ISSN (Print) : 0974-6846
ISSN (Online) : 0974-5645
Indian Journal of Science and Technology, *Vol 12(41), DOI: 10.17485/ijst/2019/v12i41/145554, November 2019*

Movie Reviews Classification Using Sentiment Analysis

Somya Dwivedi, Harsh Patel and Shweta Sharma*

Department of Computer Science & Engineering, Faculty of Engineering & Technology, Manav Rachna International Institute of Research & Studies, Faridabad, Haryana, India; Somyadwivedi21@gmail.com, harrypatel199298@gmail.com, Shweta.fet@mriu.edu.in

Abstract

Background/objectives: Movie reviews have become an essential factor when it comes to describing the success or failure of a movie. Nowaday's movie reviews influence people's choice whether to watch a movie or not. **Methods/findings:** We have studied the various algorithms of machine learning for classifying the movie review comments by semantic means. People have become very specific with their choices, and they would prefer not to squander their time and cash on a motion picture having terrible reviews. **Improvements/applications:** This study provides a comparison between various machine learning algorithms which can be used for classifying the online movie reviews by semantic meaning. The principle target of the study is to find the best algorithm based on its efficiency and accuracy.

Keywords: Movie Reviews, Sentiment Analysis, NLP, Machine Learning

1. Introduction

Semantic methodologies are portrayed by the utilisation of word, thereby introducing the polarity using sentiment analysis. The pre-processing of text is ordinarily being done by the frameworks and is responsible for gapping the same in the form of words. This is done by the evacuation mechanism in an appropriate manner which is formed from the stop words and also using standardisation based on etymology. This is thereby stemmed, and after doing all these checks the presence or absence of each term of the vocabulary, utilising the entirety of the polarity estimations of the terms for assigning the worldwide polarity estimation of the text. In a typical manner, the frameworks are likewise incorporate the various things including i) A much propelled based treatments of modified terms including as little, too and very which is liable to increase or decrease the polarity of the various accompanied terms ii) Negations or terms based on inversions including never, such as no which is helpful in in inversing the polarity of the terms to which they influence.[1,2]

In addition, the learning-based methodologies comprise with respect to training a classifier utilising any supervised learning algorithm from a gathering of clarified, where each text is usually represented by a vector of words (bag of words), n-grams or skip-grams, in blend with different sorts of semantic features that endeavor to demonstrate the syntactic structure of sentences, intensification, invalidation, subjectivity or incongruity. Systems utilise distinctive techniques, yet the most prominent classifiers based on SVM (Support Vector Machines), Naive Bayes and KNN (K-Nearest Neighbor). Further developed methods appear in the latest investigations, for example LSA (Latent Semantic Analysis) and Deep Learning.[3,4]

Since the rise of the social media platforms there is a enormous amount of textual data that is being generated and which keeps on mounting each day. The text data is the predominant arrangement on the web as it is easy to produce and circulate. This textual data can be separated into two fields: facts and opinions. The fact focuses on the objective data whereas the opinion tells us about the sentiment of their author. Earlier the research work

*Author for correspondence

mainly focused on the classification of the factual part of the data. Now we have web search engines which use keywords for searching. These keywords depict the subject of the content entered. Let's take an example of the word "Avengers", if we google search the word "avengers" it will find more than 2.8 million pages. These pages contain both objective facts and subjective opinions. The factual data can include information about the movie like its cast, crew, budget etc. The opinion part of the data can include movie reviews from the users.

In the past few years, the rise of social networking sites and creation of new websites has helped the users to express their personal opinions about certain things[5] or topics. The opinions expressed can be in the form of blogs, articles, posts and comments. Rotten Tomatoes is one example of movie reviewing sites where users can leave personal views in form of comments or long paragraphs along with a rating which can be related to the opinion expressed by the user.[6]

The main objective of sentiment analysis is to discover a user's attitude on a specific topic based on a textual content which can be in the form of a status update on Facebook, A tweet on Twitter or a comment on any other website. Sentiment analysis can also be denoted as "Opinion Mining". Opinion Mining helps in finding statistical and/or linguistic patterns in the text that reveals the attitude of the author. Opinion Mining is achieved using natural language processing, text analysis and machine learning techniques. It has become very popular in the past few years because of its prompt appropriateness in business environment, such as analyzing the reviews and outlining criticism from the surveys, finding collaborative recommendations etc.[7,8]

2. Existing Techniques for Sentiment Analysis

There are various machine learning algorithms used in the process to analyze the sentiment of movie reviews. Some of them that we used are:

2.1. Decision Trees

One of the most popular machine learning algorithms is the Decision tree which is used in classifications problems. It is best suited for both categorical and continuous target input and output variables. This technique involves the division of population or sample into 2 or more homogeneous sets.

