International Journal of Modern Developments in Engineering and Science
Volume 2, Issue 2, February 2023
https://www.ijmdes.com | ISSN (Online): 2583-3138

Various Classification Based Sentiment Analysis on Movies – A Review

H. L. Roopa[*]

Assistant Professor, Department of Information Science and Engineering, Brindavan College of Engineering, Bangalore, India

***Abstract*: Opinion mining, another name for sentiment analysis, categorizes text perspectives and analyses sentiments. It has to do with how text, linguistic, and natural language processing are used. With the quick development of digital technology, enormous amounts of data are produced. Social networking platforms are now widely used and commonplace venues where people may express their feelings through brief messages. These feelings range from joy to sorrow to fear to anxiety. Short text analysis frequently captures the sentiment of the audience. IMDb movie reviews' sentiment analysis explains a reviewer's overall sentiment or opinion of a film. Since consumer perceptions affect a product's performance and a movie's success or failure depends on its reviews, prices are rising and a good sentiment analysis model that categorizes consumer opinions is needed.**

***Keywords*: Sentimental analysis, opinion mining, text analytics.**

1. Introduction

The digital era has changed how people express themselves today. Blog postings, forums, blogs for product testing, social media, and other platforms help to achieve this more. Today, social networking sites like Facebook, Twitter, Google +, and others are used by millions of people to express their emotions. Online communities that inform and sway others through forums provide us with immersive media. The vast majority of sentimentally rich content is produced via social media in the form of tweets, updates, blog posts, notes, feedback, etc. Social media also gives businesses the chance to interact with the people who buy their ads. The majority of individuals primarily rely on user-generated web content. For instance, if someone wants to buy or utilise a product, they first research internet reviews and then discuss them on social media. A typical user cannot examine the amount of user data available. Several techniques are frequently used to process feelings; hence this must be automated [1]. Social media platforms have recently been utilised to discuss current events, business, education, celebrities, and other topics.

The sentimental analysis, often known as "opinion mining," includes the use of instruments for systematic evaluation, review, and study as well as natural language processing (NTP), text mining, electronic linguistics, and bio-metallization. The computational processing and subjectivity of the document are dealt with by SA, also known as sentiment classification, sentiment mining, examination mining, subjectivity analysis, opinion mining, and evaluation extraction and polarity classification. The goal of the textual analysis is to pinpoint the speaker's judgement on a reliable matter. The position could stand in for an appraisal, evaluation, or judgement, a sentiment, or the anticipated distressing communication. Additionally, it shows that anything being "subjective" in this context does not imply that it is unreal [2]. Text, NLP, and language analysis can be used to analyse sentiments as a way of identifying or sensing them.

ML is a technique used to teach computers how to use data more simply and to perform better. After presenting the dataset, there are times when it is difficult to understand the model or extract information from the data. In these situations, we forecast the outcomes using machine learning approaches. Machine learning is necessary and can be collected from multiple sources with a huge number of datasets. From the military to the medical industry, machine learning is utilised to extract practical information from readily available data sets.

2. Literature Review

S. Robila and M. Butler (2016) The development and implementation of the IMDb dataset file extraction and import to database approach are described in this study. Relational databases were used in the previous study, which is different from most published methodologies or studies. With this approach, anyone can add or modify structures in accordance with their demands because it uses document-driven data structures. The creation of the project involved the use of currently required technologies for software engineers and site developers, enabling other developers to fork variations of the work and utilize that for their particular study.

For five high-performance models, A. Yenter & A. Verma (2017) are experimenting with a wide range of regularization techniques, network structures, and kernel sizes. These models have a prediction accuracy of above 89% for the polarity of the feedback from the IMDb dataset. First off, the model with the highest performance is more precise than the previous models, and secondly, it greatly improves the standard CNN+LSTM model. Further sentimental analytics or text classification datasets may benefit from the simultaneous kernel functionality from a number of divisions of the LSTM-based CNN architecture. The proposed architecture may also be used for visual and vocal machine learning.

[*]Corresponding author: roopahl@brindavancollege.com

Sandesh Tripathi and others (2020) To demonstrate how valuable insights may be obtained from a variety of textual data acquired over the Internet, they ran SA on the Kaggle Bag of Words dataset of IMDb film reviews for the research schemes. These concepts are generated from the use of the NB, LR, RF, and DT conventional machine learning algorithms. Also, the performance of all four algorithms was assessed using six test measures: confusion matrix, accuracy, precision, recall, F1 measure, and AUC.

3. Sentimental Analysis

Sentiment Analysis is the process of [2] classifying whether a block of text is positive, negative, or, neutral. Sentiment analysis is contextual mining of words which indicates the social sentiment of a brand and also helps the business to determine whether the product which they are manufacturing is going to make a demand in the market or not. The goal which Sentiment analysis tries to gain is to analyze people's opinion in a way that it can help the businesses expand. It focuses not only on polarity (positive, negative & neutral) but also on emotions (happy, sad, angry, etc.). It uses various Natural Language Processing algorithms such as Rule-based, Automatic, and Hybrid.

