The Electronic Library

Capturing user sentiments for online Indian movie reviews: A comparative analysis of different machine-learning models

Journal:	*The Electronic Library*
Manuscript ID	EL-04-2017-0075.R1
Manuscript Type:	Article
Keywords:	User sentiment analysis, Opinion mining, Machine learning classifiers, Indian movie reviews

Capturing Users' Sentiments for Online Indian Movie Reviews: A Comparative Analysis of Different Machine-learning Models

Abstract

Purpose: Sentiment analysis and opinion mining are emerging areas of research for analyzing Web data and capturing users' sentiments. This research presents sentiment analysis of an Indian movie-review corpus using natural language processing and various machine-learning classifiers.

Design/methodology/approach: In this paper, a comparative study between three machine-learning classifiers (Bayesian, naïve Bayesian and support vector machine (SVM) was performed. All the classifiers were trained on the words/features of the corpus extracted, using five different feature-selection algorithms (chi-square, info-gain, gain-ratio, one-R, and relief-F attributes), and a comparative study was performed between them. The classifiers and feature-selection approaches were evaluated using different metrics (F-value, false-positive rate (FP rate), and training time).

Findings: The results of this study show that, for the maximum number of features, the relief-F feature-selection approach was found to be the best, with better F-values, a low FP rate, and less time needed to train the classifiers while, for the least number of features, one-R was better than relief-F. When the evaluation was performed for machine-learning classifiers, SVM was found to be superior, although the Bayesian classifier was comparable with SVM.

Keywords: Sentiments Analysis, Opinion Mining, Machine Learning Classifiers, Indian Movie Review.

Introduction

The Web has significantly transformed the world, and the rise of Web 2.0 has totally changed the situation as people can now express their thoughts and opinions digitally. People can also read specific product or service reviews, written by other users, by simply accessing the desired online portal before making a purchase decision; alternatively, if someone wants to watch a movie they can simply read the movie's reviews before making a decision. The Internet has given freedom of speech to users: they can write their feelings/sentiments in the form of reviews or blogs using online portals. Such user behavior creates opportunities for online retailers and organizations in the form of text data. Further, this text data can be analyzed using various natural language processing tools and artificial intelligence, which can help businesses make better decisions and better predict success and sustainability.

Sentiment analysis is an important area of research, utilizing a number of applications, and is found to be robust when seeking to understand customers' feelings and attitudes toward various products and services (Manek et al., 2017). The feedback provided by customers helps organizations to make informed decisions. For example, a hotel review may help a visitor to locate the most suitable hotel. In the same fashion, movie reviews may help consumers in deciding whether a movie is worth watching.

A sentiment is an expression of opinion, feeling, or emotion, or an assessment made by the individual that can be either positive, negative, or neutral. These polarities are known as sentiment orientations, opinion orientations, semantic polarity, or simply orientations. Such polarity can be classified and predicted by opinion mining, and can be distinguished in three ways.

1. Document-level sentiments. At this level, a whole document is considered as a positive or negative sentiment for specific products or services. This level is restricted to those documents that do not measure or compare various attributes because, at this level, a whole document represents a sentiment toward a single attribute (or single product) (Liu, 2012).
2. Sentence-level sentiments. At this level of classification, a sentence is used to decide the positive, negative, or neutral sentiment toward products or services. Sentence-level sentiments deal with subjective classification, and differentiate between subjective and objective sentiment classifications, whereby subjective sentences reveal the opinion or sentiment, and objective sentences convey the true information. Information-handling requirements for objective sentences are found to be greater than those for subjective sentences, e.g. "A few buttons on the remote control of a smart TV which we purchased a couple of days back are malfunctioning" (Liu, 2012).
3. Entity- and aspect-level sentiments. This analysis is based on the feature, or attributes, of the text where a feature, or word, is taken as either a positive or a negative sentiment. This is a finer-grained analysis, in which all the features, taken together, provide insight in to the overall sentiment weight of any opinion. Aspect-level sentiment analysis defines the opinion as positive, negative, or neutral, based on the words'/features' sentiment weight (Hu & Liu, 2004).

In various opinions and reviews, the sentiment of an opinion depends on the individual entities and their respective aspects (Manek et al., 2017). For example, the sentence "Although the battery backup is not that high, I still like the Samsung mobile phone" contains both partial positive sentiments and partial negative sentiments. Here, a positive sentiment is expressed for "Samsung" and a negative sentiment for "battery backup. "Hence, the aim of performing an analysis at this level is to decide which entities have which aspect. To perform such an analysis, unstructured text must be converted to structured text to capture these entities and aspects. This level of analysis imposes more challenges than document- or sentence-level analysis. Sentiment classification is a domain-specific problem (Aue & Gamon, 2005). In natural language processing, this is a special case of text classification, where text mining, natural language processing, and computational linguistics methods are all employed.

Two main approaches and techniques to analyze sentiments are: machine learning; and the semantic approach. Most research undertaken has focused on the English language, as it is internationally accepted by most of countries. In the machine-learning approach, data are converted in to feature vectors, then machine-learning classifiers are trained to infer a combination of specific features yielding a specific class (Pang&Lee, 2008) and, finally, a model is created that is used to predict the sentiment polarity of a fresh review or opinion.

This study aims to develop a movie-review filter with the help of machine-learning models. Entity- and aspect-based sentiment analysis has been preferred for this research for finer-grained analysis. Indian movies released between 2000 and 2015 were considered for this analysis, for which the sentiments of Indian users' movie reviews have been captured.

The literature identifies many challenges in analyzing sentiments; for example, it is more difficult than traditional text classification, because text classification relies on keywords and sentiment relies on opinion, aspect, and/or their entities (Pang et al., 2002). Another challenge for sentiment analysis is the authenticity of the end users, which has been addressed by some researches by incorporating more text data (Manek et al., 2017). In this study, the users' authenticity problem has been addressed by incorporating the positive reviews (seven- to nine-star rating) and negative reviews (two- to four-star rating) from Indian movie review data, with the assumption that too high a positive score, or too low a negative score, may reveal users' authenticity issue. The model presented in this paper could, in the future, be validated through the use of a larger dataset. Another problem is the differentiation between positive and negative sentiments. Sometimes, negative sentiments are expressed without the use of any negative words, and are expressed as sarcasm, or irony, which creates further challenges (Riloff et al., 2013).

This paper is divided into five sections. In first section, the introduction and background of the study are explained. The second section details related work, with the help of an extensive literature review. In section three, the methodology is examined. Section four discusses the experiment's results and analysis. Finally, in section five, conclusions are drawn.

Related Work

Sentiment analysis, also called opinion mining (Pang and Lee, 2008), is a type of text-mining approach where people's attitudes and expressions toward various products/services, organizations, individuals, or an event, are captured and analyzed using various tools and techniques. Much research has been undertaken, and much is ongoing, in this domain.

