World Wide Web
DOI 10.1007/s11280-015-0381-x

# Aspect term extraction for sentiment analysis in large movie reviews using Gini Index feature selection method and SVM classifier

**Asha S Manek**[1] · **P Deepa Shenoy**[2] · **M Chandra Mohan**[3] · **Venugopal K R**[2]

Received: 24 February 2015 / Revised: 14 July 2015 / Accepted: 14 December 2015

**Abstract** With the rapid development of the World Wide Web, electronic word-of-mouth interaction has made consumers active participants. Nowadays, a large number of reviews posted by the consumers on the Web provide valuable information to other consumers. Such information is highly essential for decision making and hence popular among the internet users. This information is very valuable not only for prospective consumers to make decisions but also for businesses in predicting the success and sustainability. In this paper, a Gini Index based feature selection method with Support Vector Machine (SVM) classifier is proposed for sentiment classification for large movie review data set. The results show that our Gini Index method has better classification performance in terms of reduced error rate and accuracy.

**Keywords** Gini Index · Feature selection · Reviews · Sentiment · Support Vector Machine (SVM)

✉ Asha S Manek
ashas100@gmail.com

[1] Research Scholar, Department of Computer Science and Engineering, Jawaharlal Nehru Technological University, Hyderabad, India

[2] Department of Computer Science and Engineering, University Visvesvaraya College of Engineering, Bangalore University, Bangalore 560 001, India

[3] Department of Computer Science and Engineering, Jawaharlal Nehru Technological University, Hyderabad, India

Published online: 04 February 2016

## 1 Introduction

Recent development of the Web has influenced every aspect of our lives and hence need of user view analysis is increasing exponentially. The flow of immense amount of information is effecting decision making processes in organizations. Analysis of people‘s aspects, reactions, emotions, etc. regarding entities such as services, products, issues, events and their attributes based on feedback from Web pages is called opinion mining. Opinion mining is also called as sentiment analysis, opinion extraction, sentiment mining, subjectivity analysis, affect analysis, emotion analysis, review mining, etc. [12]. Opinion mining becomes important for impact analysis and helps in making decisions on constructive developmental directions. It is a research area dealing with usual methods of opinion detection and extraction of sentiments presented in a text. Outcome of implementation of opinion mining methods are formation of efficient recommendation systems, financial study, market research and product growth. There is a enormous amount of opinionated data available in digital forms e.g., reviews, forum discussions, blogs, microblogs, Twitter and social networks [13]. Hence, research in sentiment analysis has an overwhelming impact on NLP, management sciences, political science, economics and social sciences as they are all affected by opinions of people.

A sentiment is a positive or negative opinion, feeling, emotion or assessment about a term, attribute or a feature from an sentiment holder. Positive, negative and neutral views are called as sentiment orientations (also called opinion orientations, semantic polarities or orientations).

In general, opinion mining has been clasified into three levels:

1. *Document level*: Document-level sentiment classification classifies a whole document as a positive or negative sentiment for a product or service. It is not relevant to documents which measures or compare several attributes at this level of analysis because it beleives that each document conveys sentiments on a single attribute (e.g., a single product) [12].
2. *Sentence level*: Sentence-level sentiment classification determines whether each sentence expresses a positive, negative or neutral opinion for a product or service. Sentence level analysis is associated with subjectivity classification which makes distinction between objective sentences and subjective sentences. The objective sentences are those sentences that express true information. The subjective sentences are those sentences that express subjective views and opinions. The objective sentences can imply more opinions than the subjective sentences. e.g., “Few buttons of the remote control of a Smart TV which we purchased a couple of days back are malfunctioning” [12].
3. *Entity and Aspect level*: Feature based opinion mining and summarization is also called as Aspect level analysis. It performs finer-grained analysis. Aspect level analysis is based on the concept that an opinion contains a sentiment (either positive or negative) and a target of opinion [7], thus directly identifies the target of opinion itself.

In many reviews, target opinions are based on aspects and/or their different entities. For example, “Although the battery backup is not that high, I still like Samsung mobile phone ” has a positive sentiment but this sentence is not completely positive. In fact, the positive sentiment is about the entity Samsung mobile (emphasized), but negative sentiment is about its battery backup (not emphasized). Thus, the objective of this level of analysis is to determine opinions on aspects and/or their entities. A structured summary of opinions about entities and their aspects can be produced which converts unstructured text to structured data which can be used for all types of quantitative and qualitative analysis. The aspect level analysis is more challenging and difficult than the document level and sentence level classifications.

Apart from these three levels of classification, regular opinions and comparative opinions are two categories of opinions. A sentiment expressed only on a particular entity or an aspect of the entity is a regular opinion, e.g., “Aamir Khan acts very well” expresses positive sentiment on the aspect of acting of Aamir Khan. A sentiment expressed by comparing multiple aspects based on some of their shared attributes is a comparative opinion [8]. For e.g., “Vanilla pastries taste better than vanilla cake”, compares pastries and cake based on their tastes (an aspect) and expresses feeling and preference for pastries.

Sentiment classification is extremely responsive to the area from which the training data are extracted. This makes it an interesting research topic which transfers learning or domain adaptation. Words and even language formats used in different areas for expressing sentiments can be somewhat different hence a classifier trained using opinionated documents from one area often performs differently from another area when it is tested or applied on opinionated documents. The same word in one area may mean positive, but in another contexual area may mean negative, making matters difficult. Thus, domain modification is needed. It is found that existing research has used labeled data class from one area, unlabeled data class from the target area and general opinion words as features for adaptation [1, 3, 20, 23].

