21-101017

21-101084

21-101030

[Email address]

Submitted to: Dr. Mohamed Taher

Faculty of Computing and Information Sciences

Egypt University of Informatics

Machine learning project

Sentiment Analysis of Movie Reviews

Abstract

This paper presents the development and evaluation of a machine learning model for sentiment analysis of movie reviews using the Natural Language Toolkit (NLTK). By leveraging NLTK for data preprocessing and feature extraction, the model demonstrates significant accuracy and robustness.

Introduction

Pattern recognition and learning theory in artificial intelligence are the foundations of machine learning, a subfield of computer science. It's now necessary for a wide range of fascinating data analytics jobs. Getting right into the data is a potent way to obtain insights and generate predictions, especially with the proliferation of data sources and the processing power to handle them.

A list of sub-problems, such as decision-making, clustering, classification, forecasting, deep learning, inductive logic programming, support vector machines, reinforcement learning, similarity and metric learning, genetic algorithms, sparse dictionary learning, etc., can be thought of as the study of machine learning. The machine learning task of determining a function from labelled data is known as supervised learning, or classification .

Sentiment analysis is an essential part of natural language processing (NLP), especially when examining reviews of films, for example. We can learn important lessons from analysing the sentiment reflected in these reviews, which will help many stakeholders in the film industry.

Sentiment analysis assists filmmakers in comprehending the reactions of viewers to their films. While negative sentiment might point out areas for improvement, positive sentiment shows that audiences are enjoying the film. Future casting decisions, marketing plans, and production decisions can all be influenced by this input. Sentiment analysis also helps critics by giving them a better grasp of how audiences respond to films. Critics can spot trends, patterns, and recurring themes in a variety of films by examining the opinions expressed in reviews

Problems and issues in sentiment based models:

Creating and implementing sentiment-based models, like the ones used to evaluate movie reviews, requires overcoming a number of obstacles. One of the main issues is handling sarcasm and ambiguity in writing. Models find it challenging to accurately interpret sentiment because words and phrases can have multiple meanings depending on the context in which they are used. For example, depending on further context, the statement "It was a bomb" could be interpreted positively or negatively. The intended meaning of sarcasm and irony is frequently the opposite of what is said literally, which adds another level of complexity. Without realizing the irony, a statement like "Great, another boring movie" could be mistakenly categorized as positive.

The use of domain-specific language in various contexts presents another difficulty. It can be difficult for a model to comprehend slang specific to the film industry or the demographic writing the reviews if it hasn't been trained on comparable data. Slang terms such as "This movie slaps" may cause confusion for the model if it is not up to date. Additionally, language is always changing. As new words and expressions appear over time, the model must be updated frequently to remain accurate.

Another big problem is managing polarity shifts and sentiment that changes depending on the situation. The context in which a word is used can alter the meaning attached to it. "Fast" can have a positive meaning in "fast performance" but a negative meaning in "fast deterioration." It's also important to recognize and understand negations correctly, such as the distinction between "I like this movie" and "I don't like this movie."

Two important aspects of training data are its quality and quantity. An imbalance between positive and negative examples is common in sentiment datasets, which can skew the model's predictions. Effective training requires high-quality annotated data, but acquiring such data is frequently expensive and time-consuming. Furthermore, sentiment analysis in multiple languages presents unique difficulties since models must comprehend the subtleties and colloquial expressions of every language. Domain adaptation is also required because, because language use and sentiment expression differ between one domain (e.g., movie reviews) and another (e.g., product reviews), models trained on data from one domain may not perform well on the other.

Dataset

The dataset used is a sample of IMDb reviews over random films. There are 50000 movie review vectors in the dataset, In order to gain insight into audience reactions and assist consumers, critics, and producers in making decisions, sentiment analysis has been applied extensively in the entertainment industry. Examining the sentiment contained in movie reviews can help identify trends in viewer preferences and enhance comprehension of the factors influencing favourable or unfavourable opinions about a film.