Pang et al.(2002) created a sentiment-analysis model using naïve Bayesian (NB), maximum entropy (ME), and support vector machine (SVM) techniques. All the models have been tested on our dataset (700 positive and 700 negative movie reviews, collected from the IMDB website (www.imdb.com)), with an accuracy of 77–82.9%. Dave et al. (2003) conducted a study using NB, SVM and ME models, which were tested on product reviews collected from Amazon (www.amazon.com), with an 88.9% performance accuracy. Mullen and Collier's (2004) research integrated point-wise mutual information (PMI) values, Osgood semantic factors (Osgood et al., 1964), and some syntactic relations in the features of SVM. Pang & Lee (2004) conducted a study with NB and SVM models that were tested on a dataset of 1,000 positive and 1,000 negative movie reviews collected from the IMDB website, with an accuracy of 86.4–87.2%.

A study by Gamon (2004) used the SVM model, and was tested on a customer-review dataset with an accuracy of 69.5–77.5%. Pang and Lee (2005) conducted a study with SVM, support vector regression (SVR), regression, and metric labeling models, that were tested on a dataset of 5,006 movie reviews collected from the IMDB website, with accuracy of 54.6–66.3%. Another study (Kennedy and Inkpen, 2006) took the SVM classifier and tested it on a dataset of 1,000 positive and 1,000negative movie reviews collected from the IMDB website, withan accuracy of 80–85.9%. Chen et al.'s (2006)study used decision trees (C4.5), SVM, and NB models, tested on the dataset of 3,168 book reviews collected from Amazon, with an accuracy of 84.59%.

A study by Boiy et al. (2007) experimented with models, such as SVM, multinomial NB, and ME, that were tested on a dataset of 1,000 positive and 1,000 negative reviews collected from the IMDB website, and were also tested on car reviews (550 positive and 222 negative),with an accuracy of 90.25%. Annett & Kondrak's(2008)study with SVM, NB, and decision trees, performed on a dataset of 1,000 positive and 1,000 negative movie reviews from the IMDB website, had an accuracy greater than 75%. Ye et al.(2009) used NB, SVM, and character-based N-gram model, and tested on a dataset of 591 negative and 600 positive travel blogs collected from the Yahoo travel page (www.yahoo.com/lifestyle/tagged/travel), with accuracy range of 80.71–85.14%. Further research by Paltoglou and Thelwall (2010) used the SVM model to test a dataset of 1,000 positive and 1,000 negative movie reviews, and a multi-domain sentiment dataset(MDSD) of 8,000 reviews, and achieved an accuracy of 96.90% for movie reviews and 96.40% for the MSMD. Thet and Khoo (2010) performed an experiment with a corpus of 520 movie reviews, and compared the textual characteristics of consumers' reviews across four different genres, and found that users express more verbs and adverbs than noun and prepositions. This study also analyzed the positive and negative sentiments of different genres. Another study, by Xia et al. (2011), tested models such as NB, ME, SVM, and a meta-classifier combination, on a dataset of 1,000positive and 1,000 negative reviews of products (Amazon) and movies (IMDB),and achieved an accuracy of 88.65%. Zhai et. al. (2011)tested an improved NB model on a dataset of 5,700 positive and 757 negative restaurant reviews, and achieved an accuracy of 83.6%.Ghorbel and Jacot (2011) experimented with a corpus of movie reviews written in the French language, and used a supervised classification, combined with SentiWordNet, to determine sentiment polarity.

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). 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 TF-IDF 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 & Lynn, 2013; Khan et al., 2014; Maharani et al., 2015;Parkhe & Biswas, 2016; Rana & 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 datasets, 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.

Methodology

In this study, a corpus of movie reviews was identified. Movie reviews have a prominent place in the sentiment-analysis and opinion-mining domains, where a review is classified either as positive or negative. The rationale behind choosing movie reviews as our potential data was its availability via many online movie sites, with a star rating also provided for each specific review. In addition, movie reviews are harder to classify than other products reviews (Turney, 2002; Dave et al., 2003), while the correct polarity of the review can be extracted directly from the rating information, i.e. the number of stars.

This study specifically analyzed Indian movies users, and gathered the data from Indian movie reviews on the IMDB website (Tripathi & Trivedi, 2016).After analyzing the content of the Indian movie reviews, it was observed that many Indian viewers may be emotionally connected with the movies, or the star(s) of the movies. Sometimes, even if movie is not good, but the actor/actress in this movie has a huge fan following, it gets constructive reviews, mixed both with positive and negative sentiments. Hence, the Indian movie-review data may be more challenging to analyze. Random Indian movies were selected, released between 2000 and 2015, to develop 1,000 negative and 1,000 positive reviews. Only reviews that had star ratings given by users were considered for preparing the corpus. The positive and negative polarity of the review was decided based on the star rating given to that review on the website. Reviewers' names and movie names have not been included in the corpus. Out of a possible ten stars, seven- to nine-star rating reviews were considered as positive reviews, and two- to four-star rating reviews were considered as negative reviews. The highest rating, i.e. ten stars, and the lowest rating, i.e. one star, have not been incorporated, owing to the possibility of biased or fake reviews. A maximum of 15 reviews per user, per sentiment category, was allowed to tackle the issue of large number of reviews written by certain individuals (Pang et al., 2002).

Pre-processing of the Corpus

Pre-processing was performed on the movie-review corpus to transform the reviews, whereby strings of characters were transformed to a binary representation to make them suitable for machine-learning classifiers. A feature-extraction process was used to extract words/features from the movie review files.

In this process, tokens/words of movie reviews are extracted by a tokenization method (splitting the text document in to a series of tokens) to develop an associated feature dictionary. The main hurdle in the tokenization process is the amount of noise present in online movie reviews, such as URLs, HTML tags, scripts, advertisements, and symbols, which are useless for the machine-learning process, and need to be removed from the feature dictionary (Manek et al., 2017). This noise was removed with methods such as case normalization (making each feature/word either uppercase or lowercase), stop-word removal (frequent words such as articles, prepositions and conjunctions, etc.), and lemmatization (reducing words to their basic form, such as "reviewing" to "review"). Extracted feature dictionary is further taken for feature selection process.

Feature-selection Techniques

The major problem with sentiment analysis is the high dimensionality of the feature space, where one dimension of a unique word is seen in many reviews. This size of the feature space creates difficulties for standard classification methods because of the high computation cost, and unreliable classification output. This large feature space is reduced in size by a dimensionality reduction method, with an accompanying feature-selection process.

Feature selection is used to obtain an informative feature subset to reduce the feature space. This research incorporates five different feature-selection algorithms for extracting features (Tripathi & Trivedi, 2016; Trivedi & Day, 2016a, 2016b, 2016c). The different feature-selection algorithms are described in the following subsections.

