Sentiment analysis for IMDB movie reviews.

Intiajul A. Shah

Email: md.intiajul.alam.shah@g.bracu.ac.bd

***Abstract*—In this research we analyze reviews from the Internet Movie Database (IMDb) using sentiment analysis, a branch of natural language processing (NLP). State of the art results in deep learning models have shown remarkable works, despite many difficulties in natural language processing. We use different machine learning techniques to categorize emotional representations of IMDb reviews including linear regression(LR), SVM and Multinomial Naive Bayes(MNB). To enhance categorization performance, stop words are eliminated and word normalization is applied to the reviews. Additionally, word embeddings and the TF-IDF are used for a number of datasets in this work. A comparative study is conducted on the experimental results obtained for the different models and input features. We hope these findings will contribute to the growing field of work on sentiment analysis, providing new insights into the application of machine learning(ML) techniques in this field.**

Keywords: Sentiment, Natural Language Processing (NLP), Emotional View, IMDb Reviews.

I. Introduction

As the time and technology is growing, ML has emerged as a key field of research since it makes processing massive volumes of data more effective. This approach has demonstrated particular utility in Natural linguistic Processing (NLP), supporting the classification and prediction of linguistic content. Sentiment analysis is a widely used technique in natural language processing (NLP) that may be used at three different levels: Sentence, Document, and Aspect levels. The objective of this research is to use machine learning to learn and find out the sentimental perspectives of Internet Movie reviews.

Challenges in NLP have made sentiment analysis less accurate and efficient, but new developments in deep learning models have shown remarkable results. To categorize the sentimental representations, we use machine learning techniques such as linear regression(LR), SVM and Multinomial Naive Bayes(MNB). To enhance classification performance, stop words are eliminated and the word normalization is applied to the reviews. The features for the classification are then represented by a word matrix created from the reviews.

Additionally, word embedding, the TF-IDF are used for a number of datasets in this work. An analysis is carried out to compare the experimental outcomes for various models and input characteristics. By offering new perspectives, this study adds to the expanding corpus of research in this area.

II. Dataset

The Internet Movie Database (IMDb) has fifty thousand (50K) reviews that have been classified as either positive or negative. The binary sentiment analysis dataset utilized in this work is called the IMDb Movie Reviews dataset. There is an equal amount of good and negative evaluations in the dataset, ensuring balance in the data. It only contains reviews that are extremely divisive, with a negative review receiving a score of four or less and a good review receiving a score of seven or higher. Reviews for each movie are limited to thirty. There is more unlabeled data in the dataset.

Because of its range and depth, sentiment analysis research has made extensive use of the IMDb Movie Reviews dataset. It offers a wealth of user-generated content, which can offer insightful information on consumer behavior and public opinion. The English-language dataset may be accessed via a variety of data loaders, including pytorch, tensorflow, and others. Because of its CC-BY-SA license, the dataset is openly accessible for use in research projects with proper acknowledgment to the original author.

The actual review texts from the dataset are preprocessed step by step to make the data interpretation easier for the study. Firstly, the punctuations, line break, white spaces, numerals, and stop words like ”a,” ”the,”, ”of” are eliminated since they don’t provide anything about the viewer’s opinion of a movie. Subsequently, to lessen vocabulary noise, all words are changed to lowercase and normalized to their real roots (e.g., played to play). Throughout this research, we employ four distinct vectorization techniques: binary, word-count, n-grams, and tf-idf vectorization.

III. Methodologies

For many years, natural language processing (NLP) has been a part of artificial intelligence (AI) research. Among its numerous uses is sentiment analysis, a branch of the science that aims to detect and extract subjective information from source materials. The IMDb Movie Reviews dataset is subjected to a series of processes in the approach that are all intended to efficiently prepare the data and use sentiment analysis algorithms.

A. Methods

- Data Preprocessing
- Vectorization
- Model Training and Testing

B. Data Preprocessing

Preparing the data is the first stage in the methodology. In order to do this, the raw text data from the IMDb Movie Reviews dataset must be cleaned. Eliminating the punctuations, line break, white spaces, numerals, and stop words like ”a,” ”the,” ”of” are all part of the process. These components are

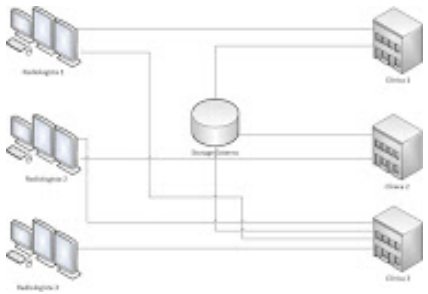

Fig. 1: Usuários acessando o Storaged externo.

eliminated as they don't reveal anything about how a user feels about a film. To lessen vocabulary noise, the words are then normalized to their real root (e.g., 'played' to 'play') and transformed to lower cases.

C. Vectorization

Following preprocessing, a word matrix representing the classification's features is created from the reviews. Converting texts input into numerical representations to make it understandable for machine learning algorithms is called vectorization. This project utilizes four distinct vectorization techniques: binary vectorization, word-count vectorization, n-gram vectorization, and tf-idf vectorization. I.

TABLE I: Um exemplo de tabela

One	Two
Three	Four

D. Model Training and Testing

The machine learning models are trained and tested on vectorized data following sentiment analysis. Evaluation criteria are used to assess how well these models perform in terms of categorizing review emotions, including accuracy, precision, recall, and F1 score.

Using the gradient descent approach, a weight vector is initialized and updated during the training phase. The initialization weight vector and training data are subjected to the gradient descent function with a learning rate of 0.01 and 100 epochs. By changing weights at each epoch, the goal is to minimize the cost function which calculates the difference between the predicted outputs and actual outputs. The cost function is computed using the cross-entropy loss, which is a frequently used function for binary classification issues.

Test data is used to evaluate the model after it has been trained. By contrasting the expected and actual feelings in the test data, the correctness of the model is determined. If the projected feeling is more than or equal to 0.5, the sentiment is defined as positive; if not, it is labeled as negative.

The confusion matrix of the model is calculated in addition to accuracy. When assessing the model's performance on a set of test data with known real values, this matrix is a helpful table. It improves our comprehension of the model's effectiveness by offering a thorough explanation of false-positives, false-negatives, true-positives, and true-negatives.

As we study so far, the training and testing processes are repeated, and the results are then compared, in order to determine which algorithm performs the best in sentiment categorization.

IV. Result

From our findings it demonstrates the potential of using several machine learning algorithms for sentiment analysis based on the model's ratings of accuracy and other scores.

We found the output from the Logistic Regression 0.75 F1 score, 0.75 accuracy, 0.75 precision, and 0.75 recall. The results for Support Vector Machines (SVM) showed 0.58 F1 score, 0.74 precision, 0.58 recall, and 0.58 accuracy. With 0.75 accuracy, 0.75 precision, 0.76 recall, and a 0.75 F1 score, Multinomial Naive Bayes fared better.

It's important to remember that the outcomes might change based on the particular dataset and preprocessing techniques applied. To determine which setup and algorithm is ideal for a particular task, it is therefore essential to experiment with many options.

These findings highlight the significance of choosing the right machine learning algorithm for sentiment analysis jobs overall. While deep learning models frequently outperform classic machine learning algorithms, they may nevertheless provide good results, indicating the potential of these cutting-edge methods in natural language processing.

References