International Journal of Computer Trends and Technology
ISSN: 2231 – 2803 / https://doi.org/10.14445/22312803/IJCTT-V70I9P104

Volume 70 Issue 9, 33-37, September 2022
© 2022 Seventh Sense Research Group®

Review Article

A Study on Sentiment Analysis of Movie Reviews using ML Algorithms

Md. Sirajul Huque[1], V. Kiran Kumar[2]

[1,2] Department of Computer Science and Technology, Dravidian University, Andhra Pradesh, India.

Received: 03 August 2022 Revised: 04 September 2022 Accepted: 15 September 2022 Published: 30 September 2022

Abstract *- To understand customer preferences, it is now a routine trend in the modern world to gather opinions and recommendations from individuals using a variety of surveys, polls, and social media platforms. Therefore, an accurate and classical mechanism for making assumptions and anticipating sentiments that can fabricate a positive or negative impact in the market is required to understand the sentiments of customers and their view of the services offered by producers. This type of analysis is important for the relationship between producers and consumers. In order to improve customer satisfaction, the key objective of this paper is to study the recommendations that viewers have left for various movies. This study will be used better to comprehend the mindsets and market behavior of the audience. This study uses two algorithms— Logistic Regression and Naive Bayes to analyze consumer perception of various movies and offers concluding observations.*

Keywords *- Recommendations, Sentiments, Naive Bayes, Logistic Regression, Perception.*

1. Introduction

Every human being bases their decisions on previous experiences, feelings, or advice given by other humans. Every time someone wants to purchase a new good or service, they ask for recommendations from others. Similarly, every company wants to provide the best product possible to the market. Thus they use surveys to get feedback from customers. The examination of someone's ideas, sentiments, or emotions in relation to a book or movie is known as sentiment analysis.

Technology has a significant influence on daily living in the modern world. The adoption of new technology has increased significantly due to research and development. Technology has advanced to the point that it now plays a role in our daily lives. Because of the significant improvements made to the web, there has been a significant rise in the amount of emotive content available there. Such a wide range of information may be obtained daily on social media, websites, and public networks through movie or product evaluations or ratings, consumer comments, endorsements, criticisms in discussion forums, etc. One may significantly alter the market by studying market trends, and organizations or producers can tailor their products in accordance with client preferences by using this type of information gathered from the web on a right manner and with the help of appropriate technologies. Sentiment analysis is the name for this form of analysis [1][2].

Sentiment analysis is the conceptual study, identification, extraction, and evaluation of views inside text using natural language processing (NLP) and text analysis. It creates mechanisms that work to determine if a remark has a neutral impact, a positive impact, a negative impact, or no influence at all. Opinion mining is another name for sentiment analysis. There has been a sharp increase in interest in studying and developing different Analysis and Prediction approaches since sentiment analysis has many real-world applications[3][4]. The huge volume of texts expressing opinions that are accessible both publicly and privately from different review sites, forums, blogs, and social media platforms like Facebook, Twitter, LinkedIn, Quora, etc., serve as the input for this process.

1.1. Sentiment Analysis Model

For everything, opinions are frequently stated. A service, a product, a person, a topic, or an organization, for instance. The object being seen consists of several components and some subcomponents. The entity is therefore referred to as an object for sentiment analysis. The hierarchical approach is used in feature-based sentiment analysis because objects, by their very nature, are hierarchical. Sub-components and characteristics of the item are possible.

Consequently, it is challenging for the general public to comprehend these complex phrases (attributes or components). Therefore, the term "Feature" refers to featured-based opinion mining. A paragraph of phrases can communicate a single opinion or attitude. The direction of a viewpoint is determined by the word used to express it. Even one word can express one or many opinions.

2. Related Work

According to Liu [5], sentiment computing is a field of research that examines people's perceptions of things, including goods, services, organizations, people, issues, events, subjects, and their qualities. It also analyses people's feelings regarding such things. Nakov [6] achieved excellent results by combining word-level and character-level models into one model. Word, sentiment and document levels might be used to categorize the text sentiment processing. To do sentiment analysis at the word level, we look for terms that can be mapped to semantic lexicons. Liet [7][15] examined the effectiveness of a statistical machine learning-based classifier versus a lexicon-based classifier.

