Analysing the sentiment of IMDB Movie Reviews using NLP techniques.

Intiajul A. Shah, Sohrab Hossain, and Annajiat Alim Rasel
BRAC University, Merul Badda, Dhaka

(md.intiajul.alam.shah, sohrab.hossain)@g.bracu.ac.bd, annajiat@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 logistic regression(LR), Random Forest(RF) and Gradient Boosting(GB). 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.[1] 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 logistic regression(LR), Random Forest(RF) and Gradient Boosting(GB). 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.[2]

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.[3]

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.[4] We divided the data into half and chose 70 percent for train and 30 percent for testing.

A. Methods

- Data Preprocessing

review	sentiment
One of the other reviewers has mentioned that after watching just 1 Oz episode you'll be hooked. The...	positive
A wonderful little production. The filming technique is very unassuming- very old-time-B...	positive
I thought this was a wonderful way to spend time on a too hot summer weekend, sitting in the air con...	positive

Fig. 1: Compact view of the dataset.

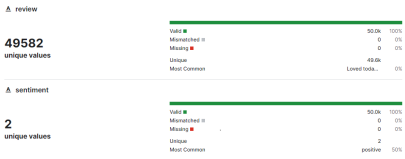

Fig. 2: Column view of the dataset.

- Vectorization
- Model Training and Testing

B. Data Preprocessing

In the initial stage of the methodology for sentiment analysis on the IMDb Movie Reviews dataset, data preparation is very crucial. Firstly, concatenate all reviews into a single string and generate a word cloud to visualize the most frequent words in all reviews. Extract all reviews for the specified sentiment from the dataset and generated a word cloud for sentiment-specific reviews. We converted all text to lowercase which helps in treating uppercase and lowercase versions of the same word as identical. All the HTML text were romoved using regular expressions, leaving only the plain text content. We also removed hashtags and digits from the text, It helps in getting rid of social media hashtags and numerical values. We also removed non-alphabetic characters and URLS. We Downloads and sets up the English stopwords to remove common stopwords such as ”a,” ”the,” ”of,” etc. These words are frequent but generally do not contribute much to the sentiment and can be considered noise. We performed

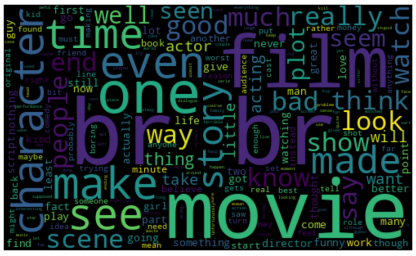

Fig. 3: Wordcloud of negative data.

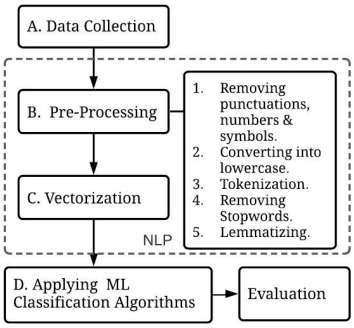

Fig. 4: Work process.

stemming on the text using the Stemmer from NLTK to group together variations of words that have the same meaning but different inflections. For example, ”running,” ”ran,” and ”runs” would all be stemmed to the base form ”run”. Strip leading and trailing spaces were done to ensures that there are no unnecessary spaces at the beginning or end of the processed text. It helps in maintaining consistency and cleanliness.

C. Vectorization

Vectorization is a technique used to convert a collection of text documents into numerical vectors. By using TF-IDF vectorization we are capturing the significance of each word in relation to the entire corpus of documents, with each document being represented as a vector. In order to prepare text data for machine learning models, this step is essential.

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

- Random Forest Classifier:

As part of an ensemble learning technique called Random Forest, multiple decision trees are constructed during training, and their predictions are combined to increase overall performance. The training data that was processed by TF-IDF transformed is used to train the Random Forest classifier. A random collection of features and a random subset of training data are used to build each decision tree. Following training, the predictions of each individual decision tree are combined

by the Random Forest to provide predictions for the test data . The model's performance on the test data is assessed using a variety of classification measures, including accuracy, precision, recall, and F1-score. A bar chart's accuracy score is represented by a green bar, which gives an immediate graphical representation of the Random Forest model's performance.

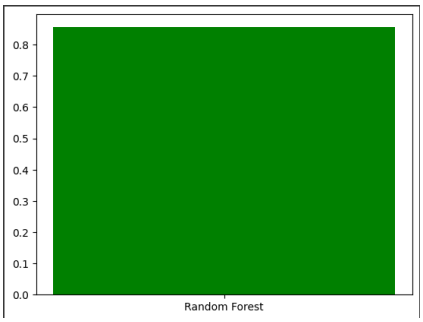

Fig. 5: Random Forest Classifier.

- Gradient Boosting Classifier:

Another ensemble learning technique is gradient boosting, which creates decision trees one after the other by fixing the mistakes of the preceding ones. Like Random Forest, the Gradient Boosting classifier is trained on TF-IDF transformed training data. Each tree focuses on the errors committed by the others, and predictions are made on the test data by integrating the predictions of individual trees in a sequential manner. Classification metrics are calculated to evaluate the model's performance on the test data, similar to the Random Forest. A red bar in a bar chart represents the accuracy score, making it possible to compare the model's performance to that of other models.