Most recent studies on opinion mining found that sentiment analysis has become an area of active study due to many demanding and interesting research problems. Due to its multiple practical applications, the enormous amount of start-up companies offering sentiment analysis or opinion mining services. Every company wants to know how consumers consider their products and services and those of their competitors. Thus, there is a actual and indeed requirement in the industry for such services. These technical challenges and practical requirement will keep the area active and dynamic for years to come.
There are many applications for Sentiment Analysis. Some of them are:

1. Financial markets:
   (a) To predict society movement based on news, blogs, reviews and social sentiment channel.
   (b) To recognise clients expressing negative emotions in social media or newscast and to raise the business transactions with them for default security.
   (c) To identify sentiments of the analyst and investors’ emotions about the stocks of a company and price trends. It is a crucial information for investors.
2. Computing customer satisfaction metrics:
   To get an idea of how happy customers are with the products, from the ratio of positive to negative reviews.
3. Identifying attackers and advertisers:
   It can be used for providing better consumer service to spot displeasure or problems with goods from customers’ end. It can also be used by analysts to find people who are happy with their products or services and the customers’ experiences can be used to promote their products.
4. Planning for a tourist spot:
   Tourists would like to know the best locations to visit or good restaurants to dine in. Applying opinion mining can assist in retreiving related information for planning a tour.
5. Opinion analysis on elections:
   Opinion analysis can be used to find out voters’ sentiments about a particular contender.

6. Sentiment analysis on softwares or film reviews:
   To identify users‘ opinions from reviews published in specific websites.

The applications for sentiment analysis are endless. Sentiment analysis is in demand because of its efficiency. Thousands of text documents can be processed for sentiment (and other features including named entities, topics, themes, etc.) in seconds, compared to the hours it would take a team of people to process the same manually. Many businesses are adopting text and sentiment analysis and incorporating it into their processes because of its efficiency and accuracy.

In this work, Section 2 gives a brief summary of related work. Section 3 describes the features and existing techniques used in opinion mining. Section 4 presents real motivational example relevant to opinion mining. The details regarding implementation of the proposed framework of opinion extraction and classification and the data set details are explained in Section 5. Results and performance analysis are discussed in Section 6. Conclusion and future work is presented in Section 7.

## 2 Related work

In paper [7], Hu, M. and Liu, B., proposed a set of mining techniques to summarize reviews of products on the basis of data mining and natural language processing methods. The purpose is to provide abstract of more number of customer reviews of a online shopping business based on aspects. They consider that such summarizing will become more crucial as more people are purchasing products online and expressing their feelings on the Web about the product in the form of reviews. Summarizing the reviews is not only helpful for buyers, but also important to product manufacturers and sellers.

Online review system is becoming very useful and serving as a vital source of information for people. As a result, at present computerized review mining system and summarization have become a challenging research area. Unlike traditional text review system, online review mining and summarization intends at obtaining the attributes on which the reviewers express their sentiments and finding out positive or negative opinions. L., Zhuang, et al. [24] recommended a multi-knowledge based method of review mining system for movie reviews. They focused on a particular movie review domain. A multi-knowledge based method is integration of WordNet, statistical analysis and knowledge of movie. In addition, with the proposed approach they claimed that it would be simple to build a summary with names of people in the film industry as sub-headlines and that it would be of considerable interest to movie fans.

Pang, et al. mined film reviews using a variety of machine-learning techniques to study whether they were as efficient for opinion classification as other classification problems such as movie review mining system [17]. They achieved the classification accuracies ranging from 77.4 % to 82.9 % by changing input features (i.e. unigrams, bigrams, unigrams + bigrams).

Pimwadee Chaovalit and Lina Zhou adapted two approaches namely machine learning and semantic orientation in the movie review domain for comparison. They concluded that due to scarcity of words in movie reviews, it is hard to use bag-of-words features using supervised learning methods. Their results were 85.54 % for 3-fold cross validation and 66.27 % when applied on the test data set [4].

In paper [10], Chenghua Lin and Yulan He presented a joint sentiment/topic (JST) model. This model can identify document level sentiment and at the same time extracts mixture of topics from text. The proposed JST model is completely unsupervised and evaluates the performance based on the movie review data set.

Li, F., Han, C., et al. [9] formulated mining the review task as a joint structure tagging problem, focussing on object feature based summarization. A framework based on Conditional Random Fields(CRFs), with features to extract object features and positive and negative opinions simultaneously was proposed. With this framework, chain structure, conjunction structure and syntactic tree structure were investigated by the authors for review mining. The result of their study was a unified model called Skip Tree CRFs for review mining.

In a paper [6] by Yulan He, Chenghua Lin and Harith Alani, the authors investigated polarity-bearing topics of the JST model. They showed that supervised classifiers learned from unique feature space enhanced with polarity-bearing topics achieved good performance on the review data as well as the multi-domain sentiment data set. Moreover, with enhanced feature space and selection based on information gain criteria for cross-domain opinion classification, their model outperformed Structural Correspondence Learning (SCL) algorithm and produced results comparable with Spectral Feature Alignment (SFA) method.

In a paper [11] by Liu, C.L. et al., the authors reported designing and studying a movie-rating and review-summarization system for mobile environment. In their study, rating was based on the results obtained by applying sentiment classification to movie reviews. As product-feature identification is important for feature-based summarization, the authors proposed a method based on Latent Semantic Analysis (LSA) to detect related product features. Opinion words that were identified through a statistical approach and product features were used for feature-based summarization.

In this work, we have collected movie reviews containing rating information from different movie review websites, blogs, discussion forums and social networking sites. Sentiment analysis is performed and performance of the classifier is evaluated. Although this paper focuses on movie reviews, this approach can be applied to reviews of any other domain such as kitchen appliances, books, electronic products, restaurants, etc.