Information gain (IG)

Information gain is used to extract informative features from the entire feature dictionary. The information of features are evaluated by calculating information gain (Trivedi, & Dey, 2013a, 2016). Information gain is calculated by measuring changes in the overall entropy by including a new feature for classification. Basically, entropy is an expected value of information required to classify an instance. This method works on association of features. Let us consider X and Y as the desecrate random variables/features. Entropy of Y before and after inclusion of X is calculated as:

$$H(Y) = -\sum_{y \in Y} p(y) \log_2 p(y) \tag{1}$$

$$H\left(\tfrac{Y}{X}\right) = -\sum_{x \in X}(x) \sum_{y \in Y} p\left(\tfrac{y}{x}\right) \log_2 p\left(\tfrac{y}{x}\right) \tag{2}$$

IG is the value of additional information of Yprovided by X, for which the entropy of Y decreases. It is computed by the following formulae:

$$IG = H(Y) - H\left(\tfrac{Y}{X}\right) \tag{3}$$

$$IG = H(X) - H\left(\tfrac{X}{Y}\right) \tag{4}$$

$$IG = H(Y) + H(X) - H(Y, X) \tag{5}$$

IG is a symmetrical measurement, hence the value of this for Y after observing X will be similar to the value for X after observing Y.

Gain ratio (GR)

This is an extension of IG. The weakness of IG is its bias toward the selection of features that have a higher numerical value with less information.

$$IG = H(Y) - H\left(\tfrac{Y}{X}\right) = H(X) - H\left(\tfrac{Y}{X}\right) \tag{6}$$

To compensate for the bias of IG, GR is used, which is a non-symmetrical measurement (Hall & Smith, 1998).

$$GR = \frac{IG}{H(X)} \tag{7}$$

From equation (5), when variable Y is to be predicted, IG will be normalized by dividing by the entropy of X. The normalization process gives GR values between 0 and 1. When the GR value is 1, the information in X will completely predict Y, and if it is 0, then X and Y will have no relation with each other. GR differs from IG in that it can accept features even if they have a lower numerical value.

Chi-square (χ^2)

This is a well-known and commonly used technique to select informative features (Liu & Setiono, 1995). The χ^2 method provides valuable features from the feature space with respect to the class by analyzing the value of χ^2 statistics. This method tests the initial hypothesis H_0, which assumes that “two features are dissimilar”.

$$\chi^2 = \sum_{i=1}^{r} \sum_{j=1}^{c} \left(\frac{O^{ij} - E^{ij}}{E^{ij}}\right)^2 \tag{8}$$

Here, the notations O^{ij} is the observed frequency, and E^{ij} is the expected frequency, justified by the null hypothesis. Higher values of the χ^2 provide significant evidence against the initial hypothesis H_0.

Relief-F (RF)

This algorithm randomly selects features from movie reviews and observes their nearest neighbors to adjust a final feature weighting vector (Trivedi & Dey, 2013a,2016 a, b,c). In this way, it gives larger weight to the features that better discriminate between the instances from other neighbors of different classes. Specifically, it attempts to observe the best estimation of W^f from the given probabilities to assign the weight for each feature f.

$$W^f = P\left(\text{Different value for } f / \text{Nearest Instance from different class}\right) - P\left(\text{Different value for } f / \text{Nearest Instance from same class}\right) \tag{9}$$

One rule (OR)

This algorithm was proposed by Holte (1993), a professor from University of Ottawa. This method works by taking a set of instances with many features and different classes. It selects a single best feature iteratively and bases the rules solely on that feature. The algorithm is described as follows.

For each feature f^x :

1. For each value v^x from the domain f^x
2. Choose the set of examples where feature f^x has value v^x
3. Let us consider c^x = most frequent class within the set.
4. Add the condition "for feature f^x with value v^x the class will be c^x".

(Rule for feature f^x)

Representation

After capturing the informative feature subset, a representation process was employed, where words/features of movie review files are represented by binary representation. In this method, review files and words together form a binary matrix called a term-document matrix (TDM).This method is called a term-weighting method. This binary matrix takes binary values (1 and 0), where 1 indicates the presence of a particular feature/word in a specific movie review file, and 0 otherwise (Tripathi & Trivedi, 2016).

Let us assume that each review file is represented as a column vector D^x, which is defined as the words extracted from the review files i.e. $D^x = (w^1, w^2, w^3 ....)$ where w^i is termed as i^{th} word/feature of the movie review d^x. The combination of all review files and words form an $M \times N$ matrix where M represents to the number of distinct features and N represents the number of movie reviews. Table 1 represents the term–document relationship as a a^{ij} matrix that is defined as the degree of relationship between term i and instance j.

Table 1: Term-to-documents binary representation (W^{ij} =1 or 0)

	Word#1	Word#2	Word#3	....
Movie Review#1	W^{11}	W^{21}	W^{31}	....
Movie Review#2	W^{12}	W^{22}	W^{32}	....
Movie Review#3	W^{13}	W^{23}	W^{33}	....
....	....	....	....	....

Opinion classification

This research incorporates machine-learning classifiers to classify movie reviews in to positive and negative sentiments. Machine learning is a method in which a specific algorithm learns or "trains" from previous data/experience to test the present scenario and to predict the future. In this research, three state-of-the-art machine-learning algorithms, i.e. Bayesian, NB, and SVM classifiers, are taken to classify the opinions of Indian movie reviews. These three algorithms are quite different in their working procedure but, in the literature, each of one has shown a significant contribution to the text-categorization domain. The aim of this study is to test the movie-review corpus with these techniques and, after comparison, a robust opinion classifier of movie review can be developed.

Probabilistic classifiers

The Bayesian classifier was initially proposed by Lewis (1998). He has suggested the term $P\left(\frac{c_i}{d_j}\right)$ as the probability of a document recognized by a vector $d_j = w_j^1, w_j^2, \ldots, w_j^n$ of terms falling in a specific class c_i. This probability is evaluated by the Bayes theorem:

$$P\left(\frac{c_i}{d_j}\right) = \frac{P(c_i) * P\left(\frac{d_j}{c_i}\right)}{P(d_j)} \tag{10}$$

where $P(d_j)$ is the probability of randomly selected documents represented by the documents vectors d_j and $P(c_i)$, termed as probability of randomly selected documents d_j, falling in a particular class c_i. This classification method is generally called a "Bayesian classification". The Bayesian method is widely used, but can have drawbacks for a high dimensional data vector d_j. The challenge is addressed by assuming that any two randomly selected coordinates of document vector d_j (tokens) are independent to each other. This assumption is represented by the given equation:

$$P\left(\frac{d_j}{c_i}\right) = \prod_{l=1}^{n} P\left(\frac{w_j^l}{c_i}\right) \tag{11}$$

This assumption creates a different probabilistic classifier known as "Naïve Bayes", a well-known classifier in the text-mining domain (Trivedi & Dey, 2013a, 2013b, 2014, 2016).

SVM

SVM is effective in the text-mining research (Joachims, 1998) that works by developing a hyper-plane to separate two classes (such as positive and negative reviews), while maximizing the margin between them. This margin is calculated by support vectors that are constructed, one on each side of the hyper-plane. The main hurdle from SVM is the large amount of time needed, which is closely related to the number of training instances, and found impractical for large-scale applications such as sentiment analysis. SVM has widely been used in the research of classification.