Sentiment analysis can occur at different levels: document level, sentence level or aspect/feature level. In this process, sentiment is extracted from the entire review, and a whole opinion is classified based on the overall sentiment of the opinion holder.

4. IMDb Movie Reviews

Movie Reviews dataset is a binary sentiment analysis dataset consisting of 50,000 reviews from the Internet Movie Database (IMDb) labeled as positive or negative. The dataset contains an even number of positive and negative reviews. Only highly polarizing reviews are considered. A negative review has a score ≤ 4 out of 10, and a positive review has a score ≥ 7 out of 10. No more than 30 reviews [4] are included per movie. The dataset contains additional unlabeled data.

5. Machine Learning

In Supervised Learning, the machine learns under supervision. It contains a model that is able to predict with the help of a labeled dataset. A labeled dataset is one where you already know the target answer.

In Unsupervised Learning, the machine uses unlabeled data and learns on itself without any supervision. The machine tries to find a pattern in the unlabeled data and gives a response.

6. Various ML Classifiers

A. Linear Regression

To understand the working functionality of Linear Regression, imagine how you would arrange random logs of wood in increasing order of their weight. There is a catch; however – you cannot weigh each log. You have to guess its weight just by looking at the height and girth of the log (visual analysis) and arranging them using a combination of these visible parameters. This is what linear regression in machine learning is like.

In this process, a relationship is established between independent and dependent variables by fitting them to a line. This line is known as the regression [3] line and is represented by a linear equation Y= a*X + b.

In this equation:

- Y – Dependent Variable
- a – Slope
- X – Independent variable
- b – Intercept

The coefficients a & b are derived by minimizing the sum of the squared difference of distance between data points and the regression line.

B. Logistic Regression

Logistic Regression is used to estimate discrete values (usually binary values like 0/1) from a set of independent variables. It helps predict the probability of an event by fitting data to a logit function. It is also called logit regression.

These methods listed below are often used to help improve logistic regression models:

- include interaction terms
- eliminate features
- regularize techniques
- use a non-linear model

C. Decision Tree

Decision Tree algorithm in machine learning is one of the most popular algorithms in use today; this is a supervised learning algorithm that is used for classifying problems. It works well in classifying both categorical and continuous dependent variables. This algorithm [5] divides the population into two or more homogeneous sets based on the most significant attributes/ independent variables.

D. SVM (Support Vector Machine) Algorithm

SVM algorithm is a method of a classification algorithm in which you plot raw data as points in an n-dimensional space (where n is the number of features you have). The value of each feature is then tied to [9] a particular coordinate, making it easy to classify the data. Lines called classifiers can be used to split the data and plot them on a graph.

E. Naive Bayes Algorithm

A Naive Bayes classifier assumes that the presence of a particular feature in a class is unrelated to the presence of any other feature.

Even if these features are related to each other, a Naive Bayes classifier would consider all of these properties independently when calculating the probability [6] of a particular outcome.

A Naive Bayesian model is easy to build and useful for massive datasets. It's simple and is known to outperform even highly sophisticated classification methods.

F. KNN (K- Nearest Neighbors) Algorithm

This algorithm can be applied to both classification and regression problems. Apparently, within the Data Science industry, it's more widely used to solve classification problems. It’s a simple algorithm that stores all available cases and classifies any new cases by taking a majority vote of its k neighbors. The case is then assigned to the class with which it has the most in common. A distance function performs this measurement.

KNN can be easily understood by comparing it to real life. For example, if you want information about a person, it makes sense to talk to his or her friends and colleagues!

Things to consider before selecting K Nearest Neighbours Algorithm:

- KNN is computationally expensive.
- Variables should be normalized, or else higher range variables can bias the algorithm.
- Data still needs to be pre-processed.

G. K-Means

It is an unsupervised learning algorithm [8] that solves clustering problems. Data sets are classified into a particular number of clusters (let's call that number K) in such a way that all the data points within a cluster are homogenous and heterogeneous from the data in other clusters.

How K-means forms clusters:

- The K-means algorithm picks k number of points, called centroids, for each cluster.
- Each data point forms a cluster with the closest centroids, i.e., K clusters.
- It now creates new centroids based [10] on the existing cluster members.
- With these new centroids, the closest distance for each data point is determined. This process is repeated until the centroids do not change.