## 3 Features and existing techniques of opinion mining

Researchers used a number of techniques and features to train the opinion mining system [19]. The main task of sentiment classification is to obtain an effective set of features in most machine learning applications. Some of the characteristics are listed below.

- *Terms and their frequency:* Individual words or word n-grams features and their frequency counts are relatively effective in sentiment classification. In some instances, position of words may also be considered. The TF-IDF weighting scheme may also be used for information retrieval.
- *Part of speech:* Many researchers found that adjectives were important indicators of sentiments and may be considered as special features.
- *Opinion words and phrases:* Opinion words and phrases are usually used to express positive or negative opinions. For example, beautiful, wonderful, good and amazing are positive opinion words. Negative opinion words are bad, poor and terrible.

  Despite the fact that many opinion words like rubbish, nonsense and junk are adjectives, nouns, adverbs and verbs (e.g., hate and like) can also indicate sentiments. There are also sentiment phrases and idioms, e.g., live and let live apart from specific words. Sentiment words and phrases are vital to opinion analysis.

- *Negations:* Obviously, negated words are vital for the reason that they change the form of sentiment orientation. For example, the sentence "I don‘t like this movie" is negative opinion. Not all events of such terms mean negation. So negation words must be handled with care. For example, "not" in "not only . . .but also" does not change the sentiment orientation.
- *Syntactic dependency:* Several researchers have also tried extracting features by word dependency in a sentence with the help of parsing and dependency trees.

The numerous custom techniques proposed by the researchers to improve the classification accuracy. These custom techniques used feature weighing method instead of using a standard machine learning methods based on terms in positive and negative reviews for opinion classification, e.g., the score function [5] by Dave, K., et al. In a paper [16] by Paltoglou, G. and Thelwall, M., feature weighting method were used to improve the classification accuracy.

Sentiment classification has also been done based on ratings (e.g., 1-5 stars) of reviews [18] apart from positive or negative opinion classification.

The following points are taken into consideration in understanding opinion mining:

1. The identification of the relevant opinion words and sentiments applied to an entity. (eg. movies, mobiles).
2. The extraction of opinion words at sentence level and applying specific algorithm to reach to document level.
3. There are several words whose polarity varies from one domain to other. So, solution obtained for a given domain (i.e. books reviews) will not work on another domain ( i.e. movie reviews). The model should be adaptive.
4. Sometimes the review writer intentionally sets up context only to confute it at the end. The entire opinion becomes negative because of the important last sentence in spite of the presence of positive opinionated words. This would have been classified as positive in traditional text classification as term frequency is considered more crucial than the presence of the term itself.
5. To distinguish between opinionated and non-opinionated text is subjectivity identification. This is used to improve the performance of a system by including a subjectivity identification module to filter out objective evidences.
6. Even without specific use of any negative word, negation can be expressed in different ways. Negation handling is a challenging job in sentiment analysis.

## 4 Motivation

Sentiment analysis finds a large number of applications and performs very useful task of identifying customer attitude/ opinions about products/services. The feedback about the products or services helps service providers and consumers in taking informed decisions. For example, reviews about restaurants in a city may help a user visiting that city in locating a good restaurant. Similarly, movie reviews help other users in deciding whether the movie is worth watching or not.

The related work section has discussed some of the issues related to sentiment analysis of movie reviews. However, still it is a challenge to improve the sentiment classification performance and opinion detection.

The first challenge faced while carrying out opinion mining on data was in regard of authentication of the end users. We focused to overcome this problem by using larger movie review data sets which have been collected from reliable and well-known sources on the internet. The next vital challenge is to analyze such non-consistent data. People have various ways of expressing sentiments; they may or may not use shorthand, acronyms and proper grammar at different online websites. This is a major challenge faced while processing natural language and analyzing opinions.

The other major challenges faced are applying the right preprocessing filters on sheer volume of data for cleaning and pruning. We used a combination of preprocessing filters, effective code and data transformation techniques to meet these challenges.

However, more number of reviews becomes information overload in the absence of computerised methods for calculating their opinion polarities. This gap can be filled up by producing a opinionated profile from a large number of user reviews about a product or service. The social media is now a major issue on the Web and a large amount of data is unstructured textual data. Thus the opinion mining has become an vital task in text analytics, which promises for a lot of applications.

This work shows that using the training model proposed with SVM classifier described, sentiment classification performance can be effectively improved.

## 5 The proposed approach for opinion extraction and classification

We propose classification tasks to extract the sentiments in five steps shown in Figure 1.

1. Data Source: We have performed experiments on below mentioned corpora and our own data set.
2. Data Preprocessing: It consists of tokenization, transforming cases, stemming and filtering stop words and extraction of opinion oriented words.

