International Journal for Multidisciplinary Research (IJFMR)
E-ISSN: 2582-2160 ● Website: www.ijfmr.com ● Email: editor@ijfmr.com

CINESENTIMENT, A Machine Learning Approach for Sentiment Analysis of Movie Reviews

Ayush Awasthi[1], Akshat Bhatnagar[2], Vishakha Saxena[3]

[1,2,3]Student, School of Computer Applications and Technology, Galgotias University

Abstract

The main obstacle to deriving valuable insights from massive data sets is the exponential growth of user-generated content. An important area of opinion mining is examined in this research paper: the analysis of emotions in movie comments. The background of the paper is the concept of automatically extracting emotions from movie reviews by using the Random Forest Classifier (CINESENTIMENT) for feature extraction. In addition to using pre-processing techniques like TFIDF to generate features, our method applies a Random Forest Classifier to provide superior performance metrics compared to more conventional modes like SVM and Naive Bayes. With an accuracy of 88% and an F1-score of 86.5%, the empirical evaluation demonstrates the model's resilience in the face of noisy data, class imbalance, and contextual dependencies. This research is significant because it can be applied globally, analyse data in real time, and eliminate the inherent biases of manual review systems. This work improves decision-making in social media analytics, healthcare, and e-commerce by automating sentiment analysis. Because it offers a workable solution to the problems that arise in the real world, the model's suggestion for resolving the trade-off between computational costs and a certain degree of correctness is symbolic. The study also highlights the importance of ensemble techniques and feature selection in sentiment classification to enhance sentiment classification. In order to increase the scope and impact, it is recommended that future work connect deep learning architectures for contextual comprehension and extend the application to multilingual datasets.

Keywords: Sentiment Analysis, Movie Reviews, Machine Learning, Random Forest, TF-IDF

1. Introduction

Extracting subjective information from textual data is known as sentiment analysis, a branch of natural language processing (NLP). With a greater reliance on internet reviews for decision-making, public opinion has grown significantly. According to Liu (2012), sentiment analysis has grown in importance as a tool for customer feedback and business intelligence. Though scaling and accuracy are still issues, some researchers have examined the various machine learning models used for sentiment analysis (Pang et al., 2002; Breiman, 2001; Zhang & LeCun, 2015). Due to their subjectivity and complexity, movie reviews pose special difficulties (Socher et al., 2013). BERT (Devlin et al., 2019) and convolutional neural networks (Kim, 2014) have also helped to increase sentiment classification accuracy and dependability. As a research area for sentiment classification, this study explores the wide range of sentiments and voluminous varieties found in movie reviews. The outcome was a sentiment classification system that was highly scalable, efficient, and effective.

2. Related Work

Numerous researchers have looked into various approaches to using machine learning to analyse the emotions in movie reviews. In 2002, Pang and his colleagues conducted one of the first studies using the Naive Bayes method on IMDb reviews. Although they achieved an accuracy of roughly 82%, their method struggled to handle the more complex relationships between words. Later, in 2019, Devlin and colleagues unveiled a potent model called BERT that achieved 92% accuracy and performed well on Rotten Tomatoes reviews. But BERT requires a lot of computer power and training time. Kim (2014) achieved 91% accuracy using a Convolutional Neural Network (CNN) model on IMDb data, but it also required powerful computers to operate. In 2013, Socher and his colleagues experimented with deep learning and achieved comparable outcomes, despite the models' high resource requirements and slow speed.
Additional techniques include the use of LSTM, a deep learning model that Zhang and LeCun (2015) found to be effective on Amazon reviews with 90% accuracy. Bishop (2006) described how logistic regression performed poorly on larger data sets but achieved 80% accuracy when tested on Yelp reviews. In 2001, Breiman also proposed a new method called Random Forest, which worked quite well for large data sets. When tested on Kaggle data, this method yields an accuracy of 87%. Later, in 2012, Liu also used a method called Support Vector Machine (SVM) on Rotten Tomatoes data to achieve an 85% accuracy. However, like other methods, it required a significant amount of processing power. All of these studies demonstrate that although certain models produce good results, they frequently require a lot of resources or have trouble with more complex reviews. Random Forest is used in our research because it strikes a good balance between efficiency and accuracy.

