IJIRCCE

e-ISSN: 2320-9801 | p-ISSN: 2320-9798

INTERNATIONAL JOURNAL OF INNOVATIVE RESEARCH

IN COMPUTER & COMMUNICATION ENGINEERING

Volume 11, Issue 5, May 2023

INTERNATIONAL STANDARD SERIAL NUMBER INDIA

Impact Factor: 8.379

International Journal of Innovative Research in Computer and Communication Engineering

| e-ISSN: 2320-9801, p-ISSN: 2320-9798| www.ijircce.com | |Impact Factor: 8.379 |

|| Volume 11, Issue 5, May 2023 ||

| DOI: 10.15680/IJIRCCE.2023.1105183 |

Sentimental Analysis of Movie Review System Using Machine Learning Algorithms

Rajvi Nilesh Bhavsar, Ayush Sha, Anjali Kumari, Harsh Dodiya, Aditya Deori,

Dr. M K Jayanthi Kannan

UG Student, Dept. of Computer Science, Jain University, Bengaluru, India

UG Student, Dept. of Computer Science, Jain University, Bengaluru, India

UG Student, Dept. of Computer Science, Jain University, Bengaluru, India

UG Student, Dept. of Computer Science, Jain University, Bengaluru, India

UG Student, Dept. of Computer Science, Jain University, Bengaluru, India

Professor, Dept. of Computer Science, Jain University, Bengaluru, India

ABSTRACT: Entertainment plays a vital role in songs, drama, and movies. Although determine if a movie is good or bad as human life and includes various forms such as music, people typically prefer theaters to watch good movies, it is not sufficient to rely on one or two reviews to different individuals have different opinions. Therefore, a sentiment analysis technique using natural language processing (NLP) and machine learning classification algorithms is proposed for languages like GUJARATI and ASSAMESE. The sentiment expression is analyzed to classify the polarity of the movie review on a scale of 0 (highly disliked) to 3 (highly liked), and feature extraction and ranking are performed to train a multi-label classify the movie review into its classifier to correct label. Due to the lack of strong grammatical structures in movie reviews, an approach based on structured N-grams is adopted. Additionally, a comparative study on different classification approaches is conducted to determine the most suitable classifier for the problem domain. Overall, it is concluded that the proposed approach to sentimentclassification cancomplement existing movie ratin systems usedon the web and will serve as afoundation forfuture research in this area.

KEYWORDS: Decision tree, random forest, support vector machine

I.INTRODUCTION

Assessing a movie's performance through reviews is an important practice. While assigning a rating can provide a quantitative measure of success, a compilation of textual reviews can offer a deeper qualitative understanding of a movie's strengths and weaknesses. By analyzing reviews, one can gauge whether the film meets the reviewer's expectations. Sentiment Analysis is a field of machine learning that aims to extract subjective information from textual reviews, relying on natural language processing and text mining. It can determine the reviewer's attitude towards various aspects of the film and their overall sentiment. Additionally, it can identify subjectivity/objectivity, which involves classifying text into one of two categories: objective or subjective. This classification can be challenging since an objective document may contain subjective sentences, and the subjectivity of words and phrases depends on their context. Existing approaches to sentiment analysis fall into four categories: keyword spotting, lexical affinity, statistical methods, and concept-level techniques. Keyword spotting identifies text by affect categories based on the presence of affect words likehappy, sad, and afraid. Lexical affinity enhances keyword- based approaches by considering arbitrary words' probable affinity to certain emotions. Statistical methods incorporate machine learning elements, such as latent semantic analysis, support vector machines, bag of words, and Semantic Orientation.
In this project we aim to use Sentiment Analysis on a set of movie reviews given by reviewers and try to understand what their overall reaction to the movie was, i.e. if they liked the movie or they hated it. We aim to utilize the relationships of the words in the review to predict the overall polarity of the review.

II. RELATED WORK

Sahu, T. P., & Ahuja, S. (2016) In this work, they have extracted new features that have a strong impact on determining the polarity of the movie reviews and applied computation linguistic methods for the preprocessing of the data. We then performed the feature impact analysis by computing information gain for each feature in the feature set and used it to

| e-ISSN: 2320-9801, p-ISSN: 2320-9798| www.ijircce.com | |Impact Factor: 8.379 |

|| Volume 11, Issue 5, May 2023 ||

| DOI: 10.15680/IJIRCCE.2023.1105183 |

derive a reduced feature set. Among six classification techniques, they have found that the highest accuracy was given by Random Forest with an accuracy of 88.95%. In future, we would like to evaluate the effectiveness of the proposed sentiment classification features and techniques for other tasks, such as sentiment classification. They would like to apply in-depth concepts of NLP for better prediction of the polarity of the document. We would also like to extend this technique on other domains of opinion mining likes newspaper articles, product reviews, political discussion forums etc.

