Investigation of Movie Review Data Set by the Use of Opinion Mining

Manoj kumar[1], Mohit Dayal[2] , Jitender kumar[3] Anant Rajee Bara[4]
[1]Bhai Parmanand Institute of Business Studies, Shakarpur, Delhi india.
[2]Chandigarh University, Mohali, Punjab, India.
[3]Rawal Institute of Engineering and Technology, Faridabad Haryana, India.
4 Innov4Sight Health and Biomedical Systems Private Limited, Bangluru, India.

E-Mail: manoj.g2408@gmail.com

Abstract

With the growth of the internet and its users over the past decade, there is an extremely large magnitude of data online. There are various websites and forums which provide options to review a particular product or place. These reviews are helpful to all the customers, who are thinking of buying that product, as well as the organisation building that product, so they can improve its flaws. The opinions and sentiments gathered through online forums and review sites help in a more accurate way of understanding the customer's preferences and tastes. In this study, our objective is to determine the underlying opinion in the movie reviews which could be positive, neutral or negative, through various machine learning algorithms like Support Vector Machine (SVM), Decision trees and Naive Bayes classifier and to compare the results given by each algorithm. Further, we also appraise the role of pre-processing online reviews before polarising them. The outcome of this study shows the different accuracies given by the various classifiers applied to the same movie review dataset.

Keywords: Sentiment Analysis, Data Pre-processing, Feature Extraction, Mining

Introduction

The understanding and extraction of the prevailing opinion behind a sentence or paragraph are referred to as sentiment analysis of a text. It is also referred to as *opinion mining* or *sentiment mining*. Sentiment Analysis is the computational study of people's opinions, attitudes, and emotions toward an entity which can be individuals, events or topics [1].

The actions and decisions made by people today are mostly driven by someone's view or influence. Whether one likes to accept it or not, but public opinion on matters does affect our view. Before visiting a particular place or experiencing something new, we always want to look up to other people's views for assurance or advice. Nobody would like to waste their time and resources to visit any place which hasn't been well-reviewed. Therefore, review writing platforms have been established by almost all websites.

Through the spread of opinion and review writing on online platforms, there is a plethora of text being generated. Apart from helping the users in making a better and informed choice while buying a product, opinion mining also proves useful in judging the success or failure of a new product being launched [2]. By understanding the sentiments and by being in touch with its customers any organization can benefit better as the business of the product goes up if the consumers would find the information that would be useful for them in making a decision.

Say, a movie is released and it starts receiving a lot of mentions all over the web. Does it mean the movie was a big hit? No. On the contrary, it could also mean that it is receiving criticism and negative comments. But text can be misleading without understanding the sentiment behind it. There can varied number of ways in how a reviewer would like to express himself. There is plenty of difference between " This part was better than the last one" and " It could have been better. Now, to extract useful information from that pool of data is very necessary, for the review sites to fulfil their purpose. It is hard for the companies to manually perform the task of analysis and extract the trend in opinions, thus the requirement of an automated system surfaces. In the past, various models and techniques have been discussed which show different accuracies under varying conditions. Some supervised learning techniques being k-NN, Support Vector Machine (a non-probabilistic binary linear classifier), Decision Trees, and Naive Bayes Classifier (a probabilistic model), etc. This study makes a comparison between different types of techniques used for the automated classification of movie reviews.

Literature Review

The researches based on sentiment analysis have seen an increase in their number and popularity. In the past, many studies have been carried out using several methodologies and techniques.. [1] gives a detailed study of all the main algorithms which can be used for sentiment analysis. The author also talks about the applications in his study. [3] talks about the significance of the pre-processing of data by drawing a comparison in the accuracies before and after the data transformations are done. The accuracy increases from 78.33% to 81.5% after the data is pre-processed. [10] proposes a new improved model as a classifier for movie review data over existing Naive Bayes and Support Vector Machine giving accuracies of 91.15% and 91.35% respectively.

The proposed hybrid NB-SVM method promises an accuracy of 94.15%. [11] works on text categorization of reviews for opinion mining using the two standard techniques Naive Bayes and SVM. The result shows that the NB classifier has higher accuracy over SVM. [12] used SVM and Maximum Entropy to achieve accuracy up to 95% on twitter sentiment mining. [13] achieved an accuracy of 92% using frequent itemset mining and Naive Bayes classifier on product reviews.