It is based on structure to that of flowchart which contains the internal nodes, branches and the leaf nodes. In this the test on an attribute is done by every internal node and thereby the test results are represented by every branch and also a class label is represented by every leaf node.

The rules based on classification are represented by the path from root to leaf[9,10] Figure 1.

The various parts of the Decision trees are as follows:

Root Node: This node helps in representation of the entire population. This is then divided into 2 or more identical sets.

Splitting: This part is responsible for the division of a node into various sub-modes.

Decision Node: The decision node is the sub-node which divides itself into more sub-nodes.

Leaf Node: Leaf nodes are the nodes which do not split.

Branch: It is a sub section of entire tree.

Decision trees algorithms are easy to understand and easy to explain to others. Those people who lack technical knowledge will be able to draw out the hypothesis drawn from a decision tree. It can handle both categorical and numerical variables. There are no assumptions made by these algorithms on the classifier structure and space distribution. An important aspect in predictive analysis is feature selection which is implicitly performed by Decision trees.[11]

2.2. Random Forest

Random forest is supervised machine learning algorithm. In simple terms random forest builds n decision trees and combines them together to get an accurate as well as stability-based prediction. This algorithm is based on the

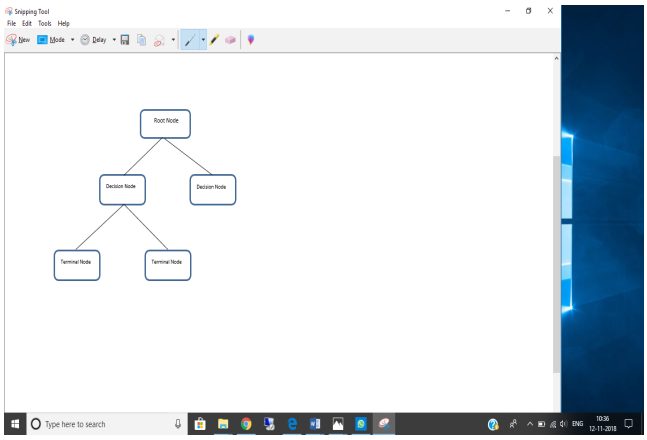

Figure 1. Decision tree model.

bootstrapping mechanism having a CART model. E.g. On having the 1000 observations in total population having 10 variables so, this algorithm build up the multiple CART models having lot of sample and a lot of initial variables[12,13] Figure 2.

Features of Random Forest:

It runs proficiently on vast databases.

It can work with enormous amount of input variables.

It gives estimates about important variables in the classification.

Prototypes are generated that derives the relation between the variables and the classification.

Dealing with missing values in random forest:

Random forest has two ways of replacing missing values. The principal way is quick. If the myth variable is not categorical, the strategy computes the median of all values of any variable in a class. The second way of replacing missing values is computationally more complete yet has given preferable yield over the first, even with a lot of missing data. It replaces missing values only in the training set.[14,15]

2.3. Naïve Bayes

It is a classification technique that assumes of independent predictors. Training data is used for calculating class and conditional probabilities and then new observations can be classified by using the values of these probabilities.

It assumes that a particular feature which is there in a class is not similar to any feature that can exist. E.g. a fruit may be called as an apple if it has features like red color, round and having 2 inches as a diameter. Although, these features are related and dependent, they individually contribute to the fact that this fruit is an apple and hence called Naïve. This sort of classifier calculates the probability for each and every factor. It also imagines that the features are not dependent (In our case words is an input). It is quite easier to build the same and the same is also suitable while we deal with huge datasets.[16,17]

$$P(A|B) = P(B|A)\ P(A)/P(B)$$

The above eqn denotes that how often A happening gives that B will also happen thereby written as P(A l B) where it is known that how often B's happening will give us denotion for A's happening and written as P(B l A) and how likely A and B are on their own.[18,19]