Yasen, M., &Tedmori, S. (2019) The research goal of this work is to address SA by constructing an approach that can classify movie reviews and then compare the results in an inclusive study of eight well known classifiers. To evaluate the proposed model, IMDB reviews real dataset was utilized. Tokenization was applied on the dataset to transfer strings into word vector, then stemming was used to extract the root of the words, afterwards gain ratio was applied on the dataset as an attribute selection algorithm. Then, the data was split into training and testing datasets using the percentages 66%, 34% respectively. To evaluate the results accuracy, precision, f- measure, recall, and AUC were used. The results showed that RF has proved its efficiency over 7 other classifiers where it got the best result in all of the evaluation metrics taken into consideration, KNN also was able to get a recall similar to RF and a very competitive f-measure and AUC. Furthermore, DT got a very competitive recall value. Finally, RRL got the worst result. The authors wish to conduct a similar study on different languages specifically on Arabic. In addition, the authors wish to experiment with different SA methods in order to increase the accuracy of the results. Nama, V., Hegde, V., & Satish Babu, B. (2021) The proposed model uses the Natural Language Toolkit (NLTK) library to conduct sentiment

analysis on the given data. NLTK is a leading platform used in building Python programs thatwork with human language and data. In the context of sentiment analysis NLTK provides a probability of a text being positive, neutral or negative. In the scope of this work, it was considered to only take into account positive and negative movie reviews only. The sentiment with the highest probability is what is finally determined to be the overall sentiment of the movie review. This library consists of various functions such as “word tokenize”, “Naïve-Bayes Classifier, “stop words” and “wordnet” that are incorporated into the model. In thismodel NLTK Wordnet was used which is imported from the nltk.corpus package. It is an extremely big database of English words consisting of Adjectives, Nouns, Adverbs and Verbs. These are grouped into synsets which are a set of cognitive synonyms. The model also uses the “movie reviews” dataset which is imported from nltk corpus. In language, the word corpus (plural corpora) or text corpus is a big and organized group of words or textual matter. In corpus linguistics, they are used to do test hypothesis and analyze statistical results, calculating occurrences or authenticating language regulations within a specific linguistic territory. This dataset consists of a collection of movie reviews which are comprised of a total of 1583820 words. The movie reviews are divided in 2 categories ‘neg’ and ‘pos’.

Permatasari, R. I., Fauzi, M. A., Adikara, P. P., & Sari, E. D. L. (2018) In this study, movie reviews in Indonesian language was classified into positive and negative class using Naïve Bayes classifier. Ensemble Features was employed in this study by combining several features i.e., Twitter specific features, textual features, part of speech features, and lexicon- based features, and Bag of Words. Based on the experiment results, it can be seen that in general, Naïve Bayes showed good performance in this task. Meanwhile, among these individual features, it was found that the Bag of Word features has the best performance with 0.96 precision, 0.92 recall, and 0.94 f-measure value. This is because the word feature in the bag of words depends on the data. Word features that often appear in training data with negative classes will have high frequency values for negative classes as well, so that the word feature can be a clue in the classification process to show which test data will enter the negative class and vice versa. Besides, some feature types like Twitter specific features and textual features very rarely appear in training data so it makes poor classification performance. However, if the bag of words features is combined with other individual features, the accuracy is lower than the accuracy of the system using only bag of words because some of the excess bag of words is covered by the lack of individual features such as Twitter specific features that rarely appear in all data.

Vrunda C. Joshi,Dr. Vipul M. Vekariya (2017) This work includes classification of tweets into two basic polarities i.e. positive and negative. From literature survey, I found that SVM and POS tagging gives higher accuracy. So I will usePOS tagging as a feature extractor and SVM as a classifier. The accuracy is the parameter that is used to measure the performance of the system. Results of this approach give accuracy 92%. This work can be extended to combination of any other feature extraction and machine learning approach.Lata Gohil, Dharmendra Patel (2019). They proposed method to generate G-SWN using Hindi SentiWordNet (H-SWN) and IndoWordNet (IWN) by exploiting synonym relations. The generated G-SWN resource will be useful for sentiment analysis of Gujarati text. The proposed method can be applied to generate sentiment lexical resources for all languages included in IWN. The Gujarati tweets corpus is developed for evaluation of the generated lexical resource. Corpus was annotated for positive and negative polarity

| e-ISSN: 2320-9801, p-ISSN: 2320-9798| www.ijircce.com | |Impact Factor: 8.379 |

|| Volume 11, Issue 5, May 2023 ||

| DOI: 10.15680/IJIRCCE.2023.1105183 |

classes by two annotators. Statistical measure inter-annotator agreement Cohen's kappa achieved for this corpus is 0.55. Resultant annotated corpus comprises of 863 tweets out of which 442 are positive tweets and 421 are negative tweets. Evaluation of G-SWN using this gold standard corpora achieved 52.72% and 52.95% accuracy for unigram presence and simple scoring classifiers respectively. Result shows moderate performance of G-SWN. G-SWN provides baseline for further study.