Similarly, [14] applies supervised opinion mining techniques on online user reviews using Naive Bayes and n-grams achieving an accuracy of 80%. [15] compares the accuracy of SVM and NB using bigram and trigram in their work. [16] also does a comparative study between the working of NB and Apache Hadoop on movie reviews and shows that the Map-Reduce of Apache Hadoop performs better than NB Classifier. [2] shows that SVM outperforms the NB and k-NN classifiers while showing an accuracy of 80% on movie reviews. In the upcoming sections of the study, our methodology and results are discussed.

Flow of the Proposed Methodology

A) Data Pre-Processing And Data Transformations

Any data collected online contains some unwanted and futile text. The removal of such text, and thus cleaning the data prior to classification is termed as Pre-processing of data. XML/HTML tags, advertisements, special symbols create no meaning but noise in the text. Removing such uninformative parts from the text increases the accuracies of the classification and accelerates the process [3]. The fundamental steps involved in processing of data involve tokenisation, stop-words removal, case normalisation, stemming, lemmatization, removal of symbols and non-English letters. The steps mentioned earlier can be easily performed through Natural Language Toolkit (NLTK).

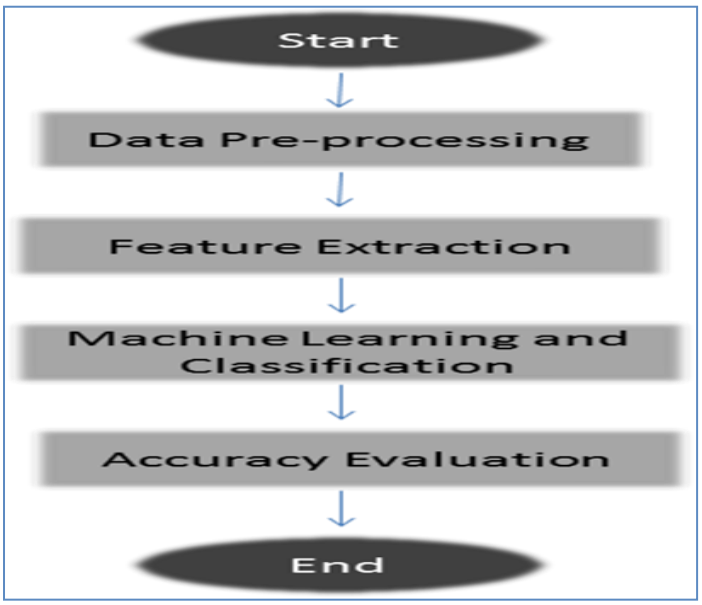

Figure 1: Proposed Methodology

Case Normalization

The dataset contains text in both lower and upper case characters so converting all the sentences into the same case (here, lower case) is referred to as case normalisation. This process converts all review documents into lower case.

Removal of abnormal white spaces and special symbols

The data collected online normally has plenty of dispensable spaces and special characters. They can be easily eliminated with the help of regular expressions, leaving behind only English characters in the documents.

Tokenisation

The reviews are in form of paragraphs is different documents. Tokenization is the process of splitting up the paragraphs of text into personal terms or tokens. For the English language, white space can be used as a delimiter to separate different tokens in a sentence [4].

###Example: this is an example will be tokenized as [this, is, an, example]

Stop words filtering

The data contains words which are insignificant and unhelpful in determining the polarity of the review, such words are named as stop words. Therefore, a list of the most occurring stop words was made to remove their existence from the text. The word list includes the usual stop words like, is, the, a, an etc as well as some movie review specific words like actor, actress, scene which have high frequency but are useless for retrieving any information [3].

Stemming

Stemming refers to the process of reducing an extended word to its original word or stem by trimming it. The significant tokens are reduced to a single type by eliminating prefixes and suffixes wherever required.

###Example: stemmed, stemmer, stemming are all reduced to stem

Sometimes, in search for the stem, the word is reduced to something unintelligible, which is not helpful.

Lemmatization

A lemmatizer does a full morphological analysis to precisely spot the lemma for every token. It uses vocabulary and attempts to eliminate endings and hence, returning tokens to their dictionary form [5]. It produces modest results in retrieving information.

