; Khan *et al.*, 2014; Maharani *et al.*, 2015; Parkhe and Biswas, 2016; Rana and Cheah, 2016), supervised and unsupervised classification of aspect-level sentiments (Manek *et al.*, 2017) and aspect-based summary generation (Samha *et al.*, 2014).

In addition, Stanford University (https://nlp.stanford.edu/sentiment/) has undertaken various researches with data sets, using unsupervised learning to cluster the words that are semantically similar to create word vectors, and many models were run, using these words, to understand the polarity of the reviews.

Khadjeh Nassirtoussi *et al.* (2015) proposed a Heuristic-Hypernyms Feature-Selection approach by creating a means to identify words with the same parent-word, to be regarded as one entity. The work improved performance accuracy by up to 83.33 per cent. Tripathy *et al.* (2016) used four different ML algorithms and unigram, bigram and trigram models to classify the sentiment of movie reviews. The performance of the classifiers was evaluated via various metrics, such as precision, recall, F-measure and accuracy.

3. Testing corpora

This study uses two different movie review corpora (comprising Indian movie reviews and movie review polarities). Such corpora have a prominent place in the literature on sentiment analysis, where a review is classified as either positive or negative. This is a useful task, because movie reviews are harder to classify compared to other product reviews (Turney, 2002; Dave et al., 2003). The polarity of the review can be extracted through the star rating information. First, the PPC approach and other ML classifiers are tested on the Indian movie reviews corpus; thereafter, the movie review polarity corpus is considered for validation purposes.

The main corpus of this study comprises Indian movie reviews, and was constructed using data from the IMDB[1] website (Tripathi and Trivedi, 2016). Indian viewers may be emotionally attached to movies or to the star(s) of the movies. Sometimes, even if the movie is not good, it still has a huge fan following because of the starring actor/actress, and therefore receives constructive reviews containing both positive and negative words. Hence, the Indian movie review data may be more challenging to analyse. Random Indian movies released between the years 2000 and 2015 were chosen from which to draw 1,000 negative and 1,000 positive reviews. The reviewers' names and the movie titles were excluded from the corpus. The positive and negative polarity of the reviews were determined by the star rating given to that review on the site. Out of a possible 10 stars, seven- to nine-star rating reviews were considered positive, and two- to four-star reviews were considered negative. The highest rating, i.e. 10 stars, and the lowest rating, i.e. one star, were not incorporated, owing to the possibility of biased or fake reviews. A maximum of 15 reviews per user, per sentiment category, was allowed to avoid bias arising from the incorporation of a large number of reviews written by certain individuals (Pang *et al.*, 2002).

The other corpus used in this study comprises movie review polarities (Pang and Lee, 2004), and contained 1,000 positive and 1,000 negative reviews, all composed before 2002 where 20 reviews per user was considered.

3.1 Pre-processing of the corpus

Pre-processing was conducted on the collected movie reviews whereby strings of characters were transformed to make them suitable for the classification task. The words/features were extracted from the movie reviews via the feature extraction method; thus, the words from the movie reviews were isolated via a string-to-word vector process to create a word dictionary. This incorporates the removal of HTML (or other) tags and stop-words (i.e. words that occur often in the reviews, such as articles, prepositions, conjunctions, etc.; e.g. The, A, An, In, That, etc.) and lemmatization (decreasing words to their actual frame; e.g. enhanced or enhancing can be written as "enhance") (Manek *et al.*, 2017).

A serious issue that arises during the sentiment analysis process is the high dimensionality of word/feature space, where one dimension of a word can be found in other reviews as well. This large feature space poses problems in standard classification methods, as it generates high computation costs and produces unreliable classification results. This problem may be tackled via a dimensionality reduction process carried out using feature selection methods.

In feature selection, the most informative word/features; (e.g. awesome, excellent, best, nice, etc., for positive polarity and worst, bad, underperforming, overacting, etc., for negative polarity) are extracted and less informative features are removed (Tripathi and Trivedi, 2016; Trivedi and Dey, 2016a, 2016b, 2016c). In this research, the greedy stepwise search method was used for the selection of informative feature subsets.

3.1.1 Greedy stepwise feature subset search. In the greedy stepwise feature subset search method (Trivedi and Dey, 2013c; Trivedi and Dey, 2014), an iterative process is applied to evaluate all features and a single informative feature is identified to include in the model. The stepwise regression method is used for feature evaluation. Informative features are extracted with three different techniques; viz., forward selection (to add most informative features to the model), backward selection (to remove the less informative features) and mixed selection (forward and backward selection together). To terminate the process, methods such as *p*-value are used. The termination process ensures that all the informative features have been added to the model or none of the features are left, which can add value.

Consider f^s as the feature set found after the dimension reduction process and f^e as the number of features participating in the evaluation process. The evaluation is conducted with the fitness value of the individual feature and computed with the fitness function; hence, the best feature set may be produced using the following fitness function equation:

$$f_*^b = \arg\max_{f^e \notin f^s} fit(f^s \cup \{f^e\}) \tag{1}$$

After collecting the informative features subset, the *feature representation method* is performed, in which words/features of the movie review are represented by the binary representation method. Therein, movie reviews and words together form a binary matrix called the Term–Document Matrix (TDM). This is as a term weighting method, and the binary matrix carries binary values (1 and 0), assigned a value of 1 if the particular feature/word is present in a specific movie review, and 0 otherwise.