Additionally, they selected NTUSD as their primary emotion lexicon, which had 7576 positives and 2510 negative terms, respectively.SentiWordNet is utilized in Aurangzeb Khan's rule-based method, which he introduced in 2011 [8], to improve the accuracy of sentiment analysis for user and software evaluations. The suggested method has an accuracy rate of 90% for documents and an accuracy rate of 87% for sentences.

Mudinas and Zhang (2012) [9] presented a hybrid strategy that performs nearly as well as a leaning-based technique while outperforming the lexicon. Hybrid approaches perform as well as machine learning-based techniques while as stable as lexical techniques. The total accuracy of the method is 82.3%. A method for rating and extracting product attributes from opinion articles was proposed by Lei Zhang et al. in 2010 [10]. They had user evaluations at first, but it was challenging for the algorithm to distinguish between positive and bad remarks. They extracted product attributes using the linked rule mining approach. An application called "PoliTwi" was suggested by Seven Rill et al. in 2014 [11] that demonstrate concept-level sentiment analysis's influence on the early discovery of new political subjects on Twitter. Before "Google Trends," hashtags were utilized in this paper's Twitter analysis to predict the outcome of the US election. Data collection and sentiment analysis are done utilising the Twitter API[12][20].

The vast volume of unstructured data generated from numerous social networking sites, such as "Facebook, Twitter, and Instagram," in today's world is challenging to assess, according to Monu Kumar and Dr. Manju Bala's 2016 [7] proposal. As a result, they employed Hadoop and cloud services to analyse and store huge amounts of data intelligently [23].

Twitter sentiment analysis is carried out in the cloud. A method to assess attitudes in the Audio-Video context of a YouTube movie was proposed by Martin Wöllmer et al. in 2013 [12][14]. To obtain user reviews as input, they used the Metacritic database. They applied the data-based method inside and across domains to evaluate the knowledge-based approach. Aspect rating distributions and language modeling were introduced by Giuseppe Di Fabbrizio in 2013 [8] and are used to summarize online product and service evaluations. They employed a cutting-edge method that considers aspect rating distributions and language modeling while extracting multi-document summarization for textual data.

Rafeeque Pandarachalil et al. developed an unsupervised learning strategy in 2014 [9] as a tool for Twitter sentiment analysis. They discovered that three sentiment lexicons—SenticNet, SentiWordNet, and SentislangNet—are used to analyse the polarity of tweets. To put this strategy into practice, they employed the parallel Python framework. App producers can benefit financially from Chirag Sangani's 2013 [10] approach for evaluating user opinions about apps through review comments and ratings. They suggest a system that offers a list of reviews for each topic that show user perspectives on that subject and a many-to-many relationship from reviews to subjects of interest.

3. Proposed Work

3.1. Data Collection

Movietalks.csv files contain reviews of movies. The goal is to foretell whether a particular review will be favorable or unfavorable. To do this, an algorithm is trained using the reviews and their classifications in train data.csv, and it then uses the reviews in test data.csv as input for making predictions. Then errors are estimated using the test data.csv actual classifications to determine our predictions' accuracy.

3.2. Data Set Preparation

The pre-processing of the document includes the dataset's generation before any algorithm is applied. It is done to eliminate any words or symbols you don't want. These words or symbols don't change the outcome, but they can make the algorithm take longer to process. Pre-processing for our dataset entails the following actions:

3.3. Stopping

Stopping is a method for removing the majority of repetitive words while using a stop-word list as a guide to lowering the size of the document. Stop words often include a, an, the, this, too, for, etc.

Porter stemming removes common or regular morphological ends from English words. It roots the words by stemming them. For instance, the words hot, hotter, and hottest are derived from the word hot.

Additionally, words that appear more than 80% of the time in the sample are disregarded since they are probably stop-words. Similarly, words with extremely low frequencies have to be discarded as well.

3.4. Splitting Tests by Train

Training data (X train, y train) and test data (X test, y test) make up both halves of the whole dataset. After learning from training data, different classifiers will be evaluated for efficiency using test data. The original dataset is broken at an 80:20 ratio.