A. Equations:
Naive Bayes classifier:
The Naive Bayes classifier is a probabilistic classifier that makes predictions based on the probability of a given input belonging to a particular class (e.g. positive or negative). The basic formula for the Naive Bayes classifier is:
P(class | document) =
P(document | class) * P(class) / P(document)
where:
- P(class | document) is the probability of the input document belonging to the given class
- P(document | class) is the probability of observing theinput document given the class
- P(class) is the prior probability of the givenclass
- P(document) is the probability of observing the input document (i.e. a normalization constant)

Support vector machines (SVMs):
SVMs are a type of supervised learning algorithm used for classification tasks. The basic formula for SVMs involves finding a hyperplane that separates the input data into the different classes, while maximizing the margin between the hyperplane and the closest data points. SVM is a popular machine learning algorithm used in sentiment analysis
because of its ability to handle high-dimensional feature spaces and its effectiveness in separating data points into different classes. The basic formula for SVMs is:
f(x) = sign(w.T * x + b) where: - x is the input feature vector
- w is the weight vector
- b is the bias term
- f(x) isthe predicted output class (i.e. +1 or -1)

III. PROPOSED SYSTEM

- **Data Collection**: Gather a diverse dataset of movie reviews from various sources like IMDB, Rotten Tomatoes,and Metacritic, to ensure that the reviews are representative of a wide range of movie genres and ratings.
- **Pre-processing**: Clean the data by removing duplicate reviews, correcting spelling errors, and converting the text to lowercase for consistency
- **Tokenization**: Split the text into individual words or tokens to prepare for further analysis
- **Stop Word Removal**: Eliminate commonly used words like "the," "and," and "is," which do not provide any meaningful sentiment information.
- **Stemming/Lemmatization**: Reduce the words to their root form to ensure that variations of the same word are treated as one
- **Feature Extraction**: Use techniques such as bag-of- words or TF-IDF to extract the most relevant features from the text data.
- **Sentiment Analysis**: Apply machine learning algorithms like Naive Bayes, Support Vector Machines, or Recurrent Neural Networks to classify the reviews into positive, negative, or neutral sentiment categories.
- **Model Evaluation**: Evaluate the performance of the model using various metrics like accuracy, precision, recall, and F1 score. Tweak the model parameters and try different techniques to enhance its performance.
- **Accuracy**: is a commonly used metric for evaluating the performance of machine learning models. It measures the proportion of correctly predictedinstances among all instances in the dataset. The formula for accuracy is:

Accuracy = (Number of Correct Predictions) / (Total Number of Predictions)

Precision: is a metric used to evaluate the performance of a machine learning model, specifically for binary classification problems. It measures the proportion of true positive predictions among all positive predictions made by the model. The formula for precision is:
Precision = (True Positives) / (True Positives + False Positives)

| e-ISSN: 2320-9801, p-ISSN: 2320-9798| www.ijircce.com | |Impact Factor: 8.379 |

|| Volume 11, Issue 5, May 2023 ||

| DOI: 10.15680/IJIRCCE.2023.1105183 |

Recall: is another commonly used metric for evaluating the performance of a machine learning model, specifically for binary classification problems. It measures the proportion of true positive predictions among all actual positive instances in the dataset. The formula for recall is:
Recall = (True Positives) / (True Positives + False Negatives)

F1 score: is a commonly used metric for evaluating the performance of a machine learning model in binary classification problems. It is a harmonic mean of precision and recall, which takes into account both metrics and provides a balanced evaluation of the model's performance. The formula for F1 score is:
F1 Score = 2 * (Precision * Recall) / (Precision + Recall)

- **Deployment**: Integrate the sentiment analysis model into a movie review system to classify new reviews and provide recommendations based on the sentiment analysis results.

Machine learning is a type of artificial intelligence that allows computer algorithms to learn from data and improve their performance over time. It involves training a model on a dataset to identify patterns, relationships, and trends, and then using that model to make predictions or decisions about new data. Machine learning algorithms can be supervised, unsupervised, or semi- supervised, depending on the level of guidance provided during the learning process.