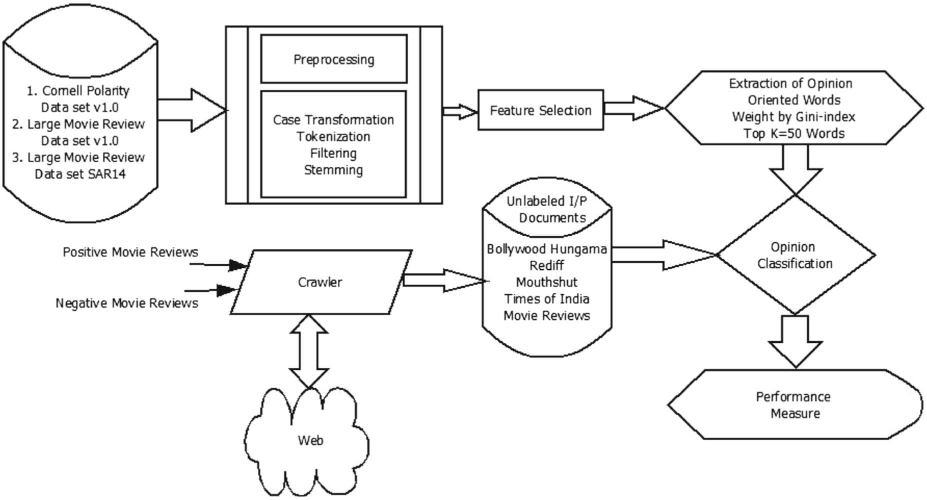

**Figure 1** Steps to extract sentiments

3. Feature Selection: The method of selecting a subset of appropriate features is called feature selection also known as attribute selection. It is generally used for model construction.
4. Representation: The relevance of the attributes based on Gini Index is calculated and weights are assigned to them accordingly.
5. Opinion Classification: It is a classification based on the polarity of the opinion. It takes input as an attribute, selects top $k$ attributes based on weights and term frequencies extracted in the previous step using SVM classifier.

### 5.1 Data source

In this section, we provide brief details of data sets used in our experiments and study. We extracted most recent review messages from websites like Bollywoodhungama, Rediff, Times of India, Rottentomatoes and Mouthshut. Table 1 shows the statistics of our data set.

### 5.2 Data preprocessing

This step involves pre-processing the data in order to make the data ready for analysis.

1. Tokenization: Splitting the text of a document into a sequence of tokens is called tokenization. The data from online reviews contain noise such as URLs, HTML tags, scripts, advertisements and symbols such as asterisks, hashes, etc., which do not have an impact and are not useful for machine learning. These have to be removed in order to keep only the text so as to improve the performance of the classifier.
2. Case Normalization: Transforms all the characters in a document to either lowercase or uppercase. As most of the reviews are in combined case i.e. lowercase and uppercase characters, the process needs to convert entire document into lowercase or uppercase one. Our process turns the entire document or sentences into lowercase one.
3. Stemming: Stemming is a process where the affixes of the word are clipped from the word to make it concise with minimum length, yet having the same meaning.
4. Filtering: This function filters English stop word from a document by removing every token which matches a word from a built-in stop words list. Stop words are words that are not critically necessary to the sentence or opinion.

### 5.3 Feature selection

1. Term Frequency (TF): The term frequency in a document is defined as measure of important terms within a given document. Term Frequency $tf_{t,d}$ of document $d$ and term $t$ is defined as the ratio of the number of occurence of a term in a Web page to the total number of words in that page.
2. Term Frequency - Inverse Document Frequency (TF-IDF): TF-IDF can be applied for filtering stop words in numerous subject areas including text summarization and classification. Term frequency and inverse document frequency are the product of two statistics of TF-IDF. There are numerous methods to find the exact values of both statistics. The term is common or rare across all documents is the inverse document frequency *idf(t,d)*. If a term occurs in all the collected documents, its *idf* is zero.
3. Extraction of opinion oriented words: TF-IDF method is widely used in document classification because it is a simple, straightforward and high processing speed feature-weighting method. But this method simply considers the words of low frequency as

important and the words of high frequency as unimportant which may not be useful as it decreases the precision of classification.

## 5.4 Representation

### *5.4.1 Weight by Gini Index*

A Gini Index based feature selection method solves the problem mentioned above. The experiments showed that the weight by Gini Index method has better classification performance.

Gini Index is an impurity splitting method. It is suitable to binary, continuous numeric type values, etc. It was proposed by Breiman in 1984 and has widely been used in algorithms

**Table 1** Details of data set used

| Data set | No. of reviews | |
|---|---|---|
| | Positive | Negative |
| Cornell sentiment polarity<br>Data Set v1.0 | 700 | 700 |
| Cornell sentiment polarity<br>Data Set v2.0<br>(http://ai.stanford.edu/amaas//data/sentiment/)<br>(Total 3400) | 1000 | 1000 |
| Large movie review<br>Data Set v1.0 IMDB11<br>(http://ai.stanford.edu/amaas//data/sentiment/)<br>(Total 50000) | 25000<br>(12500 Train<br>+12500 Test) | 25000<br>(12500 Train<br>+12500 Test) |
| Large movie review<br>Data Set v1.0<br>IMDB11 unlabeled<br>(http://ai.stanford.edu/amaas//data/sentiment/)<br>(Total 50000) | 50000 | |
| Large movie review<br>Data Set SAR14<br>(https://sites.google.com/site/nquocdai/resources)<br>(Total 233600) | 167378 | 66222 |
| Bollywoodhungama, Rediff,<br>Times of India, Rottentomatoes and<br>Mouthshut movie reviews<br>(http://www.bollywoodhungama.com/movies/reviews<br>http://www.rediff.com/movies/reviews)<br>(http://timesofindia.indiatimes.com/entertainment/movie-reviews,<br>http://www.rottentomatoes.com/,<br>http://www.mouthshut.com/movies.php) | 5370 | |

such as CART, SLIQ, SPRINT and Intelligent Miner decision tree (IBMs Data mining tool), achieving fairly good classification accuracy.

### *5.4.2 Gini Index principle*

The specific algorithm: Suppose the collection of data samples is $S$ of $s$ having $m$ different values of class label attribute which defines different classes of $C_i$, $(i = 1; 2; ...; m)$. According to the class label attribute value, $S$ can be divided into $m$ subsets $(S_i, i = 1; 2; ...; m)$. If $S_i$ is the subset of samples which belongs to class $C_i$, and $s_i$ is the number of the samples in the subset $S_i$, then the Gini Index of set $S$ is

$$GiniIndex(S) = 1 - \sum_{i=1}^{m} {P_i}^2 \tag{1}$$

Where $P_i$ is the probability of any sample of $C_i$, which is estimated by $s_i$/s. When the minimum of *GiniIndex(S)* is 0, i.e. all records belong to the same category at this collection, it indicates that the maximum useful information can be obtained. When all the samples in this collection have uniform distribution for a certain category, *GiniIndex(S)* reaches maximum, indicating the minimum useful information obtained. The initial form of the Gini Index is used to count the “impurity” of attribute for classification. The smaller its value, i.e. the lesser the “impurity”, the better the attribute. On the other hand,

$$GiniIndex(S) = \sum_{i=1}^{m} {P_i}^2 \tag{2}$$

measuring the “purity” of attribute for categorization, the bigger its value, i.e. the better the “purity”, the better the attribute.