Let us consider that each movie review is represented as a column vector, R^x, which is defined by the words extracted from the movie review; i.e. $R^x = (w^1, w^2, w^3 \ldots)$, where w^i is

termed as the i^{th} word/feature of the movie review, r^x (Kjersti and Eikvil, 1999). The combination of all movie review documents and words form an $M \times N$ matrix, where $_M$ represents the number of distinct features and N represents the number of movie review instances. Table I represents the term–document relationship as an a^{ij} matrix, defined as the degree of relationship between term $_i$ and instance $_j$.

3.2 Proposed probabilistic committee selection mechanism

3.2.1 Probabilistic classifiers. The Bayesian classifier method (Trivedi and Dey, 2013b) was proposed by Lewis (1998). He defined this term as the likelihood of a document being perceived by a vector of words falling into a specific class. This likelihood is calculated using the Bayes hypothesis:

$$P\left(\frac{c_i}{r_j}\right) = \frac{P(c_i) * P\left(\frac{r_j}{c_i}\right)}{P(r_j)} \tag{2}$$

where, $P(r_j)$ symbolizes the probability of arbitrarily selected reviews being represented by the review document vector r_j, and $P(c_i)$ is the probability of arbitrarily selected review document r_j falling into a particular class c_i. This classification method is usually known as Bayesian Classification.

The Bayesian method is popular, but is seen as challenging in the case of a high-dimensional data vector, r_j. This challenge can be tackled using the assumption that any two arbitrarily selected coordinates of review document vector r_j (tokens) are kept independent of one another. This assumption is well described by the equation:

$$P\left(\frac{r_j}{c_i}\right) = \prod_{l=1}^{n} P\left(\frac{w_j^l}{c_i}\right) \tag{3}$$

This presumption is taken by the classifier referred to as Naïve Bayes (NB).

3.2.2 Probabilistic committee selection classifier. In the PCC method, a classifier committee is developed wherein a number of classifiers are selected as members, and another classifier is selected as the president of the committee. Such a multilevel classification committee is known to comprise "stacking" (Wolpert, 1992; Sakkis *et al.*, 2001; Trivedi and Dey, 2016c). Each fresh movie review is first classified by the members of the committee. Further, the committee president receives the output of the committee members and selects the best one. The final classification decision is made by considering the individual members' decisions together with the president's decision. The benefit of this

	Word#1	Word#2	Word#3	.....
Movie Review#1	W^{11} (1 = present, 0 = absent)	W^{21} (1 = present, 0 = absent)	W^{31} (1 = present, 0 = absent)	.....
Movie Review#2	W^{12} (1 = present, 0 = absent)	W^{22} (1 = present, 0 = absent)	W^{32} (1 = present, 0 = absent)	.....
Movie Review#3	W^{13} (1 = present, 0 = absent)	W^{23} (1 = present, 0 = absent)	W^{33} (1 = present, 0 = absent)	.....
.....	.....	.....	.....	.....

Table I. Term–document binary representation

method is that, though members often do make mistakes (i.e. misclassification), the final decision of the committee is rarely incorrect (i.e. misclassifications are far rarer).

In this research, a novel PCC method (Figure 1) has been developed, wherein probabilistic classifiers – i.e. a Bayesian classifier and NB – are considered as members of the committee and NB is taken as the president of the committee.

3.3 Other machine learning classifiers

In this study, the proposed PCC model is compared with other popular ML classifiers, viz. Bayesian, NB, SVM, Decision Tree (J48) and RF. The rationale for comparing the proposed classifier with the other classifiers is derived from their popularity in the literature, as they have been tested in various text mining application domains. The ideas behind the selected state-of-the-art algorithms are quite different from one another, but each has been shown to be effective in the literature on text mining and natural language processing, as mentioned in Section 2, which also describes the algorithms.

3.3.1 Support vector machine. In the ML research, SVM has been found to be a promising and popular in classification research (Joachims, 1998, Moraes *et al.*, 2013; Trivedi and Dey, 2013a). SVM identifies a hyper-plane to separate two classes (such as positive and negative reviews) by maximizing the margin between them. This margin is computed via support vectors, where one is developed on each side of the hyper-plane. The main issue with SVM is that it is time-consuming because of the large number of training instances, which make it impractical for considering large-scale review corpora. SVM has generally been used in classification and sentiment analysis research.

The basic assumption behind SVM pertains to separating the classes (i.e. positive and negative) using the maximum margin produced by the hyper-plane. Consider a training sample, $X = \{x_i, y_i\}$, where $x_i \in R_n$ and $y_i \in \{+1, -1\}$, termed as the specific class of the i^{th} training sample. In this study, $_{+1}$ indicates positive reviews and $_{-1}$ negative reviews. The final output of the classifier is denoted via the equation:

$$y = wx - b \tag{4}$$

Where y is the final output of the classifier, w is termed as the normal vector analogous to those in the feature vector, x and b is the bias parameter that is obtained via the training method. The following optimization function is considered to maximize the separation between classes:

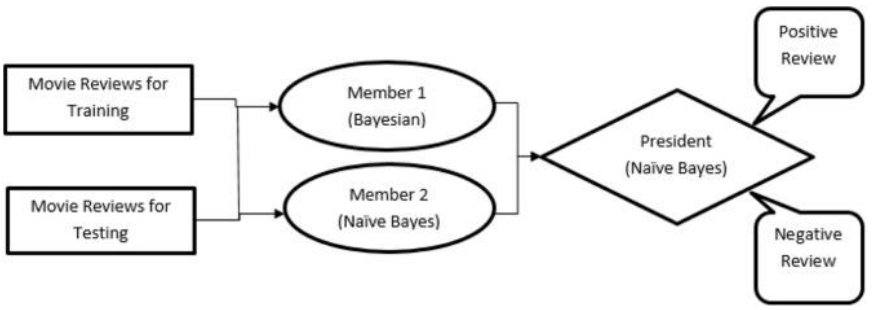

Figure 1.
Probabilistic committee selection classifier

$$\text{Minimize } \frac{1}{2}\|w\|^2 \tag{5}$$

$$\text{subject to } y_i(w \,.\, x - b) \geq 1, \ \forall i \tag{6}$$

In a few circumstances, the SVM classifier is unable to recognize a linear hyper-plane to separate the input instances into particular classes. This issue is tackled by changing the high-dimensional input instances by incorporating nonlinear transformation functions. This procedure isolates the information in such a way that a linear separable plane can be discovered in the transformed space. On the other hand, the high dimensions of the feature space make computation of the inner product of two transformed vectors practically unfeasible. To tackle this problem, "kernel functions" are involved and are used in place of the inner product of two transformed data vectors in the feature space. For viable operations, the computational effort is reduced via the appropriate use of kernel functions.

Appropriate selection of a kernel function is important for unique applications of SVM-based classification. A good choice of kernel function accords learning potential to SVM. A variety of kernel functions have been discussed in the literature. Our research incorporates a normalized polynomial kernel for the most accurate evaluation of SVM (Trivedi and Dey, 2016).

3.3.2 Decision tree (J48). Decision trees (Trivedi and Dey, 2016; Carreras and Màrquez, 2001) have likewise been broadly tested in the exploration of grouping. A decision tree builds some perception about examples to draw a classification choice. In preparation, it sets aside one perception as an opportunity to part the information. It picks the request to look at perceptions.

A basic decision tree classifier depends on C4.5 algorithms, which pick the most informative features from the feature subset. The features are preferred by normalizing the data (i.e. entropy distinction). This procedure is applied after assuming some base cases:

(1) If the entire example of the rundown has a place with a comparable classification, it creates a leaf hub in the choice tree for selecting that class.

(2) If none of the components can offer data pick up, it assembles a choice hub that is higher up on the tree via normal estimation of the class.

(3) If the illustration originates from a prior untouched class, it again makes a choice hub that is higher up the tree via normal estimation of the class.

Algorithm for C4.5

(1) Verify the above base cases.

(2) For each feature x^j, observe normalized information gain.

(3) For the best feature x_b^i (with higher normalized gain), produce a decision node that splits on x_b^i.

(4) Repeat the above steps on the sublists generated from splitting on x_b^i.

In this paper, three different kinds of decision tree classifiers are considered.

3.3.3 Random Forest. This method (Trivedi and Dey, 2013d) takes the concept of an ensemble of classifiers, where it combines the decision of several weak classifiers to produce the appropriate classification results. Similarly, RF combines decision trees to generate the highest classification accuracy. A bagging concept is used for an ensemble of weak decision

tree classifiers, and works by altering the data samples and creating a number of weak decision tree classifiers that are trained with the small subset of the data sample and selected features.

Algorithm for RF

Given: n^T = training examples, x^i = all selected features, x^e = selected features for ensembles, m^i = number of all member classifiers for ensemble.

Produce RF for m^i weak decision tree classifiers

- For m^i iterations:
- Do,
- Bagging: take sample $_{nT}$ with replacement instances from entire training set.
- Random feature selection: Grow decision tree without pruning. At each step, select informative features by incorporating x^e arbitrary selected features and computing the Gini index.

Classification:

- Use text set for m^i decision trees initiating from the root node. Assign a specific category with respect to the leaf node. Combining the individual decisions of weak classifiers by majority voting to produce the most accurate and strong classifier.

3.3 Experiment and evaluation setup

The proposed classification models and related classifiers were constructed and tested in a JAVA-based software environment and the simulation graphs were obtained from the use of MATLAB 8. All experiments were completed using a computer with Windows 7, 4 GB RAM and Intel CORE i7 processor.

The credibility of the proposed classifier was tested and compared with the other popular ML classifiers via the use of different metrics. In this study, four measures were used: F-value, false positive rate, receiver operating characteristic (ROC) curve and training time.

The simplest measure by which to test the performance of classifiers is the classification accuracy, which is explained as the percentage of correctly classified instances. The shortcoming of this metric is its failure to distinguish between false positive and false negatives. To tackle this issue, the F-value is used; that is, the harmonic sum of precision (i.e. the fraction of retrieved reviews that are relevant) and recall (the fraction of relevant reviews that are retrieved) and considered as a good measure to evaluate classification performance.

Misclassification is a serious problem in a classifier, especially when a positive review is misclassified as negative, which creates confusion in the minds of users about the product or service (in this study, movie reviews). To analyse the misclassification rate, the false positive (FP) rate is considered. The FP rate reveals the rate of reviews misclassified as positive. For the classifier to qualify as robust and accurate, this value should be as low as possible.