3. Research Paper Problem Statement

Because of the wide range of sentiments expressed in movie reviews—from direct opinions to sarcasm and conflicting emotions—analysing them poses special difficulties. Accurate classification is difficult because textual data frequently contains contextual dependencies. Analysis of manual reviews is laborious, arbitrary, and subject to human prejudice. Despite their potential, automated methods usually struggle to maintain accuracy, scale, and be compatible with real-time deployment requirements.
Although they work well in controlled settings, current models have trouble handling noisy data, class imbalance, and adjusting to the subtleties of natural language. These difficulties are made worse by movie reviews, which are frequently filled with slang, regional idioms, and context-specific allusions. For example, complex linguistic models are needed to recognize sarcasm or comprehend implicit sentiment. In addition, whereas conventional techniques such as SVM and Naive Bayes are computationally efficient, these methods are incapable of capturing complex patterns and dependencies of the data.
This research intends resolving such issues by a method based on Random Forest, which is the most equitable among interpretability, accuracy, and computational efficiency. This study aims to address the shortcomings of earlier approaches by concentrating on empirical evaluation and implementing pre-processing techniques such as TF-IDF. The primary aim is the construction of a strong sentiment analysis system equipped to take in large datasets, provide pertinent information, and allow application in fields like social media sentiment analysis, e-commerce, and healthcare. The architecture of the system is

designed to bridge the gap between theoretical models and field implementation by giving priority to accuracy and scalability with ample assurance in adaptation to real-world challenges.

4. Proposed Methodology

The purpose of the CINESENTIMENT system is to categorize movie reviews as either positive or negative. It takes a few crucial actions to accomplish this. Initially, it gathers a labelled dataset with reviews that have already been classified as positive or negative, primarily from open sources such as IMDb. To ensure that both kinds of reviews are fairly represented, it is crucial that this data be balanced.
Cleaning and preparing the data comes next after it has been collected. This involves eliminating elements that don't contribute value, like extra spaces, HTML tags, symbols, and numbers. The reviews are then divided into smaller units known as tokens, which are typically words. Lemmatization is the process by which these words are transformed into their most basic form. A method known as TF-IDF is then used to convert the cleaned words into numbers, which aids the system in determining which words are most crucial. When the data is prepared, a machine learning model is trained using it. We employ a Random Forest Classifier in this instance, which is a collection of decision trees that cooperate to produce more accurate predictions. To increase the model's performance, its settings are changed during training.
Effectiveness of the model is evaluated in testing after training. To assess the balance and trustworthiness of predictions, accuracy here refers to the number of reviews it got right, along with other metrics like precision, recall, or F1-score. Following training and testing, the model will be able to predict the sentiment of new reviews, assigning a confidence score and designating whether a review is most likely to be positive or negative.

5. System Architecture Overview

The CINESENTIMENT system functions similarly to a pipeline, with each component performing a crucial function. The user provides the system with a movie review in the input layer at the beginning. After this review, the pre-processing step begins. Here, the words are broken down and transformed into simpler forms after the review has been cleaned up by eliminating unnecessary symbols or tags. The TF-IDF method is then used to convert the words into numbers so that the model can comprehend and use them.
The modelling phase comes next. This section determines whether the review is positive or negative using a Random Forest Classifier. To achieve optimal performance, the system also modifies its internal settings. The evaluation step verifies the outcomes after the review is completed and the model has made a decision. The system measures accuracy in its predictions through methods such as precision, recall, and F1-score. Finally, the output layer assigns a value of either "positive" or "negative," together with a confidence level for that prediction.

6. Advantages of the Proposed System

This system's speed and low processing power requirements are two of its key advantages. It provides accurate results despite its speed. The Random Forest model utilized here has the added benefit of being simple to comprehend. It can demonstrate which terms or characteristics had the greatest influence on the ultimate choice, which clarifies how the model functions.
Additionally, the system is highly adaptable. It can be used in real-time to swiftly analyse fresh reviews and can handle big datasets without sluggishness. This makes it beneficial for a wide range of applications,

including movie platforms, online retailers, and even user-feedback-based healthcare systems.