The following lists are produced by this step:

- X train: review/features of the training
- X test: test review/features;
- y train: training sentiments/output (1 for positive, 0 for negative)
- y test: test output/emotions (1 for positive, 0 for negative)

3.5. Extraction of Features and Opinion Words

The sentence's whole collection of opinion words is chosen. From the movie review, the algorithm extracts all nouns, noun phrases, verbs, and adjectives and then compares them to the list of terms already in use. These words are grouped according to how polar they are. The word "good," for instance, has a positive polarity.

On the other hand, characteristics are chosen based on the frequency with which opinion words appear. Opinion words are added to the features list if they appear more frequently in the review than the threshold value. For this system, API is solely trained for movie reviews using a lexicon of words and phrases such as "excellent acting," "solid narrative," and "great action."

4. Algorithms Used

Using well-known supervised machine learning classifiers like KNN, Logistic Regression, Nave Bayes, Support Vector Machine, and many more, the problem of sentiment analysis of movie reviews is examined.

4.1. Logistic Regression Classifier

A classification algorithm is characterized as using logistic regression. When a set of independent variables is presented, it is typically used to forecast a binary result (such as 0 / 1, False / True, No / Yes, Wrong / Right, etc.). We employ a few dummy variables to represent the binary result or categorical outcome.

Hypothesis: Z=WX+B (1)

hØ(x)=sigmoid(Z)

Sigmoid (t)=1/(1+e-t)

Cost function

Cost (hØ(x),Y (actual))

{ -log(hØ(x)) if y=1

{ -log(1-hØ(x)) if y=0

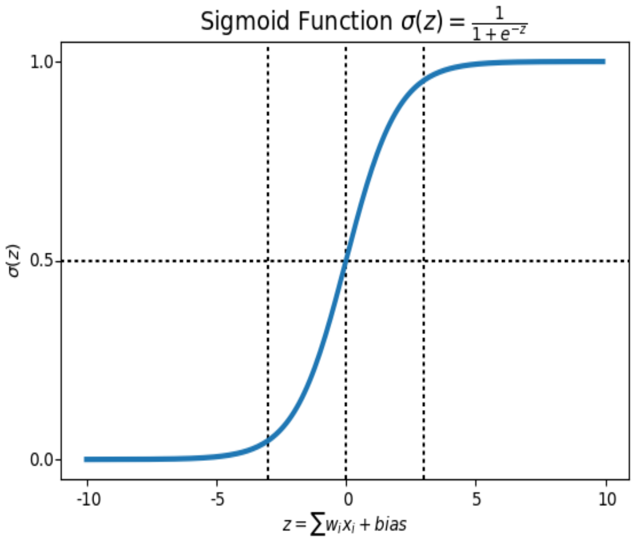

Fig. 1 Sigmoid function in logistic regression

4.2. Naive Bayes Classifier

A classification process that relies on the Bayes theorem and assumes that the characteristics are independent of one another. As a general rule, we may state that "a classifier that uses the Nave Bayes method postulates that the presence of a specified feature in a class is not connected to the occurrence of any other characteristic" to make it as easy to grasp as possible[23].

The Naive Bayes model is quick and straightforward to construct and is particularly effective in classifying text in very sizable data sets. Along with being straightforward, it has a reputation for outperforming sophisticated and very complicated categorization algorithms.

P (A|B) = P (B|A)*P (A)/P (B) (2)

Where

P (A|B) = Posterior Probability

P (A) = Class Prior Priority

P (B|C) =Likelihood

P (B) = Predictor Prior Probability

P (A|B) = P (B1/C) X (B2/C) X...P (B_n/C) X P(C)

5. Results

This study employs three machine learning classifiers—Naive Bayes, Logistic Regression, and k-Nearest Neighbors-and the analysis is at the phrase level. Data from Bigrams and Unigrams were used to evaluate the procedures of partitioning the data set into two halves. Table 1 displays the accuracy and confusion matrices for several classifiers, respectively.