A robust classifier is not decided only by detecting how accurate it is, but also how fast. For cost-sensitive evaluation, training and testing time are also used in this study to check the speed of the classifier during training and testing. These metrics were incorporated to check the rapid installation and training capability of the classification model.

4. Results and analysis

After pre-processing, TDM metrics were constructed to train and test the proposed and comparison classifiers. The entire movie review corpus was split randomly into two parts, where 66 per cent of instances were used for training and the remaining 34 per cent for testing the classification models. The results and analyses are discussed in five parts, considering different metrics: F-value, FP rate, ROC, training time and all the metrics taken together.

4.1 Analysis of F-Value

Tables II and III and Figure 2 demonstrate the test results for the F-value on the Indian Movie Review corpus and the Movie Review Polarity corpus. The PCC was one of the most promising of the classifiers used for the testing, with the highest F-value (88.5). However, the F-value for the Bayesian (88.4) and SVM (88.1) classifiers were found to be comparable with the PCC, with a highly similar F-value. The F-values for other three classifiers were also acceptable, with the NB (87.9) classifier in the middle and RF (85.7) and J48 (82.5) having the poorest performance.

When the same classifiers were tested on the Movie Review Polarity corpus (Table III and Figure 2), the F-values strongly supported the tests conducted on the Indian Movie Review corpus. Again, the PCC had the highest F-value (78.5 per cent), followed by SVM and NB (77.9 per cent). The F-values for the remaining classifiers, RF, Bayesian and J48, were satisfactory (77.1 per cent, 74.4 per cent and 69.9 per cent respectively). J48 was again the poorest performer for this data set.

4.2 Analysis of false positive rate

Tables II and III and Figure 3 demonstrate the results for the FP rate (in per cent) for all classifiers tested on the Indian Movie Review and Movie Review Polarity corpora. The best FP rate amongst all classifiers was found for the PPC (with an 11.5 per cent FP rate). The Bayesian classifier depicted similar results (11.6 per cent), followed by SVM (11.9 per cent) and NB (12.1 per cent). RF (14.3 per cent) and J48 (17.5 per cent) showed the highest FP rate.

For the Movie Review Polarity corpus, the results for the ML classifiers support those for the Indian Movie Review corpus. The FP rates for NB and the proposed PCC are comparable (20.1 per cent for NB and 21.4 per cent for PCC). The results for SVM and RF are also good, at 22.1 per cent and 22.9 per cent, respectively.

4.3 Analysis of receiver operating characteristic

Tables II and III and Figure 4 show the results for the ROC area for all classifiers tested using the Indian Movie Review and Movie Review Polarity corpora. The PCC scored the highest in this regard (95 per cent for Indian Movie Review and 86.4 per cent for Movie Review Polarity) when compared with the other classifiers used in the study. However, the Bayesian classifier and NB classifier also yielded good ROC area coverage (94.9 per cent and 83.5 per cent for Bayesian and 94.7 per cent and 88.0 per cent for NB, for the Indian Movie Review corpus and Movie Review Polarity corpus, respectively). This was followed by RF (92.9 per cent, 83.6 per cent), J48 (88.3 per cent, 74.4 per cent) and finally SVM (88.1 per cent, 77.9 per cent).

4.4 Analysis of training time

Tables II and III and Figure 5 show the results for training time taken by each respective classifier. The results clearly depict that among all the classifiers, NB and Bayesian performed best (with 0.09 s and 0.22 s for NB and 0.19 s and 0.19 s for Bayesian, for the Indian Movie Review corpus and the Movie Review Polarity corpus, respectively). SVM and

RF took up a lot of time for training (12.81 s, 12.64 s for SVM; 12.83 s, 14.63 s for RF). Training time of PCC was slightly high (2.63 s and 2.58 s for both data sets, respectively) from Bayesian, Naïve Bayes and J48 but less than SVM and RF.

4.5 Analysis of all metrics

After considering all the evaluation metrics and ML classifier results, our proposed PCC can be said to have performed satisfactorily. In terms of performance accuracy, ROC curve and

Table II.
Results for proposed PCC and other ML classifiers (Indian Movie Review)

Machine learning classifiers	F-value (%)	FP rate (%)	ROC area (%)	Training time (s)
Indian Movie Review				
Bayesian	88.4	11.6	94.9	0.19
Naïve Bayes	87.9	12.1	94.7	0.09
J48	82.5	17.5	88.3	1.59
Random Forest	85.7	14.3	92.9	12.81
SVM	88.1	11.9	88.1	12.83
Proposed committee selection	88.5	11.5	95.0	2.63

Table III.
Results for proposed PCC and other ML classifiers (Movie Review Polarity)

Machine learning classifiers	F-value (%)	FP rate (%)	ROC area (%)	Training time (s)
Movie Review Polarity				
Bayesian	74.4	25.6	83.5	0.19
Naïve Bayes	77.9	20.1	88.0	0.22
J48	69.9	30.1	74.2	1.83
Random Forest	77.1	22.9	83.6	12.64
SVM	77.9	22.1	77.9	14.63
Proposed committee selection	78.5	21.4	86.4	2.58