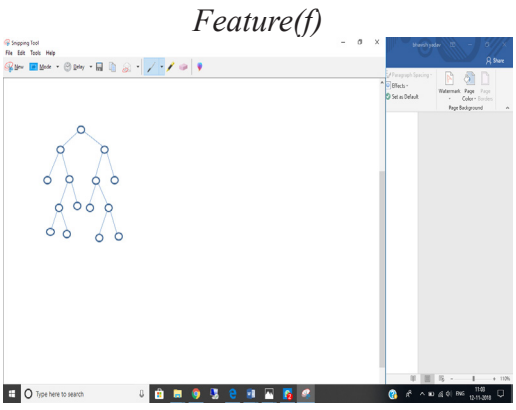

Figure 2. Random forest model.

3. Proposed Work

Dataset used- Only some of the unambiguous ratings are recognized by us. By the help of a set of ad-hoc rules these are extracted. In essence, the first rating that is identified determines classification of a file. Accurate rating decisions can be achieved if the maximum rating for both numerical and star ratings is specified explicitly. ("8/10", "four out of five", and "OUT OF ****: ***" are examples of rating indications we recognise.)

- Considering a 5-star system: Ratings which are 3.5 stars and above are considered positive and ratings that are 2 stars and below are considered negative.
- Considering a 4-star system: Ratings which are 3 stars and above are considered positive and ratings that are 1.5 stars and below are considered negative.
- Considering a letter grade system: If a grade is B or above it then it is taken as positive. If a grade is C- or below then it taken as negative.

Steps of the Algorithm:

- STEP 1 - Import the dataset
- STEP 2 - After removing stopwords we get:
- STEP 3 - To tabulate the no. of words in the corpus, we used nltk. FreqDist object, to know the top N most frequent words in the corpus
- STEP 4 - To extract first 100 most common words
- STEP 5 - To split documents into train (90%) and test (10%) sets
- STEP 6 - To feed the data into the classifier

Classifier gives the output in the form of positive or negative values.

Analysis- We see the "mess" is 4 times more prominent in negative reviews than positive ones and "touches" is 2.8 times more salient in positive reviews Figures 3 and 4.