Table 1. Evaluation Measure

ML Classifier	Precision	Recall	Accuracy
Naive Bayes	0.2278	0.0191	47%
Logistic Regression	0.4542	0.7739	41.42%

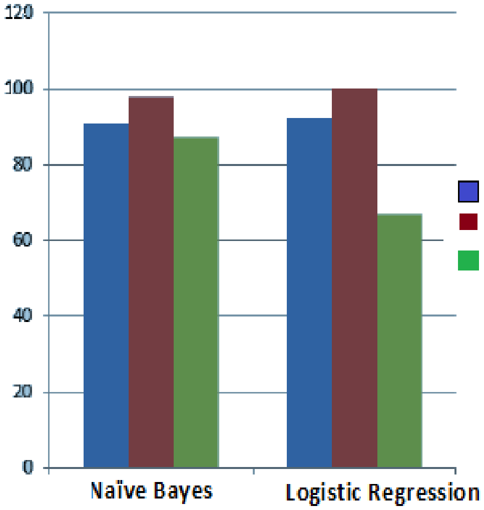

Fig. 2 Performance of ML algorithms

6. Conclusion

Most of the time, information on consumer reviews of a certain product is accessible across various disparate information domains. Users may choose whether a movie is worth their time by reading movie reviews. Users can save time by not reading all of the reviews for a movie by using a summary of all the reviews to make this decision. Critics frequently publish reviews and ratings on websites that assess movies, which aids audiences in deciding whether or not to see the movie. Based on their reviews, sentiment analysis can determine the attitude of the judges. A movie review's sentiment may be analyzed to determine if it is good or negative, affecting the movie's total score. Therefore, since the machine learns by training and evaluating the data, it is possible to automate the process of determining whether a review is favorable or negative.

The data set used in this study was obtained from various sources, and significant tokens were then extracted using the count vectorization and train test split procedures. In order to train the data set, analyze the reviews, and accurately forecast the reviews' sentiment—whether positive or negative, machine Learning methods (Logistic Regression and Naive Bayes) are used. An effective and efficient mechanism has been automated to gather the movie reviews, enabling the anatomization to be completed quickly so that the evaluation's usefulness may be employed effectively up to that point. Nave Bayes may be applied extremely effectively and correctly to sentiment analysis of movie reviews, brand reviews, and many other reviews to understand consumer emotions and behavioral preferences.

References

[1] Shengyi Jiang, Limin Kuang, Meiling Wu, and Guansong Pang, Guangdong University of Foreign Studies, School of Informatics, Guangzhou, China, vol. 39, pp. 1503-1509, 2012.

[2] N. Aston, J. Liddle, and W. Hu, "Twitter Feelings in Data Stream with [J] Perceptron," *Journal of Computer and Communications*, pp. 11–16, 2014.

[3] "Sentiment Analysis of Hollywood Films on Twitter, by Umesh Rao Hodeghatta," *IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining*, pp. 25–29, 2013.

[4] Bhumika Gupta, Monika Negi, Kanika Vishwakarma, Goldi Rawat, and Priyanka Badhani, "Study of Twitter Sentiment Analysis using Machine Learning Algorithms on Python," *International Journal of Computer Applications*, vol. 165, 2017.

[5] B. Liu, "Synthesis Lectures on Human Language Technologies," Sentiment Analysis and Operation Mining, pp. 152-153, 2016.

[6] P. Nakov Tiedemann, "Combining Word-Level and Character-Level Models for Machine Translation Between Closely Related Languages," *Association for Computational Linguistics Meeting: Short Papers*, pp. 301-305, 2012.

[7] B. Wen, T. T. He, L. Luo, L. Song, and Q. Wang, "Text Sentiment Classification Research Based on Semantic Comprehension," *Computer Science*, pp. 261-264. 2010.

[8] "Extraction and Ranking of Product Features," Lei Zhang University of Illinois, Chicago, Coling 2010 Poster Volume, Beijing, pp. 1462-1470, 2010

[9] A. Khan, B. Baharudin, and K. Khan, "Sentiment Classification from Online Customer Reviews Using Lexical Contextual Sentence Structure," *2nd International Conference on Software Engineering and Computer Systems ICSECS*, Springer, pp. 317–331, 2011.