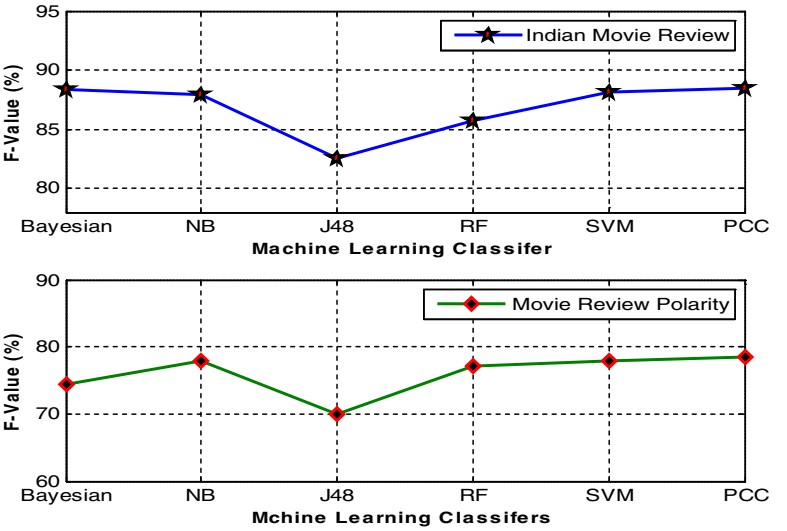

Figure 2.
F-values of proposed PCC and other ML classifiers

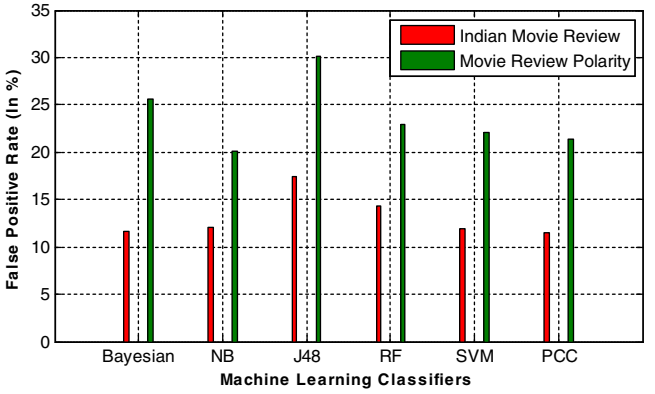

Figure 3.
FP rate of proposed PCC and other ML classifiers

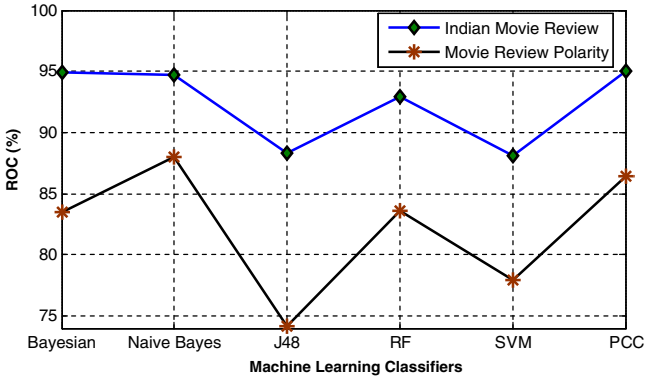

Figure 4.
ROC of proposed PCC and other ML classifiers

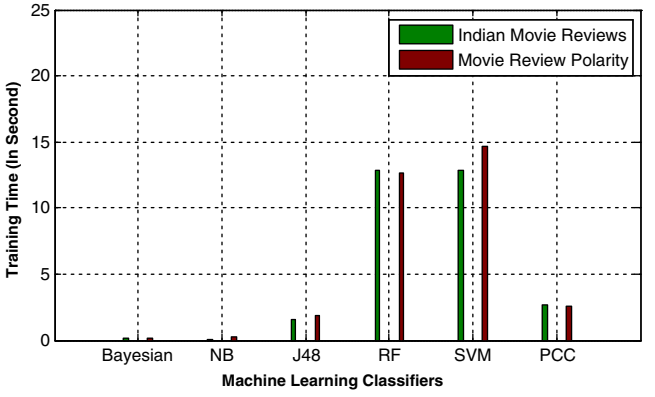

Figure 5.
Training time of proposed PCC and other ML classifiers

misclassification rate, the PCC was shown to be the best of the classifiers tested in this research. However, this classifier had an average training time. The training time of the proposed classifier is slightly high because it uses a committee of multiple classifiers to combine individual decisions. However, this metric can be ignored when measuring the strength of the proposed classifier because once the classifier has been trained, training time is not a significant issue.

5. Discussion

After analysing the results of this study, the PCC was found to perform well when compared to the other state-of-the-art classifiers tested and compared in this study. Sometimes ML classifiers make mistakes during training time and misjudge new instance categories. A committee selection mechanism is used to yield a better-trained classifier, where multiple classifier members participate in decision-making and the president classifier makes a final classification decision. Such multiple learning provides a better understanding of the training samples to the president and hence the likelihood of misclassification is reduced. The proposed PCC uses such a committee selection concept, and was found to perform well in terms of accuracy, ROC, reducing the misclassification rate, but with moderate speed of training.