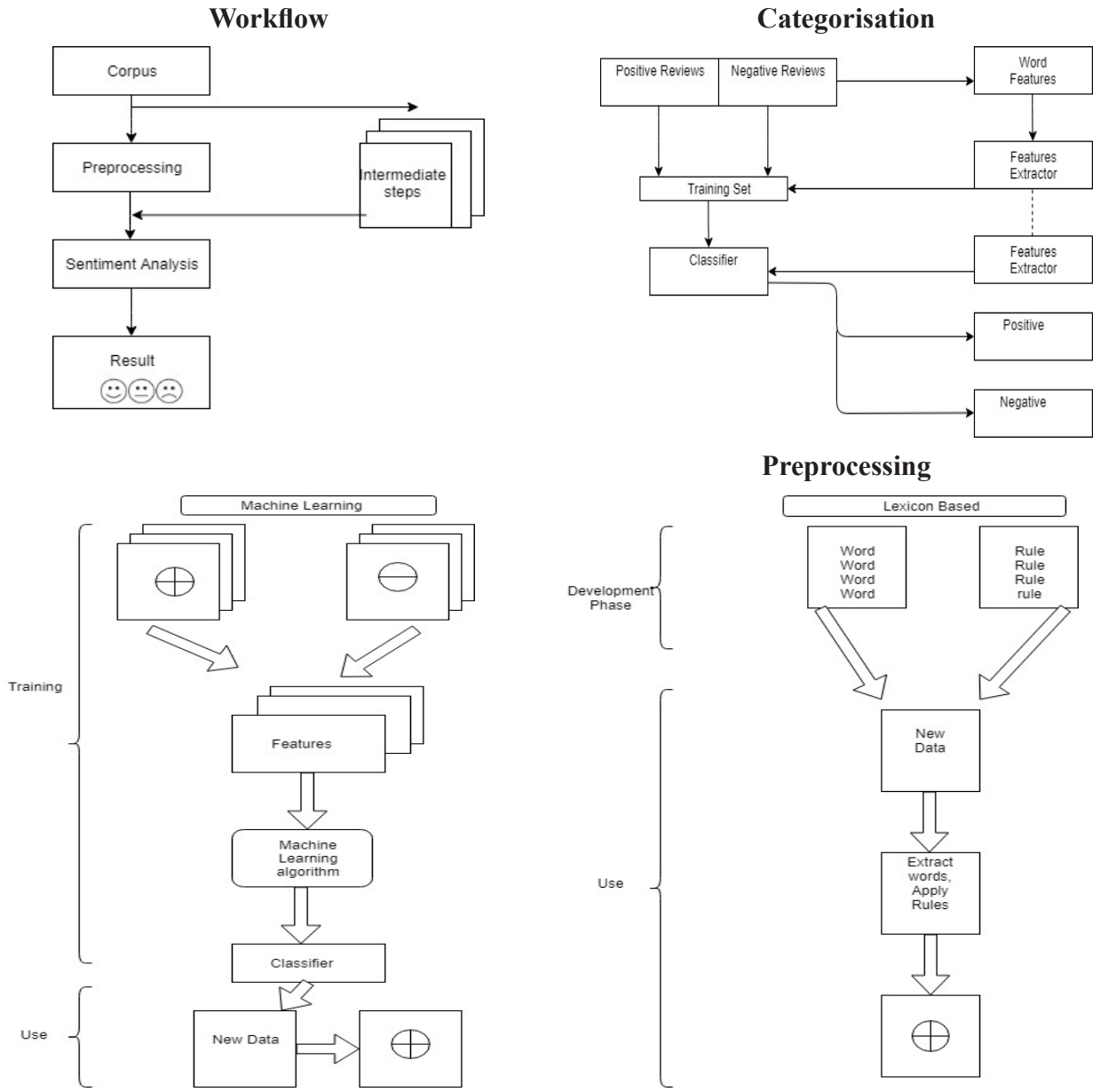

Figure 3. Workflow, categorisation, pre-processing speps.

4. Conclusion

We analyzed the data containing 1000 positive reviews and 1000 negative reviews on the algorithms mentioned above. We found that Naïve Bayes algorithm outperforms and provides best result on the test set. We are working towards increasing the accuracy of movie recommendation and also in real time.

Acknowledgement

We would like to sincerely bring our kind gratitude to Dr. Prateek Jain, Accendere Knowledge Management Services for helping and guiding us in this paper formation.