SVM works by separating the classes (i.e. positive and negative) by using a maximum margin created by the hyper-plane. Let us consider a training sample $X = \{x_i, y_i\}$, where $x_i \in R_n$ and $y_i \in \{+1, -1\}$ are defined as the particular class for i^{th} training sample. In this research, $+1$ is denoted as a positive movie review, and −1 is considered as a negative movie review. Final classification output is computed by the following equation:

$$y = w.x - b \tag{12}$$

where $y$is the final classification output, w is the normal vector that is analogous to those in the feature vector x, and $b$is the bias parameter, which is determined by the training procedure. The following optimization function is taken to maximize the separation between classes:

$$\text{minimize} \quad \tfrac{1}{2}\|w\|^2 \tag{13}$$

$$\text{subject to} \quad y_i\left(w.x-b\right)\geq 1, \forall i \tag{14}$$

Sometimes, SVM classifiers had difficulty in identifying a linear hyper-plane to separate the input data into specific classes. This problem is resolved by transforming the high dimensional input data with the help of some non-linear transformation functions. This process helps to separate the input data in such a manner that a linear separable plane is revealed in the transformed space. In addition, high dimensionality of the feature space makes the computation of the inner product of two transformed vectors practically unfeasible. To resolve this problem "kernel functions" are used in place of the inner product of two transformed data vectors in the feature space. For viable operations, the computational effort is minimized by the appropriate use of kernel functions.

An appropriate selection of kernel function is essential 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 the normalized polynomial function with SVM, as it is found to be effective in the literature (Trivedi & Day., 2013).

Evaluation metrics

This study incorporates an Indian movie-review corpus of 1,000 positive and 1,000 negative reviews, captured from the IMDB website. Three different machine-learning classifiers (Bayesian, NB, and SVM) were incorporated in this study for capturing the sentiments of reviews. Five different feature selection techniques (χ^2, IG, GR, OR, and RF) have been used in this study for capturing informative features. The whole movie review corpus was split in to 66% for training the classifiers and 34% for testing. Java and Microsoft Excel 10 platforms were used to complete this study, with a Windows8 operating system and 8GB RAM. The considered corpus was checked with XL-Minor software for oversampling problems, and it was found that the training corpus (TDM) was free from oversampling problems, as we achieved an almost 50% success and 50% non-success rate. The performance of classifiers was measured by three indicators: F-measure, false-positive rate (FP rate), and time taken for training.

F-measure (Provost & Fawcett, 2001) is defined as the harmonic mean of precision and recall. This tells the actual percentage of sentiments/words that are correctly classified.

FP rate (Viola & Jones, 2001) gives the rate of misclassified instances. For good and accurate results, the classifier's false-positive value should be as low as possible. In the opinion mining, the false-positive value means the percentage sentiments/words expressed by the peoples are misclassified.

Training time (Lim et al., 2000) for the classifier was also captured to complete the analysis. To make a time-sensitive and accurate model, the following metric was used (Table 2).

Table 2: Evaluation metrics

Instruments	Related Formulas
F- Measures	$F = \frac{2*\text{Precision}*\text{Recall}}{\text{Precision}+\text{Recall}}$
False Positive Rate	$FP_{rate} = \frac{N_{positive \to m}}{N_{positive \to m} + N_{positive \to c}}$
Training Time	Measured in second during training process.

Experiment results and analysis

After experimental analysis, a comparative study was performed, first on feature selection, and then on machine-learning classifiers. The evaluation and compression of the algorithms was performed using F-value, FP rate and training time.

Evaluation with F-value

For evaluating the performance, different feature-selection mechanisms were employed for extracting informative features from the corpus. Further, the analysis was performed on the different feature subsets (minimum to maximum number) with the help of three different machine-learning classifiers. The following subsections analyze the performance of each classifier in terms of the classification accuracy that is captured by the F-value.

Bayesian classifier

When analysis is performed with respect to the Bayesian classifier, the following observations are seen from the results.

Observation 1: From Table 3 and Figure 1, it is observed that, for the maximum number of features, the Bayesian classifier performed differently for all the feature-selection mechanisms, with different F-values. In this case, RF algorithms performed better than other feature-selection algorithms tested in this study, and gave the best results, with an 88.8% F-value when compared to other feature selection techniques, whereas OR (F-value = 81.3%) was found to be the second-best algorithm. On the other hand, a χ^2 feature-selection algorithm was not up to the mark, and gave only a 64.2% F-value.

Observation 2: From Table 3 and Figure 1, for a smaller number of features, the RF algorithm again performed better, with an F-value of 61.6%, while all other feature-selection techniques under-performed and behaved in a similar way, with an F-value of 0.332.

Table 3: F-value for the Bayesian classifier with different feature-selection algorithms

Number of Features	200	400	600	800	1000	1200
Chi-Square	0.332	0.332	0.332	0.332	0.332	0.642
Gain Ratio	0.332	0.332	0.332	0.518	0.659	0.775
Info-Gain	0.332	0.332	0.332	0.332	0.332	0.618
One-R	0.332	0.332	0.332	0.347	0.522	0.813
Relief F	0.616	0.616	0.616	0.635	0.888	0.888

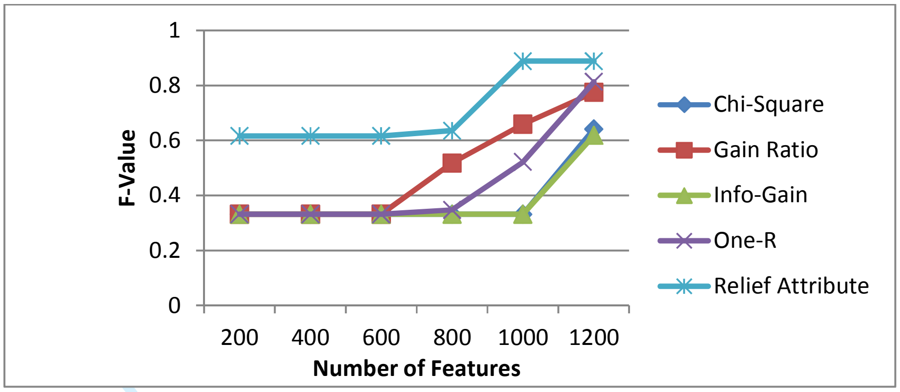


Figure 1:F-value for the Bayesian classifier with different feature-selection algorithms

NB classifier

Observation 1: Table 4 and Figure 2 show the results of the NB classifier with different feature-selection mechanisms. For the maximum number of features, it is observed that an RF algorithm gave the best results, with an F-value of 86.7%, whereas OR was the second best, with an F-value of 86%.

Observation 2: For the minimum number of features, χ^2 performed better, with an F-value of 55.5%, whereas RF and IG algorithms were the second best, with F-values of 55.1%.