All the classifiers considered in this study were tested using features selected from Indian Movie Review and Movie Review Polarity corpora, where the former was considered the primary corpus in this study. Results from this corpus show that the proposed PCC is comparable to Bayesian methods in classification tasks, where the performance of PCC is slightly higher than that of Bayesian. To validate these results, the same classifiers were again tested using features selected from the Movie Review Polarity corpus; the results again strongly supported the proposed classifier, with accurate movie review predictions. On the other hand, SVM and RF were found to perform well in terms of F-value, FP rate and ROC, though they underperformed in terms of training time. The training times for SVM and RF were very similar. The decision tree classifier was found to be the worst in this study on all measurement dimensions.

The proposed PCC model does have room for improvement, because it lacks the capacity to extract implicit aspects (Lal and Asnani, 2014). In addition, because of a lack of consideration of informal opinion carriers, such as emoticons and slang (Gamon *et al.*, 2005), during pre-processing, classification accuracy may have been affected. The proposed model is also unable to consider multiple aspects and associated sentiments present in a single sentence. For example, in the sentence "The food was very good, but it took over half an hour to be seated, and the service was terrible", "Food" and "Restaurant's ambience and services" are two different aspects, and "Good" and "Terrible" are the two different opinions expressed for these two aspects, respectively.

6. Conclusion

In this research, a novel approach to identifying the sentiment polarity of movie reviews was proposed using committee selection of probabilistic classifiers to construct an optimal sentiment classification model. The greedy stepwise method was used to select the most informative features. The proposed method was tested and compared with other popular ML models, including Bayesian, NB, J48, SVM and RF, and achieved a maximum accuracy of 88.5 per cent and 78.5 per cent for the Indian Movie Review corpus and Movie Review Polarity corpus, respectively, using a 66-34 per cent data split. The proposed classifier was also found to perform well in reducing the misclassification rate, where low FP rates, of 11.5 per cent and 21.4 per cent, were identified for the respective corpora. The

value of the ROC curve for the PCC was found to be the highest amongst all other classifiers tested in this study. However, the training time for the proposed classifier was found to be slightly high.

Because this research did not consider implicit aspects, future studies may extend this work by extracting implicit aspects of users. The features information may also be improved by considering opinion carriers, such as emoticons and slang, during pre-processing. More efficient post-processing methods using different feature extraction approaches may also be incorporated to enhance accuracy and minimize the FP rate. In future, multiple aspects, and the associated sentiments of the word, may be captured in the analysis stage. The model could be validated using different corpora and *n*-fold cross-validation processes. Some other ML classifiers and feature-selection methods could also be used to compare with our proposed model. Further, the proposed model may be tested using reviews in different languages, because in the present scenario only the English language was studied.

This study proposes a better aspect-based probability committee selection sentiment analysis model for Indian movie reviews. Various enterprises may use such a model to analyse and summarize the sentiments regarding their products and services and thereby improve customer relationships, which can enhance their position in the competitive market. In addition, the proposed sentiment classifier may be used in diverse applications, such as blog mining, spam classification and other areas of text mining.

Note

1. www.imdb.com

References

Abbasi, A., Chen, H. and Salem, A. (2008), "Sentiment analysis in multiple languages: feature selection for opinion classification in web forums", *ACM Transactions on Information Systems (Systems)*, Vol. 26 No. 3, p. 12.

Aue, A. and Gamon, M. (2005), "Customizing sentiment classifiers to new domains: a case study", *Proceedings of Recent Advances in Natural Language Processing (RANLP), September*, Vol. 1 No. 3.1, pp. 1-2.

Boiy, E., Hens, P., Deschacht, K. and Moens, M.F. (2007), "Automatic sentiment analysis in on-line text", *ELPUB Digital Library*, pp. 349-360, available at: https://elpub.architexturez.net/doc/oai-elpub-id-138-elpub2007

Carreras, X. and Màrquez, L. (2001), "Boosting trees for clause splitting", *Proceedings CoNLL-2001 Shared Task, Toulouse*, Association for Computational Linguistics Stroudsburg, PA.

Dave, K., Lawrence, S. and Pennock, D.M. (2003), "Mining the peanut gallery: opinion extraction and semantic classification of product reviews", *Proceedings of the 12th international conference on World Wide Web, ACM*, pp. 519-528, available at: www.kushaldave.com/p451-dave.pdf

Fersini, E., Messina, E. and Pozzi, F.A. (2014), "Sentiment analysis: Bayesian ensemble learning", *Decision Support Systems*, Vol. 68 No. 16, pp. 26-38.

Gamon, M., Aue, A., Corston-Oliver, S. and Ringger, E. (2005), "Pulse: mining customer opinions from free text", *Lecture Notes in Computer Science*, Vol. 3646, pp. 121-132.

Gamon, M. (2004), "Sentiment classification on customer feedback data: noisy data, large feature vectors, and the role of linguistic analysis", *Proceedings of the 20th international conference on Computational Linguistics, Association for Computational Linguistics*, p. 841, available at: https://dl.acm.org/citation.cfm?id=1220476

Htay, S.S. and Lynn, K.T. (2013), "Extracting product features and opinion words using pattern knowledge in customer reviews", *The Scientific World Journal*, Vol. 2013.

Joachims, T. (1998), *Text Categorization with Support Vector Machines: Learning with Many Relevant Features*, Springer Berlin Heidelberg, pp. 137-142, available at: https://link.springer.com/chapter/10.1007/BFb0026683

Kang, D. and Park, Y. (2014), "Review-based measurement of customer satisfaction in mobile service: Sentiment analysis and VIKOR approach", *Expert Systems with Applications*, Vol. 41 No. 4, pp. 1041-1050.