## 5.5 Opinion classification

### *5.5.1 Machine learning algorithm*

Machine learning algorithm is one of the most familiar techniques which classifies documents into positive and negative terms. Machine learning algorithm is said to learned from the training data or past experience $E$ with respect to some class of tasks $T$ and performance measure $P$. Its performance measure $P$ at tasks in $T$ improves with experience $E$. Classification problems can be solved by Machine learning in two steps:

1. Learning the model from training data set.
2. Classifying the hidden data on the basis of the trained model.

The way of organizing Machine learning algorithms is based on the desired result of the algorithm or the type of input available during training of the machine.
The following are some of the supervised Machine learning approaches commonly used for sentiment classification of reviews for selected movie reviews.

**(A) Naive Bayes Classification** A Naive Bayes classifier is a simple probabilistic classifier based on Bayes‘ theorem. This classifier assumes that the words are conditionally independent of each other for a given class (positive or negative). The assumption of Naive Bayes classifier is that the occurence or absence of a specific attribute of a class (i.e feature) is not related to the occurence or absence of any other feature. The mathematical formula to compute the probability of a review being positive or negative is

$$P(s|E) = \frac{P(s) * P(E|s)}{P(E)} \tag{3}$$

where $s$ stands for sentiment which can be either positive or negative class and $E$ (evidence) stands for the new movie reviews whose class is being predicted. *P(s)* and *P(E*|s) are obtained during training.
The Naive Bayes classifier requires only a less number of training data to calculate the mean and variance of the variables necessary for classification.

**(B) Support Vector Machine (SVM)** In Machine learning, Support Vector Machines (SVMs) are supervised learning models with associated learning algorithms. SVM analyzes data and recognize patterns which are used for classification and regression analysis. Given a set of training samples, each assigned with one of two classes. A SVM learning algorithm builds a model that marks new samples belonging to one class or the other class. A SVM learning algorithm model is a representation of the samples as points in space. These points are mapped such that the samples of an other class are seperated by a wider gap. Further, new samples are mapped into that same space and predicted to belong to a class based on which side of the gap they fall on.

SVMs are capable of performing a non-linear classification in addition to performing linear classification. Non-linear classification can be performed by using the kernel trick i.e. mapping their inputs into high-dimensional feature spaces. A better separation is obtained by the hyperplane that has the longest distance to the nearest training data point of any class.

**Algorithm 1** Pseudocode for the proposed framework

```
For each review in the data set
Do {
Call Crawler to extract review content from each page
For all the pages in the data set call Preprocessing()

Apply a Gini Index based feature selection method using this formula.

    $Gini(t) = \sum_{i=1}^{m} P(C_i|t)^2$

where, $P(C_i|t)$ is the probability of occurrence of each class in the
document set showing the feature $t$.
Select higher weight attributes for classification i.e all top $k$ attributes
}
Train model using SVM classifier.
```

The hyperplanes in a higher-dimensional space are the set of points. These set of points dot product with a vector in that space is constant. Vectors that define the hyperplanes may be linear combinations with parameters $\alpha_i$ of images of feature vectors that occur in the database. The points in the feature space that are mapped into the hyperplane are defined by the relation: $\sum_i \alpha_i K(x_i, x) = constant$ where $K(x_i, x)$ is kernel function.

### 5.6 Evaluation methods

A major aspect of Machine learning algorithm is evaluating the performance and efficiency of any model. The evaluation approaches used in our work are discussed below. Split Validation is performed in order to evaluate the performance of a learning operator. When an explicit testing set is not available Split Validation identifies the fitness of a model to a hypothetical testing set.

The Split Validation operator can use several types of sampling for building the subsets. The various types of sampling methods are linear sampling, shuffled sampling and stratified sampling.

In our model, relative linear and stratified sampling methods with 0.85 and 0.9 sample ratio respectively are used for building subset for classification. Linear sampling divides the input data set into partitions without changing the order of the examples i.e. subsets with consecutive examples are created. Stratified sampling also known as Bootstrap Sampling builds random subsets. It ensures that the distribution of class in the subsets is the same as in the whole input data set. For example stratified sampling builds random subsets in a binomial classification in such a way that each subset consist of approximately the same proportions of the two values of class labels.

## 6 Results and performance analysis

### 6.1 RapidMiner

RapidMiner is an open source software platform with integrated environment for machine learning, data mining, text mining, predictive and business analytics. It is possible to extend RapidMiner functionality with additional plugins. The RapidMiner Extensions Marketplace serves as a platform for developers to create and publish data analysis algorithms to a broader community.

### 6.2 Evaluation related terminologies

For the purpose of evaluation, several terminologies are used in this paper. The following definitions are the terms that are used or mentioned to measure performance of the classification approaches:

- Accuracy: Number of correctly classified documents divided by the total number of documents.
- Error rate: Number of incorrectly classified documents divided by the total number of documents $(1.0 - Accuracy)$.
- Recall of a class: Number of documents of a class correctly classified.
- Precision of a class: Number of correct predictions for a class.