Table 4: F-value for the NB classifier with different feature-selection algorithms

Number of Features	**200**	**400**	**600**	**800**	**1000**	**1200**
Chi-Square	0.555	0.551	0.556	0.594	0.693	0.784
Gain Ratio	0.486	0.418	0.449	0.384	0.28	0.19
Info-Gain	0.551	0.559	0.557	0.589	0.689	0.787
One-R	0.369	0.395	0.495	0.582	0.683	0.86
Relief Attribute	0.551	0.578	0.561	0.697	0.869	0.867

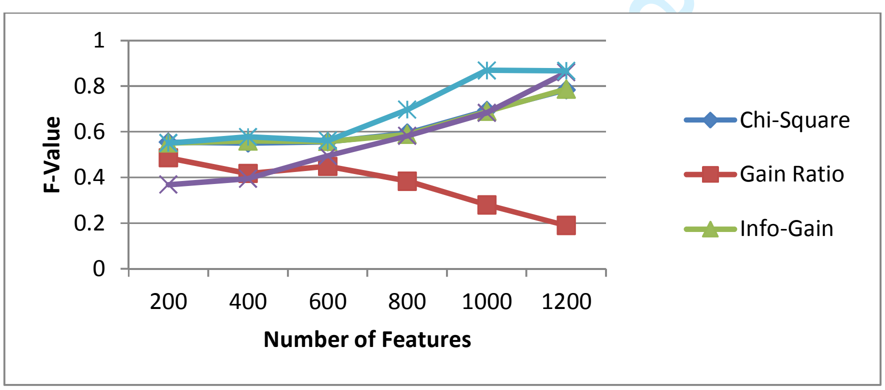


Figure 2:F-value for the NB classifier with different feature-selection algorithms

SVM classifier

Observation 1: From Table 5 and Figure 3, we observed that, for the maximum number of features for SVM classifiers, RF again performed best, with an F-value of 0.894. RF has, therefore, performed excellently for Bayesian, NB, and SVM classifiers.

Observation 2: From Table 5 and Figure 3 we observed that, for the minimum number of features for SVM classifiers, RF was found good, with an F-value of 0.593, compared to others. In addition, OR was the underperforming with F-Value (0.392) for less number of features.

Table 5: F-value for the SVM classifier with different feature-selection algorithms

Number of Features	200	400	600	800	1000	1200
Chi-Square	0.575	0.579	0.563	0.643	0.744	0.81
Gain Ratio	0.54	0.57	0.576	0.681	0.759	0.859
Info-Gain	0.562	0.569	0.575	0.657	0.753	0.815
One-R	0.392	0.456	0.596	0.706	0.766	0.86
Relief Attribute	0.593	0.596	0.598	0.737	0.897	0.894

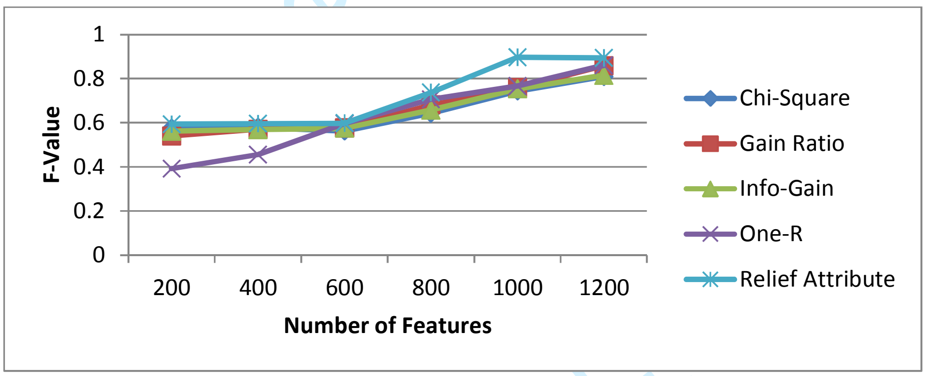


Figure 3:F-value for the SVM classifier with different feature-selection algorithms

Evaluation with FP rate

Bayesian classifier

Observation 1: From Table 6 and Figure 4, we observed that, for the maximum number of features, RF gave the lowest FP rate of 0.112, and performed excellently again when compared to other feature-selection techniques. This was followed by OR with an 0.187 FP rate.

Observation 2: From Table 6 and Figure 4 we observed that, for the minimum number of features, RF again performed outstandingly, with a 0.365 FP rate, while the others showed a similar behavior, with a 0.499 FP rate. The lower the FP rate, the more effective the results.

Table 6: FP rate for the Bayesian classifier with different features-selection algorithms

Number of Features	200	400	600	800	1000	1200
Chi-Square	0.499	0.499	0.499	0.499	0.499	0.354
Gain Ratio	0.499	0.499	0.499	0.478	0.341	0.225
Info-Gain	0.499	0.499	0.499	0.499	0.499	0.379
One-R	0.499	0.499	0.499	0.497	0.408	0.187
Relief Attribute	0.365	0.365	0.365	0.364	0.112	0.112

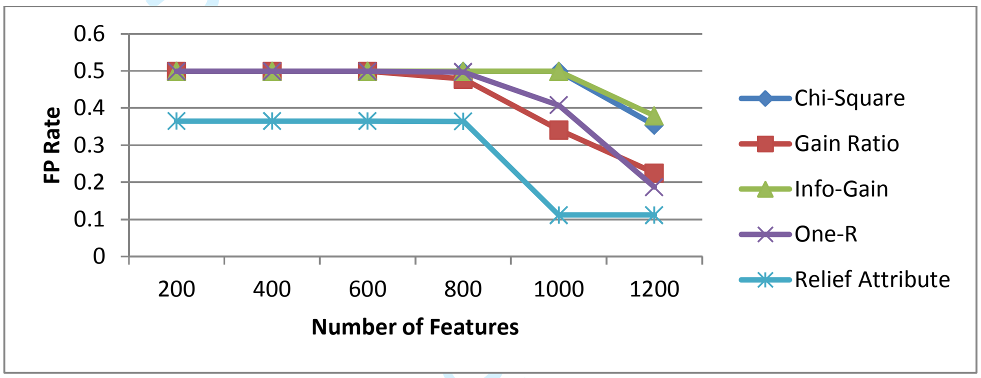


Figure 4: FP rate for the Bayesian classifier with different features-selection algorithms

NB classifier

Observation 1: From Table 7 and Figure 5, we observed that, for the maximum number of features or words, for NB classifiers, the FP rate was lowest for RF, compared to all other feature-selection techniques. RF therefore outperformed all other feature-selection techniques for the FP rate.

Observation 2: From Table 7 and Figure 5, we observed that, for the minimum number of features, χ^2 had the lowest FP rate (0.444), followed by an FP rate of 0.447 for RF. χ^2, therefore, outperformed all other feature-selection techniques.