To perform sentiment analysis using machine learning, we first need to gather a dataset of labeled text data, where each text is labeled as positive, negative, or neutral sentiment. We then preprocess the text data by removing

stop words, stemming, and converting the text into a numerical representation, such as bag-of-words or TF- IDF.Next, we split the dataset into training and testing sets and select a machine learning algorithm, such as Naive Bayes, Support Vector Machines (SVM), or Neural Networks, to train a model on the training data. During training, the algorithm learns the relationships between the input text features and the corresponding sentiment labels.

To improve the performance of the sentiment analysis model, we can experiment with different preprocessing techniques, feature representations, machine learning algorithms, and hyperparameters. We can also incorporate domain-specific knowledge, such as sentiment lexicons or topic models, to enhance the model's accuracy and generalizability.
Linear Regression: A statistical model that predicts a continuous output variable based on one or more input variables. It's commonly used in applications such as finance, economics, and healthcare.

Logistic Regression: A statistical model used for binary classification problems, where the output variable can take only two values. It's used in applications such as fraud detection, spam filtering, and credit risk analysis.

Decision Trees: A machine learning model that uses a tree- like structure to classify data based on a sequence of decisions. It's commonly used in applications such as customer segmentation, fraud detection, and medical diagnosis.

Random Forests: An ensemble learning method that combines multiple decision trees to improve accuracy and reduce overfitting. It's used in applications such as image classification, speech recognition, and bioinformatics.

Support Vector Machines (SVM): Support Vector Machines (SVM) is a popular machine learning algorithm used for classification and regression tasks. In SVM, the algorithm maps the input data to a high-dimensional feature space and identifies a hyperplane that maximally separates the data points into different classes. SVM tries to find the best decision boundary (hyperplane) that can separate the data points with the largest margin possible, which is known as the maximum margin hyperplane.

In sentimental analysis of movie reviews, SVM can be used to classify reviews into positive or negative sentiment categories. The input features can be the words used in the review, and SVM tries to identify the best hyperplane that can separate the reviews with different sentiment. SVM can also handle non- linear classification tasks by using kernel functions to transform the input data into a higher-dimensional feature space where a linear hyperplane can separate the data points.

Naive Bayes: Naive Bayes is a machine learning algorithm commonly used for sentimental analysis of movie reviews. It is based on Bayes' theorem and assumes that the features(words) in the movie reviews are independent of each other. Naive Bayes calculates the probability of each class (positive or negative sentiment) given the input features and

| e-ISSN: 2320-9801, p-ISSN: 2320-9798| www.ijircce.com | |Impact Factor: 8.379 |

|| Volume 11, Issue 5, May 2023 ||

| DOI: 10.15680/IJIRCCE.2023.1105183 |

assigns the review to the class with the highest probability. The algorithm first preprocesses the text data by removing stop words and converting the text into numerical features such as bag of words or term frequency-inverse document frequency (TF- IDF). Then, it calculates the prior probability of each class based on the frequency of each class in the training data and the conditional probability of each feature given each class.

K-Nearest Neighbors (KNN): A machine learning algorithm that predicts the class of a data point based on the classes of its k nearest neighbors. It's used in applicationssuch as anomaly detection, customer segmentation, and image recognition.

Clustering: A machine learning technique that groups data points based on their similarity. It's used in applications such as customer segmentation, anomaly detection, and image segmentation.

IV.CONCLUSION

In conclusion, the application of sentimental analysis in movie review systems has been a topic of interest for researchers and industry professionals alike. The use of natural language processing and machine learning algorithms has allowed for accurate prediction and classification of sentiments in movie reviews, enabling users to filter and access reviews based on their preferences.

By reviewing various studies and research works, in this literature paper we have examined the effectiveness of sentimental analysis in improving the accuracy of movie recommendation systems, enhancing user experience, and providing valuable insights into audience opinions and preferences.

Despite the challenges that still exist, such as the need for improved accuracy in capturing the subtleties of human language, sentimental analysis holds great promise for transforming the way we process and interpret textual data. As sentiment analysis technology continues to evolve, we can expect to see even more sophisticated applications in the movie industry and beyond.

In summary, in this literature paper we have developed the sentimental analysis in movie review systems for GUJARATI language which can be further implemented for the other regional languages of the country and highlighted its potential for revolutionizing the way we engage with movies and other forms of media.

V.ACKNOWLEDGMENT

It is a great pleasure for us to acknowledge the assistance and support of a large number of individuals who have been responsible for the successful completion of this project work.
First, we take this opportunity to express our sincere gratitudeto Faculty of Engineering & Technology, Jain Deemed to be University for providing us with a great opportunity to pursue our Bachelor's Degree in this institution.