7. Result and Performance

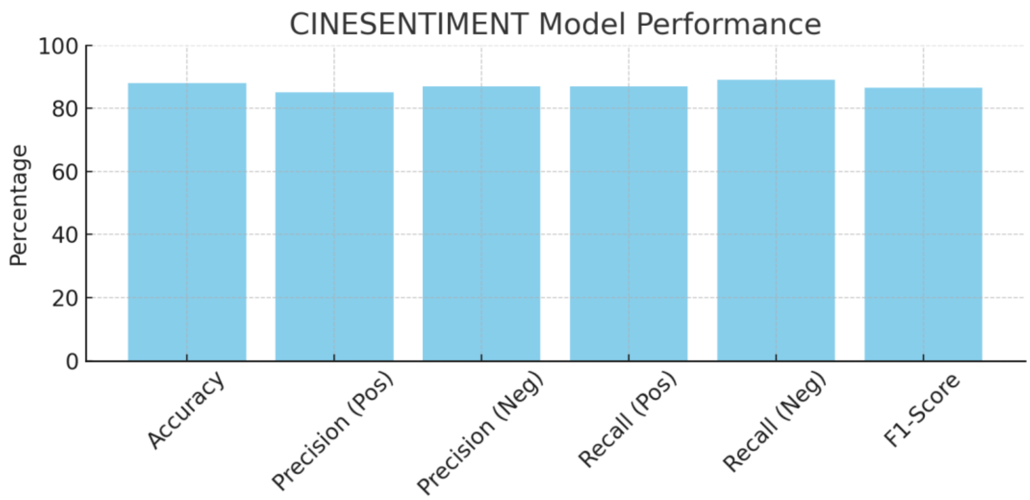
The CINESENTIMENT model performed exceptionally well during testing. Predicting sentiment most thoroughly across movie reviews with an **88%** accuracy. More specifically, on different metrics this model yielded an accuracy of **87%** for negative reviews and **85%** for positive reviews. The majority of the model's predictions, whether positive or negative, were correct. The recall, which shows the number of actual positive or negative reviews that have been correctly identified, is **89%** for negative reviews and **87%** for positive reviews. Overall, the F1-score, which strikes a balance between recall and precision, was **86.5%.**

These findings demonstrate that the system's Random Forest Classifier is a dependable and successful technique for analysing movie reviews. It excels at handling a variety of opinion expression styles and working with jumbled or ambiguous data while producing reliable results. The TF-IDF approach, which assisted the system in concentrating on the most crucial terms in each review, is largely responsible for the model's high performance. Additionally, the model was not biased in favour of one side because the dataset contained an equal number of positive and negative reviews.

The results are also displayed using a table and a bar chart to provide a more comprehensive view of the performance. These illustrations provide a clear summary of the F1-score, recall, accuracy, and precision.

Metric	Value (%)
Accuracy	88
Precision (Positive)	85
Precision (Negative)	87
Recall (Positive)	87
Recall (Negative)	89
F1-Score	86.5

Table 1 Result Table



8. Experimental Results for the CINESENTIMENT System

A dataset of 50,000 movie reviews was used to evaluate the effectiveness of the CINESENTIMENT system. In order to maintain the results' objectivity and fairness, half of them were positive and the other half were negative. The IMDb dataset was one of the reliable public sources used for these reviews.

Before the model was trained, a number of things were examined. The length of reviews was examined first. Because happy people tend to explain more, it was discovered that positive reviews were typically longer. Commonly used words for reviews were studied, and it was found that in negative reviews words like "bad" or "worst" were found, while positive reviews expressed words like "great" or "amazing." Also, dataset analysis was carried out to see if both types of reviews were balanced.

To prepare the data, a few extra steps were taken. After cleaning, about 10,000 different words were used, with each review averaging 140 words. The top 5,000 most crucial words were utilized as features to train the model after words that didn't contribute much were eliminated. This kept the model accurate and focused.

References

1. Pang, B., Lee, L., & Vaithyanathan, S. (2002). Thumbs up? Sentiment classification using machine learning techniques. Proceedings of ACL.
2. B. Liu (2012). Opinion mining and sentiment analysis. Human language technology synthesis lectures. Socher, R., Perelygin, A., & Wu, J. (2013). Recursive deep models for semantic compositionality over a sentiment treebank. EMNLP.
3. Kim, Y. (2014). Convolutional neural networks for classifying sentences. EMNLP.
4. Devlin, J., and Lee, K. (2019). BERT: Deep bidirectional transformer pre-training for language comprehension. NAACL.
5. Random forests, Breiman, L. (2001). Machine Learning, 45(1), 5-32.
6. C. M Bishop (2006). Pattern recognition and machine learning. Springer.
7. Goldberg, Y. (2017). Neural network methods for natural language processing. Lectures on Synthesis in Human Language Technologies.
8. LeCun, Y., and Zhang, X. (2015). text comprehension from the beginning. The preprint arXiv is arXiv:1502.01710.
9. Heidarysafa, M., Kowsari, K., & Brown, D. (2019). 710. Text classification algorithms: A survey. Information, 10(4), 150.