B) Feature Extraction

Bag of Words Model

In text classification, one of the prime tasks is the representation of the document. The Bag of Words representation is the most elementary and extensively used approach [6]. In this method, the frequency of words is kept in check, being indifferent even to sequence of the words. The representation is done by a vector of word counts of the document. Regardless of being an easy approach, the classification models applying this method commonly attain high performance [6].

Classification Techniques

Support Vector Machine (SVM):

SVM is a linear binary classifier which is seen as a highly potent text classifier. It predicts the class to which each of the input belongs. SVM classifies a new test example by putting it under one of the two categories the training data belonged to. The main objective of the training data is to help draw a hyperplane which is of maximum margin [7]. New examples are then given their category based on which side of the plane they lie in. There are a set of hyperplanes drawn but the plane having the largest distance to the nearest training example (of any class) is considered as the best [8].

Decision Trees Classifiers

These classifiers fragment the training data space in a hierarchical way in which division of data depends on the result of a condition on the attribute value [9]. The division of data is done in a recursive manner till a decision is reached at the leaf nodes(classes) for the classification [1]. Each leaf node has a class label, which is determined by the majority of the training examples reaching that particular leaf. Each internal node is a question of features, as it branches out according to the answers.

Naive Bayes Classifier

Bayesian classifier is completely based upon the Bayes Theorem of probability. It has a strong hypothesis of class conditional independence saying that a particular feature is independent of the value of any other feature, given the class variable. According to the theorem, the probability that needs to be computed P (A|X) can be expressed in terms of probabilities P (A), P (X|A) as follows.

$$P(A|X) = \frac{P(X|A)P(A)}{P(X)} \quad (1)$$

P (A) is the a priori probability of A, P (X) is the a priori probability of X

Similarly, P (X|A) is the a posteriori probability of X conditioned on A.

Dataset used: The dataset used for the analysis in this study has been obtained from the Cornell univerisity dataset repository. The dataset is *polarity dataset v1.0* which contains 7000 positive as well as 700 negative movie reviews.

Results

There are plenty of performance evaluation techniques. The evaluation used in this study finds the accuracy in terms of Precision, Recall and F1-score. These terms can be calculated based on the formulas given ahead.

$$precision\,(p) = \frac{tp}{tp+fp} \qquad (2)$$

$$recall(r) = \frac{tp}{tp+fn} \qquad (3)$$

$$f1 - score = 2 * p * \frac{r}{p+r} \qquad (4)$$

Where, tp denotes true positive(orignal-positive, classified-positive), fp denotes false positive(orignal-negative ,classified-positive), fn denotes false negative(orignal-positive ,classified-negative).

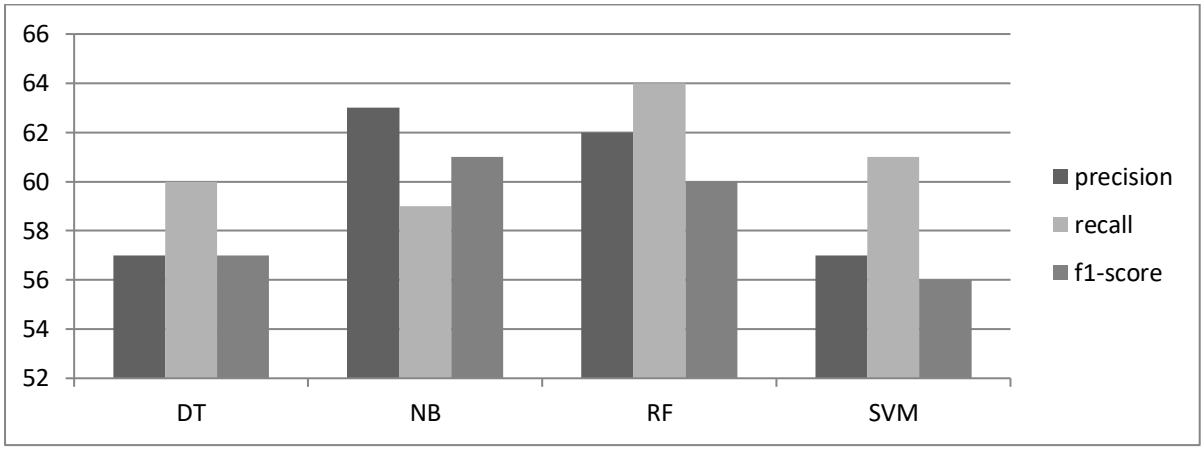

Figure 2: Graph showing results of Decision Tree, Naive Bayes, Random Forest and Support Vector Machine Classifiers

Conclusion

In this paper, we compared techniques like Decision Tree, Naive Bayes, Random Forest and Support vector Machine classifiers. The motive of this study was to compare the different accuracies given by the different types of classifiers on the dataset containing a total of 1400 movie reviews. The result shows that the Naive Bayes and Random forest classifiers perform better than the other two classifiers according to our evaluation.