[10] M. Annett and G. Kondrak, "A Comparison of Sentiment Analysis Techniques: Polarizing Movie Blogs," *Canadian Conference on AI*, pp. 25–35, 2008.

[11] "Combining Lexical and Learning-Based Techniques for Concept-Level Sentiment Analysis," ACM, New York, NY, USA, *Proceedings of the First International Workshop on Issues of Sentiment Discovery and Opinion Mining,* vol. 5, pp. 1-8, 2012.

[12] Yue Lu, C. Zhai, and H. Wang, "A Rating Regression Method for Latent Aspect Rating Analysis using Review Text Data," in *Knowledge Discovery and Data Mining: Proceedings of the 16th ACM SIGKDD International Conference, ACM*, pp. 783–792, 2010.

[13] Fouziah Hamza, S. Maria Celestin Vigila, "A Trust Management Scheme for Intrusion Detection System in MANET using Weighted Naïve Bayes Classifier," *International Journal of Engineering Trends and Technology,* vol. 70, no. 2, pp. 75-85, 2022 *Crossref,* https://doi.org/10.14445/22315381/IJETT-V70I2P211.

[14] A. Kumar, R. Khorwal, and S. Chaudhary, "A Survey on Sentiment Analysis Utilising Swarm Intelligence," *Indian Journal of Science and Technology,* vol. 9, no. 39, 2016.

[15] X. He et al., "Intelligence Science and Large Data Engineering: Image and Video Data Engineering," *5th International Conference*, IScIDE 2015 Suzhou, China, 2015 Revised Chosen Papers, Part I, Lecture Notes in Computer Science, vol. 9242, no. 1, 2015.

[16] Paramita Ray, "Document Level Sentiment Analysis for Product Review using Dictionary Based Approach," *SSRG International Journal of Computer Science and Engineering,* vol. 4, no. 6, pp. 24-29, 2017. Crossref, https://doi.org/10.14445/23488387/IJCSE-V4I6P105

[17] Prafulla Mohapatra, Rohit Kumar Singh, Shashank Pandey, Prashanth Anand Kumar, Mrs.Asha K N, "Sentiment Classification of Movie Review and Twitter Data Using Machine Learning," *SSRG International Journal of Computer and Organization Trends,* vol. 9, no. 3, pp. 1-8, 2019. Crossref, https://doi.org/10.14445/22492593/IJCOT-V9I3P301

[18] N. Muslimah and R. C. Wihandika, "Film Classification Based on Synopsis Using Improved K-Nearest Neighbor (K-NN)," *Journal of Information Technology and Computer Science Development*, J-PTIIK Universitas Brawijaya, vol. 3, no. 1, pp. 196–204, 2019.

[19] G. Portolese and V. D. Feltrin, "On the Use of Synopsis-based Features for Film Genre Classification," pp. 892-902, 2019.

[20] J. Wehrmann, M. A. Lopes, and R. C. Barros, "Self-Attention for Synopsis-Based Multi-Label Movie Genre Categorization," *Proceedings of the 31st International Florida Artificial Intelligence Research Society Conference,* FLAIRS, pp. 236-241, 2018.

[21] D. Bui and J. Doba, "Lyrics Classification Using Naive Bayes," *International Convention on Information and Communication Technology, Electronics and Microelectronics (MIPRO),* pp. 1011–1015, 2018.

[22] A. D. Hartanto, "Job Seeker Profile Classification of Twitter Data using the Nave Bayes Classifier Algorithm Based on the DISC Method," pp. 533-536, 2019.

[23] G. Portolese and V. D. Feltrin, "On the Use of Synopsis-based Features for Film Genre Classification," pp. 892– 902, 2019.

[24] J. Wehrmann, M. A. Lopes, and R. C. Barros, "Self-Attention for Synopsis-Based Multi-Label Movie Genre Classification," *Proceedings of the 31st International Florida Artificial Intelligence Research Society Conference, FLAIRS*, pp. 236–241, 2018.

[25] N. K. Verma and A. Salour, "Feature Selection," *Intelligent Condition Based Monitoring: For Turbines, Compressors, and other Rotating Machines, Springer*, Singapore, pp. 175-200, 2020.