H. Random Forest Algorithm

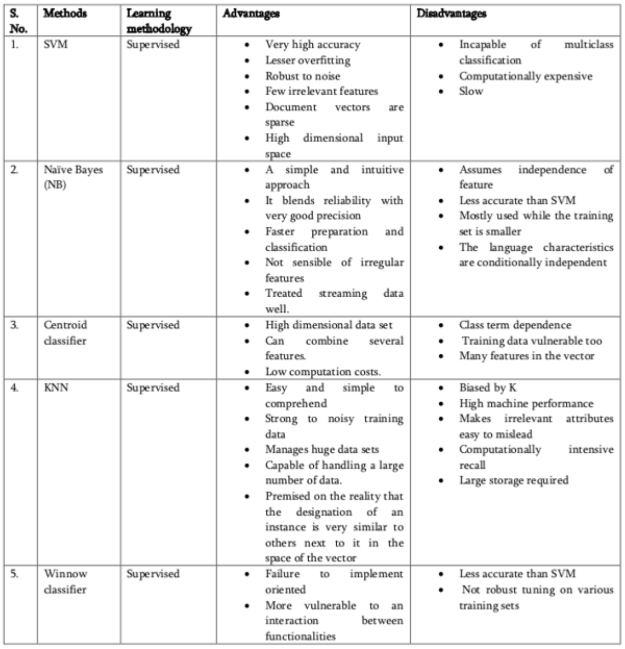
A collective of decision trees is called a Random Forest. To classify a new object based on its attributes, each tree is classified, and the tree “votes” for that class. The forest chooses the classification having the most votes (over all the trees in the forest).

Each tree is planted & grown as follows:

- If the number of cases in the training set is N, then a sample of N cases is taken at random. This sample will be the training set for growing the tree.
- If there are M input variables, a number m<<M is specified such that at each node, m variables are selected at random out of the M, and the best split on this m is used to split the node. The value of m is held constant during this process.
- Each tree is grown to the most substantial [7] extent possible. There is no pruning.

7. Comparison

Table 1
Summary and comparison of numerous methods



S. No.	Methods	Learning methodology	Advantages	Disadvantages
1.	SVM	Supervised	• Very high accuracy • Lesser overfitting • Robust to noise • Few irrelevant features • Document vectors are sparse • High dimensional input space	• Incapable of multiclass classification • Computationally expensive • Slow
2.	Naïve Bayes (NB)	Supervised	• A simple and intuitive approach • It blends reliability with very good precision • Faster preparation and classification • Not sensible of irregular features • Treated streaming data well.	• Assumes independence of feature • Less accurate than SVM • Mostly used while the training set is smaller • The language characteristics are conditionally independent
3.	Centroid classifier	Supervised	• High dimensional data set • Can combine several features. • Low computation costs.	• Class term dependence • Training data vulnerable too • Many features in the vector
4.	KNN	Supervised	• Easy and simple to comprehend • Strong to noisy training data • Manages huge data sets • Capable of handling a large number of data. • Premised on the reality that the designation of an instance is very similar to others next to it in the space of the vector	• Biased by K • High machine performance • Makes irrelevant attributes easy to mislead • Computationally intensive recall • Large storage required
5.	Winnow classifier	Supervised	• Failure to implement oriented • More vulnerable to an interaction between functionalities	• Less accurate than SVM • Not robust tuning on various training sets

8. Conclusion

This paper presented a review on various classification-based sentiment analysis on movies.

References

[1] Vishal A. Kharde, and S. S. Sonawane, “Sentiment Analysis of Twitter Data: A Survey of Techniques”, International Journal of Computer Applications, vol. 139, no.11, April 2016

[2] N. Mahendran, “A Survey: Sentiment Analysis Using Machine Learning Techniques for Social Media Analytics,” International Journal of Pure and Applied Mathematics, vol. 118, no. 8, 419-423, 2018.

[3] Suraj D. M., Rohan A. R. and Vimuktha Evangeleen Salis, “Survey on Sentiment Analysis”, International Journal of Engineering Research & Technology, vol. 8, no. 4, April 2019.

[4] Pooja Kamavisdar, “A Survey on Image Classification Approaches and Techniques”, International Journal of Advanced Research in Computer and Communication Engineering, vol. 2, no. 1, January 2013

[5] Suraj D. M., Rohan A. R et al. Survey on Sentiment Analysis,” International Journal of Engineering Research & Technology, vol. 8, no. 4, April 2019.

[6] Nirag T. Bhatt, “Sentiment Analysis using Machine Learning Technique: A Literature Survey”, International Research Journal of Engineering and Technology, vol. 7, no. 12, Dec. 2020.

[7] Vishal A. Kharde, “Sentiment Analysis of Twitter Data: A Survey of Techniques”, International Journal of Computer Applications, vol. 139, no.11, April 2016.

[8] M. Yasen and S. Tedmori, "Movies Reviews Sentiment Analysis and Classification," 2019 IEEE Jordan International Joint Conference on Electrical Engineering and Information Technology (JEEIT), Amman, Jordan, 2019, pp. 860-865.

[9] H. M. Keerthi Kumar, B. S. Harish, and H. K. Darshan, “Sentiment Analysis on IMDb Movie Reviews Using Hybrid Feature Extraction Method”, International Journal of Interactive Multimedia and Artificial Intelligence, vol. 5, no. 5, Jan 2018.