In particular we would like to thank **Dr. S A Hariprasad,** Director, Faculty of Engineering & Technology and **Dr. G Geetha**, Dean School of Computer Science and Engineering,IJIRCCE©2023 | An ISO 9001:2008 Certified Journal | 5Jain (Deemed-to-be) University for their encouragement and expert advice.

It is a matter of immense pleasure to express our sincere thanks to **Dr. MK Jayanthi Kannan, Head of the department, Computer Science & Engineering,** Jain (Deemed-to-be) University, for providing right academic guidance that made our task possible.

We would like to thank our guide **Dr. MK Jayanthi Kannan**, **Professor, Dept. of Information Science and Engineering,** Jain (Deemed-to-be) University, for sparing his/her valuable time to extend help in every step of our project work, which paved the way for smooth progress and fruitful culmination of the project. We would like to thank our Project Coordinator **Dr. MK Jayanthi Kannan and Prof. M S Sowmya** and all the staff members of Information Science and Engineering for their support.

We are also grateful to our family and friends who provided us with every requirement throughout the course.
We would like to thank one and all who directly or indirectly helped us in completing the Project work successfully.

| e-ISSN: 2320-9801, p-ISSN: 2320-9798| www.ijircce.com | |Impact Factor: 8.379 |

|| Volume 11, Issue 5, May 2023 ||

| DOI: 10.15680/IJIRCCE.2023.1105183 |

REFERENCES

1. Sahu, T. P., & Ahuja, S. (2016). Sentiment analysis of movie reviews: A study on feature selection & classification algorithms. 2016 International Conference on Microelectronics, Computing and Communications.
2. Yasen, M., &Tedmori, S. (2019). Movies Reviews Sentiment Analysis and Classification. 2019 IEEE Jordan International Joint Conference on Electrical Engineering and Information Technology (JEEIT).
3. Nama, V., Hegde, V., & Satish Babu, B. (2021). Sentiment Analysis of Movie Reviews: A Comparative Study between the Naive-Bayes Classifier and a Rule-based Approach. 2021 International Conference on Innovative Trends in Information Technology (ICITIIT).
4. Nanda, C., Dua, M., & Nanda, G. (2018). Sentiment Analysis of Movie Reviews in Hindi Language Using Machine Learning. 2018International Conference on Communication and Signal Processing (ICCSP).
5. Nagamma, P., Pruthvi, H. R., Nisha, K. K., & Shwetha, N. H. (2015). An improved sentiment analysis of online movie reviews based on clustering for box-office prediction. International Conference on Computing, Communication & Automation.
6. Permatasari, R. I., Fauzi, M. A., Adikara, P. P., & Sari, E. D. L. (2018). Twitter Sentiment Analysis of Movie Reviews Using Ensemble Features Based Naïve Bayes. 2018 International Conference on Sustainable Information Engineering and Technology (SIET).
7. Vrunda C. Joshi, Dr. Vipul M. Vekariya An Approach to Sentiment Analysis on Gujarati Tweets Advances in Computational Sciences and Technology ISSN 0973-6107 Volume 10, Number 5 (2017).
8. Lata Gohil, Dharmendra Patel A Sentiment Analysis of Gujarati Text using Gujarati Senti word Net International Journal of Innovative Technology and Exploring Engineering (IJITEE)
9. ISSN: 2278-3075, Volume-8, Issue-9, July2019
10. Qaisar, S. M. (2020). Sentiment Analysis of IMDb Movie Reviews Using Long Short- Term Memory. 2020 2nd International Conference on Computer and Information Sciences (ICCIS).
11. Parita Shah, Priya Swaminarayan, Maitri Patel, Nimisha Patel Sentiment Analysis on Movie Reviews in Regional Language Gujarati Using Machine Learning Algorithm Volume 70 Issue 1, 313-326, January, 2022.
12. Wang,J.-H.,&Liu,T.-W.(2017).Improvingsentimentratingof movie review comments for recommendation. 2017 IEEE International Conference on Consumer Electronics - Taiwan (ICCE- TW)

INNO SPACE
SJIF Scientific Journal Impact Factor

Impact Factor: 8.379

doi cross ref

ISSN INTERNATIONAL STANDARD SERIAL NUMBER INDIA

निस्केयर NISCAIR

INTERNATIONAL JOURNAL OF INNOVATIVE RESEARCH

IN COMPUTER & COMMUNICATION ENGINEERING

9940 572 462 6381 907 438 ijircce@gmail.com

www.ijircce.com

Scan to save the contact details