Kang, H., Yoo, S.J. and Han, D. (2012), "Senti-lexicon and improved Naïve Bayes algorithms for sentiment analysis of restaurant reviews", *Expert Systems with Applications*, Vol. 39 No. 5, pp. 6000-6010.

Kennedy, A. and Inkpen, D. (2006), "Sentiment classification of movie reviews using contextual valence shifters", *Computational Intelligence*, Vol. 22 No. 2, pp. 110-125.

Khadjeh Nassirtoussi, A., Aghabozorgi, S., Ying Wah, T. and Ngo, D.C.L. (2015), "Text mining of news-headlines for FOREX market prediction", *Expert Systems with Applications: An International Journal*, Vol. 42 No. 1, pp. 306-324.

Khan, K., Baharudin, B. and Khan, A. (2014), "Identifying product features from customer reviews using hybrid patterns", *Int. Arab J. Inf. Technol*, Vol. 11 No. 3, pp. 281-286.

Kjersti, A. and Eikvil, L. (1999), *Text Categorisation: A Survey*, Norwegian Computing Center, NR, p. 941, available at: www.nr.no/%7Eeikvil/tm_survey.pdf

Kontopoulos, E., Berberidis, C., Dergiades, T. and Bassiliades, N. (2013), "Ontology-based sentiment analysis of twitter posts", *Expert Systems with Applications*, Vol. 40 No. 10, pp. 4065-4074.

Lal, M. and Asnani, K. (2014), "Implicit aspect identification techniques for mining opinions: a survey", *International Journal of Computer Applications*, Vol. 98 No. 4.

Lewis, D.D. (1998), "Naive (Bayes) at forty: the independence assumption in information retrieval", In *Machine Learning: ECML-98*, Springer Berlin Heidelberg, pp. 4-15, available at: https://link.springer.com/chapter/10.1007/BFb0026666

Liu, S.M. and Chen, J.H. (2015), "A multi-label classification based approach for sentiment classification", *Expert Systems with Applications*, Vol. 42 No. 3, pp. 1083-1093.

Maharani, W., Widyantoro, D.H. and Khodra, M.L. (2015), "Aspect extraction in customer reviews using syntactic pattern", *Procedia Computer Science*, Vol. 59, pp. 244-253.

Manek, A.S., Shenoy, P.D., Mohan, M.C. and Venugopal, K.R. (2017), "Aspect term extraction for sentiment analysis in large movie reviews using gini index feature selection method and SVM classifier", *World Wide Web*, Vol. 20 No. 2, pp. 135-154.

Mesnil, G., Mikolov, T., Ranzato, M.A. and Bengio, Y. (2014), "Ensemble of generative and discriminative techniques for sentiment analysis of movie reviews", arXiv preprint arXiv:1412.5335.

Moraes, R., Valiati, J.F. and Neto, W.P.G. (2013), "Document-level sentiment classification: an empirical comparison between SVM and ANN", *Expert Systems with Applications*, Vol. 40 No. 2, pp. 621-633.

Mostafa, M.M. (2013), "More than words: social networks' text mining for consumer brand sentiments", *Expert Systems with Applications*, Vol. 40 No. 10, pp. 4241-4251.

Nagamma, P.H.R., Pruthvi, K.K.N. and Shwetha, N.H. (2015), "An improved sentiment analysis of online movie reviews based on clustering for box-office prediction", *Computing, Communication and Automation (ICCCA), 2015 International Conference on, IEEE*, pp. 933-937, available at: http://ieeexplore.ieee.org/document/7148530/

Paltoglou, G. and Thelwall, M. (2010), "A study of information retrieval weighting schemes for sentiment analysis", *Proceedings of the 48th Annual Meeting of the Association for*

Computational Linguistics, Association for Computational Linguistics, pp. 1386-1395, available at: https://dl.acm.org/citation.cfm?id=1858822

Pang, B. and Lee, L. (2004), "A sentimental education: sentiment analysis using subjectivity summarization based on minimum cuts", *Proceedings of the 42nd annual meeting on Association for Computational Linguistics, Association for Computational Linguistics*, p. 271, available at: www.aclweb.org/anthology/P04-1035

Pang, B. and Lee, L. (2008), "Opinion mining and sentiment analysis", *Foundations and Trends® in Information Retrieval*, Vol. 2 Nos 1/2, pp. 1-135.

Pang, B., Lee, L. and Vaithyanathan, S. (2002), "Thumbs up? Sentiment classification using machine learning techniques", *Proceedings of the ACL-02 conference on Empirical methods in natural language processing-Volume 10, Association for Computational Linguistics*, pp. 79-86, available at: www.aclweb.org/anthology/W02-1011

Parkhe, V. and Biswas, B. (2016), "Sentiment analysis of movie reviews: finding most important movie aspects using driving factors", *Soft Computing*, Vol. 20 No. 9, pp. 3373-3379.