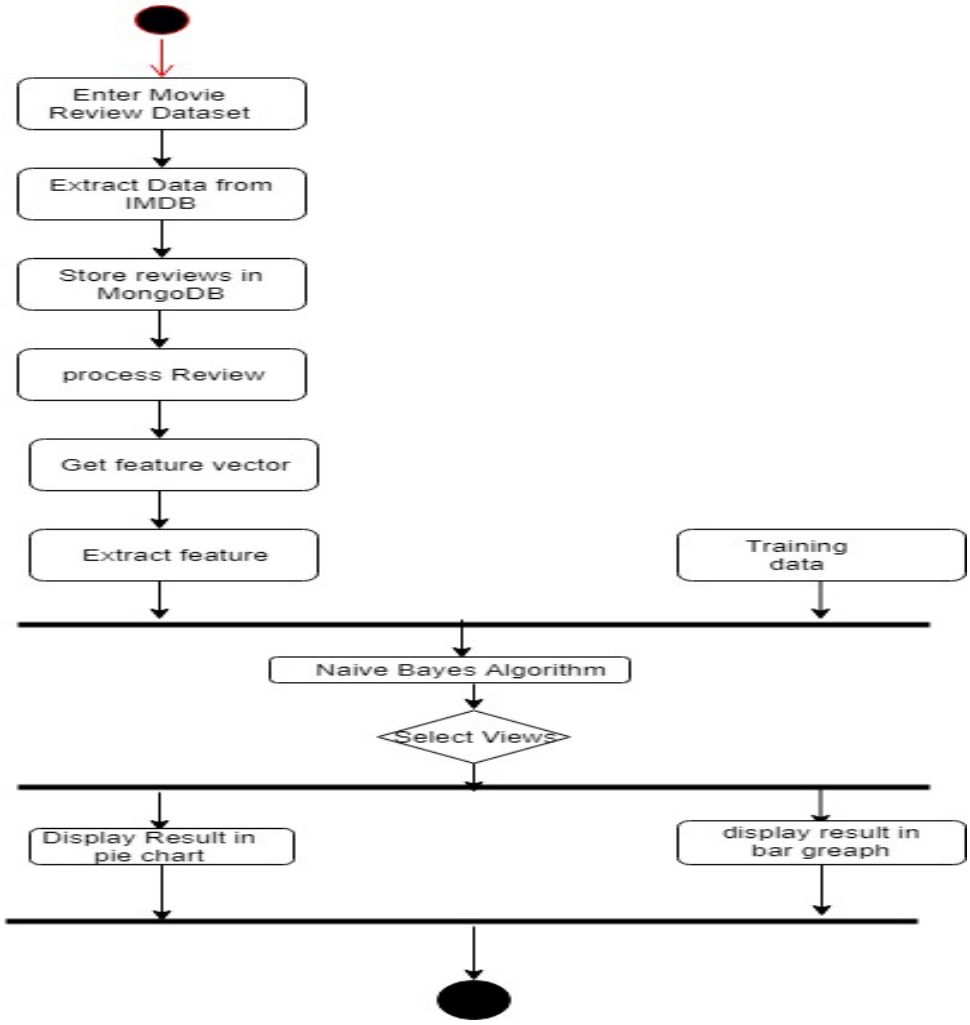

Figure 4. Activity diagram of proposed system.

References

1. Narendra B, Uday Sai K, Rajesh G, Hemanth K, Chaitanya Teja MV, Deva Kumar K. Sentiment analysis on movie reviews: a comparative study of machine learning algorithms and open source technologies. Intell Syst Appl. 2016;8:66–70
2. Singh V, Saxena P, Singh S, Rajendran S. Opinion mining and analysis of movie reviews. Indian J Sci Technol. 2017;10(19):1–6.
3. Taheri S, Mammadov M. Learning the naive bayes classifier with optimization models. Int J Appl Math Comp Sci. 2013;23(4):787–95.
4. An empirical study of the naïve Bayes classifier. [cited 2001]. http://citeseerx.ist.psu.edu/viewdoc/summary?doi=10.1.1.330.2788.
5. Zainab Mirza, Mehwash Khan, Saima Khan, Khurshid Khatri. Movie rating system based on opinion mining. Int J Interdiscip Res Innov. 2015;3(2):34–40.
6. Yessenov K. Sentiment analysis of movie review comments. 6.863 Spring 2009 final project; 2009. P. 1–17.
7. Manikandan R, Sivakumar R. Machine learning algorithms for text-documents classification: a review. Int J Acad Res Dev. 2010;1(1):1–10.
8. Jain A, Kulkarni G, Vraj Shah. Natural language processing. Int J Comp Sci Eng. 2018;6(1):1–7.
9. Tutorial: predicting movie review sentiment with naive bayes. [cited 2019 Jun 13]. https://www.dataquest.io/blog/naive-bayes-tutorial/.
10. How to prepare movie review data for sentiment analysis (text classification). [cited 2017 Oct 16]. https://machinelearningmastery.com/prepare-movie-review-data-sentiment-analysis/.
11. Brbić M, Podnar arko I. Tuning machine learning algorithms for content-based movie recommendation. Intell Decis Technol. 2015;9(3):233–42.

12. Nguyen TH. Machine learning algorithms application to road defects classification. Intell Decis Technol. 2018;12(1):59–66.
13. Devi, Sathiya S, Parthasarathy G. A hybrid approach for movie recommendation system using feature engineering. In: Second international conference on inventive communication and computational technologies; 2018.
14. Inan E. Moreopt: a goal programming based movie recommender system. J Comput Sci. 2018;28:43–50.
15. Kaur L, Kumari N. A research on user recommendation system based upon semantic analysis. Int J Adv Res Comp Sci Softw Eng. 2017;7(11):72–8.
16. Ghazanfar MA, Prügel-Bennett A. Building switching hybrid recommender system using machine learning classifiers and collaborative filtering; 2010.
17. Liu J. A personalized information filtering method based on simple bayesian classifier; 2012. P. 609–14.
18. Özbal G. A content-boosted collaborative filtering approach for movie recommendation based on local and global similarity and missing data prediction. Comp J. 2011;54(9):1535–46.
19. Arsan T. Comparison of collaborative filtering algorithms with various similarity measures for movie recommendation. Int J Comp Sci Eng Appl. 2016;6(3):1–20.