References

1. Walaa Medhat, Ahmed Hassan, Hoda Korashy, " Sentiment analysis algorithms and applications:A survey" , Ain Shams Engineering Journal (2014) 5, 1093– 1113
2. P.Kalaivani, Dr. K.L.Shunmuganathan," Sentiment Classification Of Movie Reviews By Supervised Machine Learning Approaches" , P.Kalaivani et.al / Indian Journal of Computer Science and Engineering (IJCSE)(2013)
3. Emma Haddia, Xiaohui Liua, Yong Shib, " Role of Text Pre-processing in Sentiment Analysis" , Information Technology and Quantitative Management (ITQM2013)

4. Abd. Samad Hasan Basaria, Burairah Hussina, I. Gede Pramudya Anantaa, Junta Zeniarjab, " Opinion Mining of Movie Review using Hybrid Method of Support Vector Machine and Particle Swarm Optimization" , Malaysian Technical Universities Conference on Engineering & Technology 2012, MUCET 2012
5. Vimala Balakrishnan, Ethel Lloyd-Yemoh, " Stemming and Lemmatization: A Comparison of Retrieval Performances" , *Lecture Notes on Software Engineering, Vol. 2, No. 3, August 2014*
6. Constantinos Boulis, Mari Ostendorf," Text Classification by Augmenting the Bag-of-Words Representation with Redundancy-Compensated Bigrams,
7. Jayashri Khairnar, Mayura Kinikar, " Machine Learning Algorithms for Opinion Mining and Sentiment Classification " , International Journal of Scientific and Research Publications, Volume 3, Issue 6, June 2013
8. Chetashri Bhadane,Hardi Dalal, Heenal Doshi," Sentiment analysis: Measuring opinions" ,International Conference on Advanced Computing Technologies and Applications (ICACTA-2015)
9. Quinlan JR." Induction of decision trees. Machine Learn" 1986;1:81– 106.
10. M . Govindarjan," Sentiment Classification Of Movie Reviews Using Hybrid Method" , International Journal of Advances in Science Engineering and Technology, ISSN: 2321-9009 Volume- 1, Issue-3, Jan.-2014
11. Humera Shaziya , G.Kavitha , Raniah Zaheer," Text Categorization of Movie Reviews for Sentiment Analysis" , International Journal of Innovative Research in Science,Engineering and Technology Vol. 4, Issue 11, November 2015
12. M.Ravichandran, G.Kulanthaivel," Twitter Sentiment Mining (Tsm) Framework Based Learners Emotional State Classification And Visualization For E-Learning System" , Journal of Theoretical and Applied Information Technology, 2014
13. A.Jeyapriya, C.S.Kanimozhi Selvi," Extracting Aspects And Mining Opinions In Product Reviews Using Supervised Learning Algorithm" , IEEE, 2015
14. Ion Smeureanu, Cristian Bucur ," Applying Supervised Opinion Mining Techniques On Online User Reviews", Informatica Economică, 2012
15. Jalel Akaichi, Zeineb Dhouioui, Maria Jose Lopez-Huertas Perez,"Text Mining Facebook Status Updates for Sentiment Classification, 2013", Proceedings of System Theory, Control and Computing(ICSTCC), 2013 17th International ConferenceMr. B. Narendra, Mr. K. Uday Sai, Mr. G. Rajesh, Mr. K. Hemanth, Mr. M. V. Chaitanya Teja, Mr. K. Deva Kumar," Sentiment Analysis on Movie Reviews: A Comparative Study of Machine Learning Algorithms and Open Source Technologies " , *I.J. Intelligent Systems and Applications,* 2016, 8, 66-70 Published Online August 2016 in MECS.

Source of Support: Nil
Conflict of Interest: None