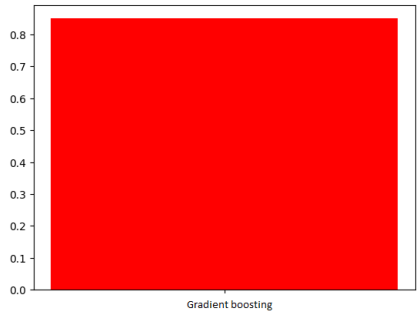

Fig. 6: Gradient Boosting Classifier.

- Logistic Regression:

Logistic Regression is a linear classification algorithm that models the probability of a binary outcome. The training data that has been TF-IDF converted is used to train the Logistic Regression model. Using a logistic function to convert the linear combination of attributes into probabilities, predictions are produced using the test data. To assess the model's performance on the test data, classification metrics are computed for accuracy, precision, recall, and F1-score. A blue bar on a bar chart represents the accuracy score and shows how the model performs in comparison to other models.

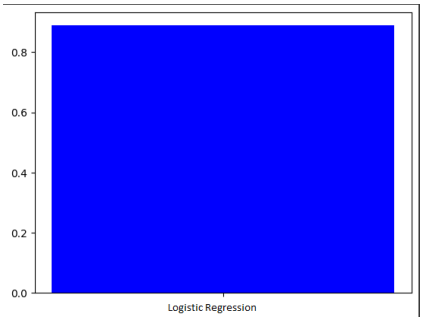

Fig. 7: Logistic Regression Classifier.

```
Rvw_train, Rvw_test , Snt_train , Snt_test= train_test_split(Rvw , Snt , train_size=0.7)
```

Fig. 8: Function for split, train, test.

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 Random Forest 0.86 F1 score, 0.86 accuracy, 0.86 precision, and 0.85 recall. The results for Gradient boosting showed 0.86 F1 score, 0.86 precision, 0.86 recall, and 0.86 accuracy. With 0.89 accuracy, 0.89 precision, 0.89 recall, and a 0.89 F1 score, Logistic regression 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.

	precision	recall	f1-score	support
0	0.86	0.85	0.85	7395
1	0.85	0.86	0.86	7480
accuracy			0.86	14875
macro avg	0.86	0.86	0.86	14875
weighted avg	0.86	0.86	0.86	14875

Fig. 9: Accuracy of Random Forest.

	precision	recall	f1-score	support
0	0.87	0.82	0.84	7395
1	0.83	0.88	0.86	7480
accuracy			0.85	14875
macro avg	0.85	0.85	0.85	14875
weighted avg	0.85	0.85	0.85	14875

Fig. 10: Accuracy of Gradient Boosting.

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

	precision	recall	f1-score	support
0	0.90	0.87	0.89	7395
1	0.88	0.90	0.89	7480
accuracy			0.89	14875
macro avg	0.89	0.89	0.89	14875
weighted avg	0.89	0.89	0.89	14875

Fig. 11: Accuracy of Logistic Regression.

provide good results, indicating the potential of these cutting-edge methods in natural language processing.

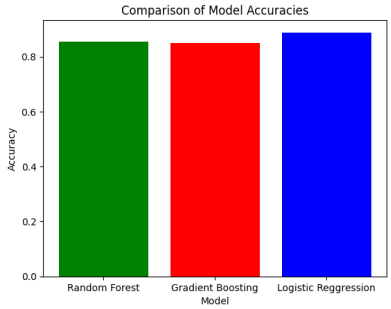

Fig. 12: Bar chart of comparing the results.

V. Future Work

We believe there are many scopes for further investigation and implementing the models. To expand our knowledge of the advantages and disadvantages of the machine learning techniques applied in sentiment analysis may be the main goal of future research. This can entail researching the fundamental principles of existing algorithms, refining their specifications, or creating new iterations. We have also opened the dataset for anyone to use and make their own type of modification.

VI. Conclusion

We have used this dataset in the study to illustrate sentiment analysis using machine learning techniques such as Random Forest, Gradient boosting and logistic regression. This result highlights the promise of these cutting-edge methods in the field of natural language processing (NLP). Even with the amazing outcomes, the study also identifies a number of areas that need more research. The models' performance might differ based on the particular dataset and preprocessing techniques applied, indicating the need for more thorough testing and validation. Furthermore, compared to standard models, deep learning models performed better but also required more processing power and longer training cycles.

Further study may be done in the future to improve existing models and create more accurate algorithms designed specifically for sentiment analysis jobs. It would also be advantageous to look at how various feature extraction and data preprocessing strategies affect the models' functionality. In order to evaluate these models' effectiveness and practicality, the research might also investigate how these models are applied in actual situations, such as social media monitoring or consumer feedback analysis.

References

[1] "Papers with Code - Exploring the Limits of Transfer Learning with a Unified Text-to-Text Transformer," 10 2019. [Online]. Available: https://paperswithcode.com/paper/exploring-the-limits-of-transfer-learning

[2] "Machine Learning based classification for Sentimental analysis of IMDb reviews," *Machine Learning based classification for Sentimental analysis of IMDb reviews*.

[3] "Papers with Code - IMDb Movie Reviews Dataset." [Online]. Available: https://paperswithcode.com/dataset/imdb-movie-reviews

[4] "IMDB dataset of 50K movie reviews," 3 2019. [Online]. Available: https://www.kaggle.com/datasets/lakshmi25npathi/imdb-dataset-of-50k-movie-reviews