In (Table 2), the accuracy of the classifier is: Accuracy $= (42 + 36)/100 = 78\ \%$
The error rate is: Error Rate $= 1.0 - Accuracy = 22\ \%$

**Table 2** Confusion matrix

| Classified as | True positive | True negative |
|---|---|---|
| Pred. Positive | 42 | 20 |
| Pred. Negative | 2 | 36 |

**Table 3** Results of the proposed model for cornell sentiment polarity data set [15]

| Classifier | SVM (Linear) | | | | |
|---|---|---|---|---|---|
| Weight by method | Accuracy % | Precision % | Recall % | Error-rate % | F-measure % |
| Maximum Relevance | 96.95 | 100 | 96.95 | 3.05 | 98.44 |
| Correlation | 97.25 | 100 | 97.25 | 2.75 | 98.60 |
| Chi-Squared Statistic | 85.19 | 100 | 85.19 | 14.81 | 91.98 |
| Info Gain | 94.12 | 100 | 94.12 | 5.88 | 96.96 |
| Gini Index | 92.81 | 100 | 92.81 | 7.19 | 96.26 |
| Classifier | Naive Bayes | | | | |
| Maximum Relevance | 85.19 | 100 | 85.19 | 14.81 | 91.98 |
| Correlation | 87.06 | 100 | 87.06 | 12.94 | 93.08 |
| Chi-Squared Statistic | 85.62 | 100 | 85.62 | 14.38 | 92.23 |
| Info Gain | 84.53 | 100 | 84.53 | 15.47 | 91.59 |
| Gini Index | 83.88 | 100 | 83.88 | 16.12 | 91.21 |

Recall of positive class: $Recall_{positive} = 42/62 = 67.7\ \%$
Precision of positive class: $Precision_{positive} = 42/44 = 95.4\ \%$

The results and performance shown in Table 3 are best proved with SVM Linear algorithm with an Accuracy of 97.25 %, Precision of 100 % and Recall rate of 97.25 % for 1700 positive and 1700 negative movie reviews. Similarly Table 4 shows Accuracy of 94.46 %, Precision of 100 % and Recall rate of 94.46 % for large movie review data set v1.0 (http://ai.stanford.edu/amaas//data/sentiment/). The word list is generated through the pre-processing technique and weight of each attribute is calculated using squared correlation feature extraction method by selecting “top $K$” attributes by weight. Selected “top $K$” attributes whose

**Table 4** Results of the proposed model for large movie review dataset V1.0 (http://ai.stanford.edu/amaas//data/sentiment/)

| Classifier | SVM (Linear) | | | | |
|---|---|---|---|---|---|
| Weight by method | Accuracy % | Precision % | Recall % | Error-rate % | F-measure % |
| Maximum relevance | 92.68 | 100 | 92.68 | 7.32 | 96.21 |
| Correlation | 92.68 | 100 | 92.68 | 7.32 | 96.21 |
| Chi-squared statistic | 94.19 | 100 | 94.19 | 5.81 | 97.00 |
| Info gain | 94.43 | 100 | 94.43 | 5.57 | 97.13 |
| Gini index | 94.46 | 100 | 94.46 | 5.54 | 97.14 |
| Classifier | Naive Bayes | | | | |
| Correlation | 87.02 | 100 | 87.02 | 12.98 | 93.07 |
| Chi-squared statistic | 87.35 | 100 | 87.35 | 12.65 | 93.26 |
| Info gain | 87.97 | 100 | 87.97 | 12.03 | 93.61 |
| Gini index | 87.50 | 100 | 87.50 | 12.5 | 93.34 |

**Table 5** Results of the proposed model for large movie review data set SAR14 (https://sites.google.com/site/nquocdai/resources)

| Classifier | SVM (Linear) | | | | |
|---|---|---|---|---|---|
| Weight by method | Accuracy % | Precision % | Recall% | Error-Rate % | F-measure % |
| Maximum Relevance | 97.30 | 100 | 97.30 | 2.7 | 98.63 |
| Correlation | 95.21 | 100 | 95.21 | 4.79 | 97.54 |
| Chi-Squared Statistic | 94.96 | 100 | 94.96 | 5.04 | 97.41 |
| Info Gain | 95.82 | 100 | 95.82 | 4.18 | 97.86 |
| Gini Index | 97.32 | 100 | 97.30 | 2.68 | 98.65 |

weight fulfill a given weight relation (K=20) with respect to the input attribute weight are evaluated using SVM Linear and Naive Bayes classifiers. The model is cross-validated using Split Validation operator, considering the split ratio as 0.85 for training samples and the sampling type as linear.

SAR14 movie review data set is used in this experiment using Gini Index feature selection method for sentiment analysis which produce good results in comparison with other methods. Results show that proposed approach improved accuracy in many times particularly while using SVM linear classifier with Split Validation of 0.85. Table 5 shows results of classification using different weighting methods on SAR14 movie reviews.

We have exploited the proposed method of feature extraction in sentiment analysis by using a real data set of moderate size collected by crawling the Web on our own. To evaluate the real performance, we collected 5370 reviews from five different movie review database websites, namely, Bollywoodhungama, Rediff, Times of India, Rottentomatoes and Mouthshut. We labeled all these reviews using our model and classified positive and negative sentiment polarity as shown in Table 6.