Because an accurately generalised sentiment analysis model can give producers detailed feedback on how their films are received, this classification task is crucial. By recognising patterns in both positive and negative sentiment, producers can use this information to inform their decisions about upcoming projects and marketing campaigns. Critics can increase the relevance and impact of their critiques by using sentiment analysis to match their reviews with audience perceptions. Sentiment analysis helps customers choose films that fit their tastes by providing a brief synopsis of the overall attitude surrounding a film.

Furthermore, sentiment analysis can lessen the possibility of human error when gauging audience sentiment. A producer or critic's subjective biases or narrow viewpoint may cause them to misinterpret the overall sentiment, but an automated system offers a thorough and impartial analysis. The accuracy and dependability of sentiment interpretation are improved by this marriage of human experience and machine learning, which improves decision-making and produces more fruitful results in the film industry.

Methodology

Our methodology is centered around the use of the TF-IDF (Term Frequency-Inverse Document Frequency) vectorizer coupled with a machine learning classifier to categorize sentiments in movie reviews. The dataset is composed of a single column of textual reviews. To prepare the data, preprocessing steps are essential and include converting all text to lowercase to ensure consistency, removing punctuation to prevent sentiment scores from being skewed, and removing stop words that do not add to the sentiment. Each review is then processed by the TF-IDF vectorizer, which transforms the text into numerical vectors representing the importance of each word in the review. These vectors are then fed into a machine learning classifier for sentiment classification.

Experimental Design

The steps involved in our experimental design include:

1. Data Cleaning: Removing any null entries and ensuring uniformity in the text.

2. Text Preprocessing: Applying functions to convert text to lowercase, remove punctuation, and eliminate stop words.

3. Sentiment Analysis: Using the TF-IDF vectorizer to transform text into numerical vectors and a machine learning classifier to predict sentiment.

4. Model Evaluation: Splitting the dataset into training and testing sets, training the classifier, and evaluating its performance using accuracy and classification report metrics.

Experimental Results

Training Accuracy: 0.911975

precision recall f1-score support

negative 0.92 0.91 0.91 20039

positive 0.91 0.92 0.91 19961

accuracy 0.91 40000

macro avg 0.91 0.91 0.91 40000

weighted avg 0.91 0.91 0.91 40000

Validation Accuracy: 0.8959

precision recall f1-score support

negative 0.90 0.88 0.89 4961

positive 0.89 0.91 0.90 5039

accuracy 0.90 10000

macro avg 0.90 0.90 0.90 10000

weighted avg 0.90

0.90 0.90 10000

The model appears to be performing well with:

• High training accuracy (91.2%) and validation accuracy (89.6%).

• Good precision, recall, and F1-scores for both classes on training and validation data.

• The slight drop in validation accuracy compared to training accuracy suggests a well-fitting model without significant overfitting.

However, it's noteworthy that initially, we attempted to utilize the VADER sentiment analyzer for sentiment classification. However, due to its inability to accurately classify sentiments in our movie review dataset, we opted for the TF-IDF vectorizer coupled with a machine learning classifier, which provided more reliable results.

Conclusion

Our project demonstrates the applicability of the TF-IDF vectorizer coupled with a machine learning classifier for sentiment analysis of movie reviews, achieving a reasonable accuracy for sentiment classification. The preprocessing steps significantly improve the quality of input data, and the results suggest that TF-IDF, despite its simplicity, can serve as an effective approach for sentiment analysis in this context. Future work could involve exploring hybrid models that combine rule-based and machine learning approaches to better capture the sentiment nuances in longer texts.

References

Mohri, M., Rostamizadeh, A. and Talwalkar, A. (no date) Foundations of Machine Learning. Second Edition. Available at: https://mitpress.ublish.com/ebook/foundations-of-machine-learning--2-preview/7093/i.

‘IMDb dataset’ (no date). Available at: https://www.kaggle.com/datasets/lakshmi25npathi/imdb-dataset-of-50k-movie-reviews.