Table 7: FP rate for the NB classifier with different feature-selection algorithms

Number of Features	200	400	600	800	1000	1200
Chi-Square	0.444	0.446	0.442	0.402	0.301	0.215
Gain Ratio	0.486	0.418	0.449	0.384	0.28	0.19
Info-Gain	0.449	0.439	0.44	0.407	0.305	0.212
One-R	0.631	0.601	0.491	0.405	0.307	0.14
Relief Attribute	0.447	0.418	0.437	0.302	0.131	0.133

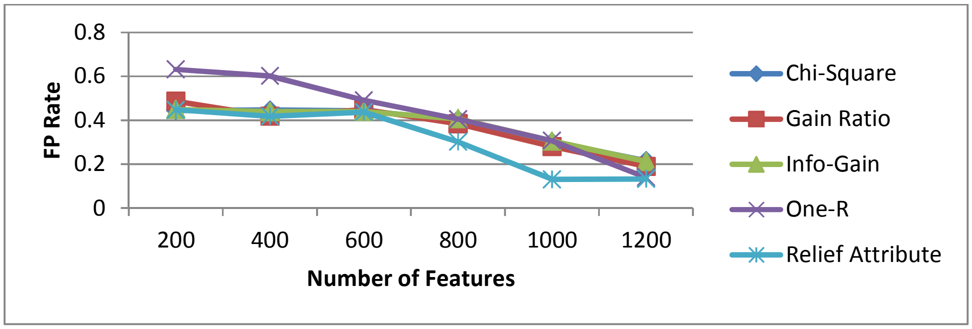


Figure 5: FP rate for the NB classifier with different feature-selection algorithms

SVM classifier

Observation1: From Table 8 and Figure 6 we observed that, for maximum number of features, RF had the lowest FP rate (0.106). RF performed excellently and outperformed all feature-selection algorithms.

Observation 2: From Table 8 and Figure 8, we observed that, for the minimum number of features, RF had the lowest FP rate (0.407) and outperformed all other feature-selection algorithms, showing extraordinary results.

Table 8: FP rate for the SVM classifier with different feature-selection algorithms

Number of Features	**200**	**400**	**600**	**800**	**1000**	**1200**
Chi-Square	0.425	0.421	0.437	0.357	0.256	0.19
Gain Ratio	0.459	0.43	0.424	0.319	0.241	0.141
Info-Gain	0.438	0.431	0.425	0.343	0.247	0.185
One-R	0.608	0.544	0.403	0.294	0.234	0.14
Relief Attribute	0.407	0.403	0.401	0.263	0.103	0.106

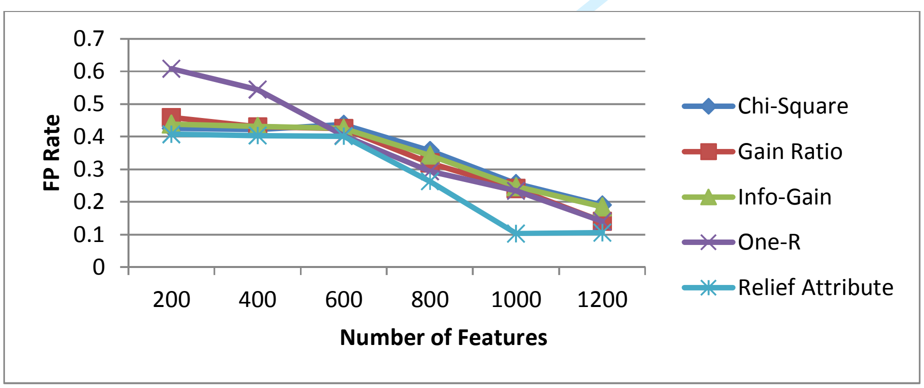


Figure 6: FP rate for the SVM classifier with different feature-selection algorithms

Evaluation with training time

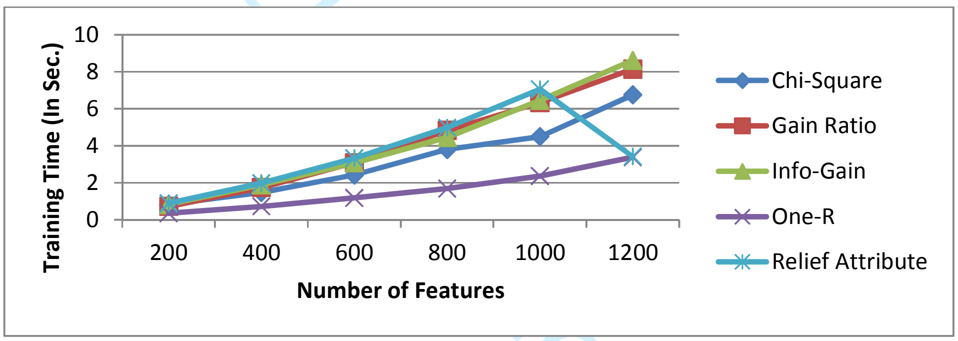
Bayesian classifier

Observation 1: From Table 9 and Figure 7, we observed that training time is one of the most important attributes to examine. For the maximum number of features, OR took the shortest time to train the data (3.38 seconds), followed by RF (3.42 seconds). Other feature-selection techniques, e.g.GR and IG, took longer time (8.14 seconds and 8.61 seconds respectively) to train the data and complete the operation.

Observation 2: From Table 9 and Figure 7, we observed that, for the minimum number of features, OR took the shortest time to train data (0.36 seconds), while RF took the longest time (0.89 seconds) to train the data, delaying the operation.

Table 9: Training time, in seconds, for the Bayesian classifier with different feature-selection algorithms

Number of Features	200	400	600	800	1000	1200
Chi-Square	0.83	1.48	2.42	3.8	4.49	6.75
Gain Ratio	0.72	1.75	3.09	4.83	6.33	8.14
Info-Gain	0.83	1.89	3.08	4.45	6.45	8.61
One-R	0.36	0.72	1.19	1.69	2.36	3.38
Relief Attribute	0.89	1.98	3.32	4.98	7.05	3.42



Figue7: Training time, in seconds, for the Bayesian classifier with different feature-selection algorithms

NB classifier

Observation 1: From Table 10 and Figure 8, we observed that, for the maximum number of features, RF outperformed all other feature-selection algorithms, performing excellently, based on the shorter time taken to train the data.

Observation 2: From Table 10 and Figure 8, we observed that, for the minimum number of features, RF took the longest time to complete the training of dataset, while OR took the least time, with an outstanding performance of 0.19 seconds to complete its operation.

Table 10: Training time, in seconds, for the NB classifier with different feature-selection algorithms

Number of Features	200	400	600	800	1000	1200
Chi-Square	0.44	0.84	1.39	2.14	2.66	3.57
Gain Ratio	0.41	1.02	1.8	2.75	3.33	4.22
Info-Gain	0.47	1.08	1.77	2.52	3.81	4.17
One-R	0.19	0.41	0.7	1.02	1.28	1.74
Relief Attribute	0.52	1.25	1.94	2.67	3.48	1.72

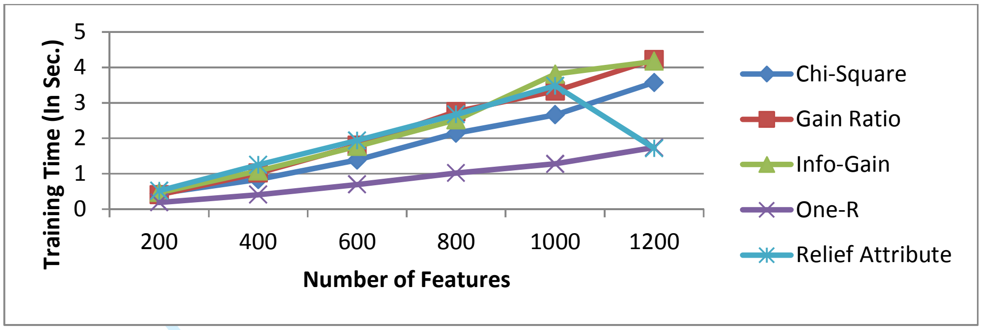


Figure 8: Training time, in seconds, for the NB classifier with different feature-selection algorithms

SVM classifier

Observation1: From Table 11 and Figure 9, we observed that, for the maximum number of features, RF took the shortest time (1.72 seconds), and also gave outstanding results for F-value and F-measures. RF trained its data in less time and produced outstanding results.