### 6.3 Performance evaluation

(A) Comparison of our approach with work by Pham, S.B., et al. [21]

**Table 6** Results of the proposed model for various movie review data set (http://www.bollywoodhungama.com/movies/reviews,http://www.rediff.com/movies/reviews), (http://timesofindia.indiatimes.com/entertainment/movie-reviews, http://www.rottentomatoes.com/, http://www.mouthshut.com/movies.php)

| Data Set | Sentiment | Documents |
|---|---|---|
| Bollywoodhungama movie reviews | Positive | 1309 |
| | Negative | 561 |
| Rediff movie reviews | Positive | 290 |
| | Negative | 70 |
| Times of India movie reviews | Positive | 2650 |
| | Negative | 40 |
| Rottentomatoes movie reviews | Positive | 290 |
| | Negative | 40 |
| Mouthshut movie reviews | Positive | 100 |
| | Negative | 20 |

**Table 7** Comparison of performance of proposed approach with Pham, S.B., et al. [21]

| | Data set used | Classifier | Feature selection method | Accuracy |
|---|---|---|---|---|
| Pham, S.B., et al. [21] | 2000 Reviews<br>proposed by<br>Pang, B. and Lee, L. [15] | LIBLINEAR | N-gram features and rating based features | 91.6 % |
| | 50000 reviews<br>created by Maas et al. (http://ai.stanford.edu/amaas//data/sentiment/) | | | 89.87 % |
| | 233000 reviews<br>created by SAR14 (https://sites.google.com/site/nquocdai/resources) | | | 93.24 % |
| Proposed approach | 2000 reviews<br>proposed by Pang, B. and Lee, L. [15] | SVM(Linear) | Weight by Correlation | 96.95 % |
| | 50000 reviews<br>created by Maas et al. (http://ai.stanford.edu/amaas//data/sentiment/) | | Weight by Gini Index | 94.46 % |
| | 233000 reviews<br>created by SAR14 (https://sites.google.com/site/nquocdai/resources) | | | 97.32 % |

Table 8 Comparison of performance of proposed approach with Basari., et al. [2]

| | Data set used | Classifier | Feature selection method | Accuracy |
|---|---|---|---|---|
| Basari et al. [2] | RAW data from twitter messages for movie reviews in the Web site http://www.standford.edu/~aleemgo/cs224n/trainingandtestdata.zip | Support Vector Machine with Particle Swarm Optimization (SVM-PSO) | TF and TF-IDF | 77 % |
| Proposed approach | 3400 reviews proposed by Pang, B. and Lee, L. [15] | SVM(Linear) | Weight by Correlation | 96.95 % |
| | 50000 reviews created by Maas et al. (http://ai.stanford.edu/amaas//data/sentiment/) | | Weight by Gini Index | 94.46 % |
| | 233000 reviews created by SAR14 (https://sites.google.com/site/nquocdai/resources) | | | 97.32 % |

**Table 9** Comparison of performance of proposed approach with Weichselbraun A., et al. [22]

| | Data set used | Classifier | Feature selection method | Accuracy |
|---|---|---|---|---|
| Weichselbraun A., et al. [22] | 2000 reviews from IMDB comedy | Naive Bayes | SentiNet to transform into sentiment lexicon | 83 % |
| | 2000 Reviews from IMDB crime | | | 73% |
| | 2000 reviews from IMDB drama | | | 93.24 % |
| Proposed Approach | 3400 reviews proposed by Pang, B. and Lee, L. [15] | SVM(Linear) | Weight by Correlation | 96.95 % |
| | 50000 Reviews created by Maas et al. (http://ai.stanford.edu/amaas//data/sentiment/) | | Weight by Gini Index | 94.46 % |
| | 233000 reviews created by SAR14 (https://sites.google.com/site/nquocdai/resources) | | | 97.32 % |

In an experimental study conducted by Pham, S.B., et al. [21] on sentiment polarity classification, rating-based feature was estimated based on regression model learned from data set SAR14 of 233600 movie reviews and accuracy of 93.24 % was reported. They also examined the contribution of the rating-based feature and N-grams in a machine learning-based approach on two data set PL04 and IMDB11 and got accuracies 91.6 % and 89.87 % respectively.

SVMs achieve significantly higher search accuracy than traditional query refinement schemes after just three to four rounds of relevant feedback. Hence, SVM is selected as machine learning method in our proposed model. Besides, we had worked on movie review Cornell data set v1.0 and in this work [14] SVM Linear outperformed the other classifiers. We apply Gini Index feature selection to reduce the dimension of term and help in selecting the best combination of parameter settings to reduce the high dimension of features and obtain a good classification result for opinions.

The response of our approach when used with Support Vector Machine (SVM) Linear classifiers with split ratio of 0.85 with weight by Gini Index feature selection method is sensitive and accurate in sentiment analysis when compared with N-gram and rating based features using LIBLINEAR proposed in work by Pham, S.B., et al. [21]. Table 7 illustrates the same.

(B) Comparison of our model with work by Basari., et al. [2]

This study has shown that Particle Swarm Optimization(PSO) affect the accuracy of SVM after the hybridization of SVM-PSO. The best accuracy level given in this study is 77 % which had been achieved by SVM-PSO after data cleansing. On the other hand, the accuracy level of Machine learning-based approach shows the effectiveness of our proposed approach using various data set in Table 8.

(C) Comparison with work by A. Weichselbraun., et al., [22]

This research introduces a novel method using an enriched version of SenticNet for polarity classification. The accuracy values of Table 9 shows the performance of the proposed approach with work done by A. Weichselbraun., et al., [22]. Table 9 shows our proposed model achieved good classification accuracy, approaching the performance of supervised sentiment classification technique.

## 7 Conclusion and future work

It is observed that sentiment analysis has become an important and essential area of Web data mining and has attracted lots of research interest over the past decade. Though there are many algorithms and techniques that are available to analyze the sentiments, there are no techniques that can provide a solution.

In this study, we proposed a statistical method using weight by Gini Index method for feature selection in sentiment analysis while at the same time improving the accuracy of sentiment polarity prediction using various large movie data set. Our proposed framework for sentiment analysis using SVM classifier is compared with other feature selection methods on movie reviews and results have shown that classification by using this efficient and novel method has improved the accuracy.