Rana, T.A. and Cheah, Y.N. (2016), "Exploiting sequential patterns to detect objective aspects from online reviews", *Advanced Informatics: Concepts, Theory And Application (ICAICTA), 2016 International Conference On, IEEE*, pp. 1-5, available at: https://ieeexplore.ieee.org/document/7803101/

Sakkis, G., Androutsopoulos, I., Paliouras, G., Karkaletsis, V., Spy-ropoulos, C.D. and Stamatopoulos, P. (2001), "Stacking classifiers for anti-spam filtering of e-mail", *Proceedings of Empirical Methods in Natural Language Processing, Cornell University Library*, Vol. 1, pp. 44-50, available at: https://arxiv.org/abs/cs/0106040

Samha, A.K. Li, Y. and Zhang, J. (2014), "Aspect-based opinion extraction from customer reviews", arXiv preprint arXiv:1404.1982.

Singh, V.K., Piryani, R., Uddin, A. and Waila, P. (2013), "Sentiment analysis of movie reviews: a new feature-based heuristic for aspect-level sentiment classification", *In Automation, Computing, Communication, Control and Compressed Sensing (iMac4s), 2013 International Multi-Conference on, IEEE*, pp. 712-717, available at: https://ieeexplore.ieee.org/document/6526500/

Tripathi, A. and Trivedi, S.K. (2016), "Sentiment analyis of Indian movie review with various feature selection techniques", In *Advances in Computer Applications (ICACA), IEEE International Conference on, IEEE*, pp. 181-185, available at: https://ieeexplore.ieee.org/document/7887947/

Tripathy, A., Agrawal, A. and Rath, S.K. (2016), "Classification of sentiment reviews using n-gram machine learning approach", *Expert Systems with Applications*, Vol. 57, pp. 117-126.

Trivedi, S.K. (2016), "A study of machine learning classifiers for spam detection", *Computational and Business Intelligence (ISCBI), 2016 4th International Symposium on, IEEE*, pp. 176-180, available at: https://ieeexplore.ieee.org/document/7743279/

Trivedi, S.K. and Dey, S. (2013a), "Effect of various kernels and feature selection methods on SVM performance for detecting email spams", *International Journal of Computer Applications*, Vol. 66 No. 21.

Trivedi, S.K. and Dey, S. (2013b), "Interplay between probabilistic classifiers and boosting algorithms for detecting complex unsolicited emails", *Journal of Advances in Computer Networks*, Vol. 1 No. 2.

Trivedi, S.K. and Dey, S. (2013c), "Effect of feature selection methods on machine learning classifiers for detecting email spams", *Proceedings of the 2013 Research in Adaptive and Convergent Systems, ACM*, pp. 35-40, available at: https://dl.acm.org/citation.cfm?id=2513313

Trivedi, S.K. and Dey, S. (2013d), "An enhanced genetic programming approach for detecting unsolicited emails", *Proc. 2013 IEEE 16th International Conference on Computational Science and Engineering, Australia Published by IEEE Computer Society, Sydney*, 978-0-7695-5096-1/13 $31.00 © 2013 IEEE, doi10.1109/CSE.2013.171.

Trivedi, S.K. and Dey, S. (2014), "Interaction between feature subset selection techniques and machine learning classifiers for detecting unsolicited emails", *ACM SIGAPP Applied Computing Review*, Vol. 14 No. 1, pp. 53-61.

Trivedi, S.K. and Dey, S. (2016a), "A Comparative Study of Various Supervised Feature Selection Methods for Spam Classification", *Proceedings of the Second International Conference on Information and Communication Technology for Competitive Strategies, ACM*, p. 64, available at: https://dl.acm.org/citation.cfm?id=2905122

Trivedi, S.K. and Dey, S. (2016b), "A combining classifiers approach for detecting email spams", *Advanced Information Networking and Applications Workshops (WAINA), 2016 30th International Conference on, IEEE*, pp. 355-360, available at: https://ieeexplore.ieee.org/document/7471226/

Trivedi, S.K. and Dey, S. (2016c), "A novel committee selection mechanism for combining classifiers to detect unsolicited emails", *VINE Journal of Information and Knowledge Management Systems*, Vol. 46 No. 4, pp. 524-548.

Turney, P.D. (2002), "Thumbs up or thumbs down? Semantic orientation applied to unsupervised classification of reviews", *Proceedings of the 40th annual meeting on association for computational linguistics, Association for Computational Linguistics*, pp. 417-424, available at: https://dl.acm.org/citation.cfm?id=1073153

Wan, Y. and Gao, Q. (2015), "An ensemble sentiment classification system of twitter data for airline services analysis", *2015 IEEE International Conference on Data Mining Workshop (ICDMW), IEEE*, pp. 1318-1325, available at: https://ieeexplore.ieee.org/document/7395820/

Wolpert, D. (1992), "Stacked generalization", *Neural Networks*, Vol. 5 No. 2, pp. 241-260.

Xia, R., Zong, C. and Li, S. (2011), "Ensemble of feature sets and classification algorithms for sentiment classification", *Information Sciences*, Vol. 181 No. 6, pp. 1138-1152.

Ye, Q., Zhang, Z. and Law, R. (2009), "Sentiment classification of online reviews to travel destinations by supervised machine learning approaches", *Expert Systems with Applications*, Vol. 36 No. 3, pp. 6527-6535.

Further reading

Zhai, Z., Xu, H., Kang, B. and Jia, P. (2011), "Exploiting effective features for Chinese sentiment classification", *Expert Systems with Applications*, Vol. 38 No. 8, pp. 9139-9146.

Corresponding author

Shrawan Kumar Trivedi can be contacted at: f10shrawank@iimidr.ac.in