Observation 2: From Table 11 and Figure 9, we observed that, for the minimum number of features, OR took the least time, while RF performed worst, in this case, by taking the most time. OR performed outstandingly in this case.

Table 11: Training time, in seconds, for the SVM classifier with different feature-selection algorithms

Number of Features	200	400	600	800	1000	1200
Chi-Square	0.44	0.84	1.39	2.14	2.66	3.57
Gain Ratio	0.41	1.02	1.8	2.75	3.33	4.22
Info-Gain	0.47	1.08	1.77	2.52	3.81	4.17
One-R	0.19	0.41	0.7	1.02	1.28	1.74
Relief Attribute	0.52	1.25	1.94	2.67	3.48	1.72

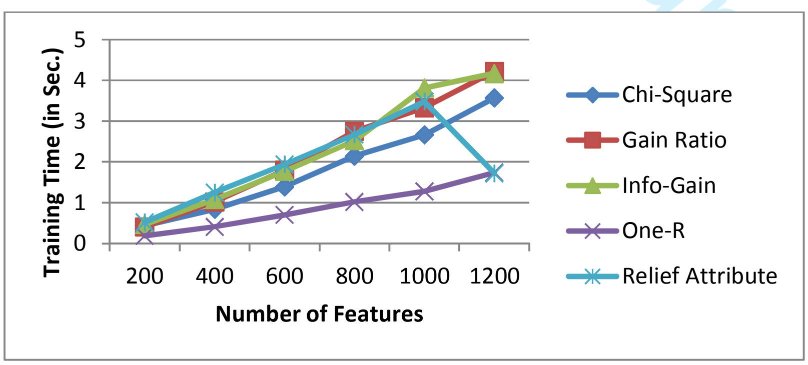


Figure 9: Training time, in seconds, for the SVM classifier with different feature-selection algorithms

Conclusion

This study identified an effective sentiment-classification model for aspect-based opinion mining, tested on Indian movie-review data. For identifying the proposed model, a comparative analysis between feature-selection methods, as well as between machine-learning classifiers, was performed. After analysis of the opinion of the different movie users, a robust and sensitive classification model has been developed. Among the feature-selection methods, RF was found to be the most effective while, among the machine-learning classifiers, SVM classifiers were found to be the most effective, good F-values, low FP rates and less training time required. In addition, the results for Bayesian classifiers were comparable with the best SVM classifiers. This research concludes that SVM, with an RF feature-selection technique, together construct a promising model for sentiment analysis of Indian movie reviews.

Limitations and Future Work

This study has a number of limitations. The proposed machine-learning model lacks the capacity to extract implicit aspects (Lal & Asnani, 2014). In addition, owing to a lack of consideration of informal opinion carriers such as emoticons and slang (Gamon et al., 2005) during pre-processing, classification accuracy may be affected. Also, the proposed model fails 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 “Restaurants ambience and services” are two different aspects, and “Good” and “Terrible” are the two different opinions expressed for these two aspects, respectively. Finally, the scope of this study was limited to the English language sentiments only.

Possible future research may extend this work by extracting the implicit aspect of the users. Opinion carriers such as emoticons and slang, may also be considered during pre-processing to enhance information on the features. Some more efficient post-processing methods may be incorporated to enhance the accuracy, and to minimize the FP rate. In future, multiple aspects, and their associated sentiments for a word, may be captured and incorporated in the analysis. The same research may be validated through different corpuses and *n*-fold cross-validation processes, where the proposed model may be verified. Some other machine-learning classifiers and feature-selection methods may also be used to compare with our proposed model. Further, the proposed model may also be tested on different language reviews as, in the present scenario, only the English language has been studied.

This study proposes a better aspect based sentiment analysis model for Indian Movie Reviews. Different enterprises can use such models to analyze and summarize the sentiments of their product and services to improve customer relationship, and can thus make their position stronger in the competitive market. In addition, the proposed sentiment classifier can also be used in diverse applications like blog mining, spam classification and other areas text mining.

References

Annett, M., & Kondrak, G. (2008). A Comparison of Sentiment analysis Techniques: Polarizing Movie Blogs. In Proceedings of the 21st Canadian Conference on Artificial Intelligence (CCAI 2008) , pages 25–35.

Aue, A., &Gamon, M. (2005, September). Customizing sentiment classifiers to new domains: A case study. In Proceedings of recent advances in natural language processing (RANLP) (Vol. 1, No. 1-3, pp. 2-1).

Boiy E, Hens P, Deschacht K, Moens M (2007) Automatic sentiment analysis in on-line text. In: Proceedings of the 11th international conference on electronic publishing, Vienna pp 349–360

Chen, C., Ibekwe-SanJuan, F., SanJuan, E., & Weaver, C. (2006, October). Visual analysis of conflicting opinions. In Visual Analytics Science And Technology, 2006 IEEE Symposium On (pp. 59-66). IEEE.

Dave, K., Lawrence, S., &Pennock, D. M. (2003, May). Mining the peanut gallery: Opinion extraction and semantic classification of product reviews. In Proceedings of the 12th international conference on World Wide Web (pp. 519-528). ACM.

Fersini, E., Messina, E., &Pozzi, F. A. (2014). Sentiment analysis: Bayesian ensemble learning. Decision support systems, 68, 26-38.

Gamon, M. (2004, August). Linguistic correlates of style: authorship classification with deep linguistic analysis features. In Proceedings of the 20th international conference on Computational Linguistics (p. 611). Association for Computational Linguistics.

Gamon, M., Aue, A., Corston-Oliver, S., &Ringger, E. (2005). Pulse: Mining customer opinions from free text. Lecture notes in computer science, 3646, 121-132.

Ghorbel, H., &Jacot, D. (2011). Sentiment analysis of French movie reviews. In Advances in Distributed Agent-Based Retrieval Tools (pp. 97-108). Springer Berlin Heidelberg.

Hall, M.A., & Smith, L.A. (1998). Practical feature subset selection for machine learning. In Proceedings of the 21st Australian Computer Science Conference, 181–191

Holte, R. C. (1993). Very simple classification rules perform well on most commonly used datasets. Machine learning, 11(1), 63-90.

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

Hu, M., & Liu, B. (2004, August). Mining and summarizing customer reviews. In Proceedings of the tenth ACM SIGKDD international conference on Knowledge discovery and data mining (pp. 168-177). ACM.

Kennedy, A., &Inkpen, D. (2006). Sentiment classification of movie reviews using contextual valence shifters. Computational intelligence, 22(2), 110-125.