More future research is needed to solve the opinion mining challenges in the form of identifying what a noun and pronoun is, what a phrase refers to, meaning of text containing sarcastic sentences or hidden emotions, linguistic issues, opinion spamming and variations over time. Sometimes a single sentiment word may convey an opposite polarity, depending on the context. In interrogative and conditional sentences, it is often observed that a sentence containing sentiment word does not express any opinion. In other cases, some sentences without any sentiment word may express opinions. Finer points such as these need to be addressed in future research work.

## References

1. Aue, A., Gamon, M.: Customizing Sentiment Classifiers to New Domains: A Case Study. In: Proceedings of Recent Advances in Natural Language Processing (RANLP-2005) (2005)
2. Basari, A.S.H., Hussin, B., Ananta, I., Zeniarja, J.: Opinion Mining of Movie Review using Hybrid Method of Support Vector Machine and Particle Swarm Optimization. Procedia Engineering, 453–462 (2013)
3. Blitzer, J., Dredze, M., Pereira, F.: Biographies, Bollywood, Boomboxes and Blenders: Domain Adaptation for Sentiment Classification. In: Proceedings of Annual Meeting of the Association for Computational Linguistics (ACL-2007) (2007)
4. Chaovalit P., Zhou, L.: Movie Review Mining: A Comparison between Supervised and Unsupervised Classification Approaches. In: Proceedings of the 38th Hawaii International Conference on System Sciences (2005)
5. Dave, K., Lawrence, S., Pennock, D.: Mining the Peanut Gallery: Opinion Extraction and Semantic Classification of Product Reviews. In: Proceedings of International Conference on World Wide Web (WWW-2003) (2003)
6. He, Y., Lin, C., Alani, H.: Automatically Extracting Polarity-bearing Topics for Cross-domain Sentiment Classification. In: Proceedings 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies, pp. 19–24, Portland (2011)
7. Hu, M., Liu, B.: Mining and Summarizing Customer Reviews. In: Proceedings of the tenth ACM SIGKDD international conference on Knowledge Discovery and Data Mining (168–177). ACM (2004)
8. Jindal, N., Liu, B.: Mining comparative sentences and relations. In AAAI **22**, 1331–1336 (2006)
9. Li, F., Han, C., Huang, M., Zhu, X., Xia, Y.J., Zhang, S., Yu, H.: Structure-aware Review Mining and Summarization. In: Proceedings of the 23rd International Conference on Computational Linguistics (pp. 653-661). Association for Computational Linguistics (2010)
10. Lin, C., He, Y.: Joint Sentiment/Topic Model for Sentiment Analysis. In: Proceedings of the 18th ACM conference on Information and Knowledge Management (pp. 375–384). ACM (2009)
11. Liu, C.L., Hsaio, W.H., Lee, C.H., Lu, G.C., Jou, E.: Movie rating and review summarization in mobile environment, Systems, Man and cybernetics, Part C: Applications and reviews. IEEE Trans., 397–407 (2012)
12. Liu, B.: Sentiment Analysis and Opinion Mining, p. 7. Morgan and Claypool Publishers, USA (2012)
13. Liu, B.: Sentiment analysis and opinion mining. Synth. Lect. Human Lang. Technol. **5.1**, 1–167 (2012)
14. Manek, A.S., Pallavi, R.P., Bhat, V.H., Shenoy, P.D., Chandra Mohan M., Venugopal, K.R., Patnaik, L.M.: SentReP: Sentiment Classification of Movie Reviews using Efficient Repetitive Pre-Processing. In: TENCON 2013-2013 IEEE Region 10 Conference (31194), pp. 1–5. IEEE (2013)
15. Movie review data set [online], Available http://www.cs.cornell.edu/people/pabo/movie-review-data/
16. Paltoglou, G., Thelwall, M.: A Study of Information Retrieval Weighting Schemes for Sentiment Analysis. In: Proceedings of Annual Meeting of the Association for Computational Linguistics (ACL- 2010) (2010)
17. Pang, B., Lee, L., Vaithyanathan S.: Thumbs up? Sentiment Classification using Machine Learning Techniques. In: Conference on Empirical Methods in Natural Language Processing (EMNLP) (2002)
18. Pang, B., Lee, L.: Seeing Stars: Exploiting Class Relationships for Sentiment Categorization with Respect to Rating Scales. In: Proceedings of Meeting of the Association for Computational Linguistics (ACL-2005) (2005)
19. Pang, B., Lee, L.: Opinion mining and sentiment analysis. Found. Trends Inf. Retr. **2**(1–2), 1-135 (2008)
20. Pan, S., Ni, X., Sun, J., Yang, Q., Chen, Z.: Cross-domain Sentiment Classification via Spectral Feature Alignment. In: Proceedings of International Conference on World Wide Web (WWW-2010) (2010)
21. Pham, S.B. et al.: Sentiment Classification on Polarity Reviews: An Empirical Study Using Rating-based Features. In: Proceedings of the 5th Workshop on Computational Approaches to Subjectivity, Sentiment and Social Media Analysis, p. 128135. Association for Computational Linguistics, Maryland, USA (2014)

22. Weichselbraun A. et al.: Enriching Semantic Knowledge Bases for Opinion Mining in Big Data Applications, Knowledge Based Systems. doi:10.1016/j.knosys.2014.04.039 (2014)
23. Yang, H., Si, L., Callan, J.: Knowledge Transfer and Opinion Detection in the TREC2006 Blog Track. In: Proceedings of TREC (2006)
24. Zhuang, L., Jing, F., Zhu, X.Y.: Movie Review Mining and Summarization. In: Proceedings of the 15th ACM international conference on Information and Knowledge Management (pp. 43–50). ACM (2006)