Khan, K., Baharudin, B., & Khan, A. (2014). Identifying product features from customer reviews using hybrid patterns. Int. Arab J. Inf. Technol., 11(3), 281-286.

Maharani, W., Widyantoro, D. H., &Khodra, M. L. (2015). Aspect Extraction in Customer Reviews Using Syntactic Pattern. Procedia Computer Science, 59, 244-253.

Lal, M., &Asnani, K. (2014). Implicit Aspect Identification Techniques for Mining Opinions: A Survey. International Journal of Computer Applications, 98(4).

Lewis, D. D. (1998). Naive (Bayes) at forty: The independence assumption in information retrieval. In Machine learning: ECML-98 (pp. 4-15). Springer Berlin Heidelberg.

Lim, T. S., Loh, W. Y., & Shih, Y. S. (2000). A comparison of prediction accuracy, complexity, and training time of thirty-three old and new classification algorithms. Machine learning, 40(3), 203-228.

Liu, B. (2012). Sentiment analysis and opinion mining. Synthesis lectures on human language technologies, 5(1), 1-167.

Liu, H., &Setiono, R. (1995, November). Chi2: Feature selection and discretization of numeric attributes. In tai (p. 388). IEEE.

Manek, A. S., Shenoy, P. D., Mohan, M. C., &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, 20(2), 135-154.

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

Mullen, T., & Collier, N. (2004, July). Sentiment Analysis using Support Vector Machines with Diverse Information Sources. In EMNLP (Vol. 4, pp. 412-418).

Nagamma, P., H. R. Pruthvi, K. K. Nisha, and N. H. Shwetha. "An improved sentiment analysis of online movie reviews based on clustering for box-office prediction." In Computing, Communication & Automation (ICCCA), 2015 International Conference on, pp. 933-937. IEEE, 2015.

Osgood, C. E., Suci, G. J., & Tannenbaum, P. H. (1964). The Measurement of Meaning [by] Charles E. Osgood, George J. Suci [and] Percy H. Tannenbaum.

Paltoglou, G., &Thelwall, M. (2010, July). A study of information retrieval weighting schemes for sentiment analysis. In Proceedings of the 48th annual meeting of the association for computational linguistics (pp. 1386-1395). Association for Computational Linguistics.

Pang, B., Lee, L., &Vaithyanathan, S. (2002, July). Thumbs up?: sentiment classification using machine learning techniques. In Proceedings of the ACL-02 conference on Empirical methods in natural language processing-Volume 10 (pp. 79-86). Association for Computational Linguistics.

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.

Pang, B., & Lee, L. (2008). Opinion mining and sentiment analysis. Foundations and Trends® in Information Retrieval, 2(1–2), 1-135.

Parkhe, V., & Biswas, B. (2016). Sentiment analysis of movie reviews: finding most important movie aspects using driving factors. Soft Computing, 20(9), 3373-3379.

Provost, F., & Fawcett, T. (2001). Robust classification for imprecise environments. Machine learning, 42(3), 203-231.

Rana, T. A., &Cheah, Y. N. (2016, August). Exploiting sequential patterns to detect objective aspects from online reviews. In Advanced Informatics: Concepts, Theory And Application (ICAICTA), 2016 International Conference On (pp. 1-5). IEEE.

Riloff, E., Qadir, A., Surve, P., De Silva, L., Gilbert, N., & Huang, R. (2013, October). Sarcasm as Contrast between a Positive Sentiment and Negative Situation. In EMNLP (Vol. 13, pp. 704-714).

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

Singh, V. K., Piryani, R., Uddin, A., &Waila, P. (2013, March). 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 (pp. 712-717). IEEE.

Thet, T. T., Na, J. C., &Khoo, C. S. (2010). Aspect-based sentiment analysis of movie reviews on discussion boards. Journal of information science, 36(6), 823-848.

Trivedi, S. K., &Dey, S. (2013a). Effect of Various Kernels and Feature Selection Methods on SVM Performance for Detecting Email Spams. International Journal of Computer Applications, 66(21).

Trivedi, S. K., &Dey, S. (2013b). Interplay between Probabilistic Classifiers and Boosting Algorithms for Detecting Complex Unsolicited Emails. Journal of Advances in Computer Networks, 1(2).

Trivedi, S. K., &Dey, S. (2013c). Effect of feature selection methods on machine learning classifiers for detecting email spams. In Proceedings of the 2013 Research in Adaptive and Convergent Systems (pp. 35-40). ACM.

Trivedi, S. K., &Dey, S. (2013d). An Enhanced Genetic Programming Approach for Detecting Unsolicited Emails. in Proc. 2013 IEEE 16th International Conference on Computational Science and Engineering, Sydney, Australia published by IEEE computer society. 978-0-7695-5096-1/13 $31.00 © 2013 IEEE DOI 10.1109/CSE.2013.171

Trivedi, S. K., &Dey, S. (2014). Interaction between feature subset selection techniques and machine learning classifiers for detecting unsolicited emails. ACM SIGAPP Applied Computing Review, 14(1), 53-61.

Trivedi, S. K., &Dey, S. (2016a). A Comparative Study of Various Supervised Feature Selection Methods for Spam Classification. In Proceedings of the Second International Conference on Information and Communication Technology for Competitive Strategies (p. 64). ACM.

Trivedi, S. K. (2016, September). A study of machine learning classifiers for spam detection. In Computational and Business Intelligence (ISCBI), 2016 4th International Symposium on (pp. 176-180). IEEE.

Tripathi, A., & 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 (pp. 181-185). IEEE.

Trivedi, S. K., &Dey, S. (2016b). A Combining Classifiers Approach for Detecting Email Spams. In Advanced Information Networking and Applications Workshops (WAINA), 2016 30th International Conference on (pp. 355-360). IEEE.

Trivedi, S. K., Trivedi, S. K., Dey, S., &Dey, S. (2016c). A novel committee selection mechanism for combining classifiers to detect unsolicited emails. VINE Journal of Information and Knowledge Management Systems, 46(4), 524-548.

Turney, Peter. 2002. Thumbs up or thumbs down? Semantic orientation applied to unsupervised classification of reviews. In ACL, pages 417–424.

Viola, P., & Jones, M. (2001). Rapid object detection using a boosted cascade of simple features. In Computer Vision and Pattern Recognition, 2001. CVPR 2001. Proceedings of the 2001 IEEE Computer Society Conference on (Vol. 1, pp. I-I). IEEE.

Xia, R., Zong, C., & Li, S. (2011). Ensemble of feature sets and classification algorithms for sentiment classification. Information Sciences, 181(6), 1138-1152.

Ye, Q., Zhang, Z., & Law, R. (2009). Sentiment classification of online reviews to travel destinations by supervised machine learning approaches. Expert systems with applications, 36(3), 6527-6535.

Zhai, Q. G., Jahn, B. M., Zhang, R. Y., Wang, J., & Su, L. (2011). Triassic subduction of the Paleo-Tethys in northern Tibet, China: evidence from the geochemical and isotopic characteristics of eclogites and blueschists of the Qiangtang Block. Journal of Asian Earth Sciences, 42(6), 1356-1370.