Sentiment Analysis study on IMDB dataset

Waddah Alhajar / Sapienza University of Rome
alhajar.2049298@studenti.uniroma1.it

Abstract

This project investigates sentiment analysis on the IMDB dataset which consists of 50,000 movie reviews. The primary objective of this study is to evaluate and compare the performance of sentiment classification models using two distinct approaches: traditional TF-IDF vectorization with machine learning algorithms, and BERT text classification model.

The study begins with cleaning, preprocessing, and analysis of the IMDB dataset, then the dataset is split into training and testing sets. Traditional machine learning models, including Multinomial Naive Bayes, Support Vector Machine (SVM), and Multi-Layer Perceptron (MLP) were trained using TF-IDF vectorization, a widely adopted technique for text data.

Subsequently, the research explores the utilization of BERT-based model for sequence classification, which is a state-of-the-art transformer-based model, in which I used the ”bert-base-uncased” pre-trained model.

Performance metrics such as accuracy, precision, and recall were used to assess the effectiveness of each model. This empirical study contributes to the broader understanding of different sentiment analysis techniques and algorithms.

1 Task description/Problem statement

Sentiment analysis is defined as the process of obtaining meaningful information and semantics from text using natural processing techniques and determining the writer’s attitude, which might be positive, negative, or neutral [1]. This NLP task holds significant importance in various domains, including marketing, customer feedback analysis, social media monitoring, and product development. Understanding the sentiment behind textual data can provide valuable insights for businesses, organizations, and researchers.In this case study we are taking into consideration Movies reviews in particular.

1.1 Examples

Table 1: Sentiment Analysis Input and Output Example

Input Sentence	Sentiment Output
”The movie was captivating, with great acting and a compelling storyline.”	Positive
”It was terrible,didn’t enjoy.”	Negative
”I love this movie! It was amazing.”	Positive
”The film was boring.”	Negative

1.2 Real-world applications

Sentiment analysis plays a crucial role in movies industry, especially for streaming platforms like Netflix. Applications of my sentiment analysis models that can be applied in the context of movie platforms are:

1- Content Recommendation: Netflix and similar platforms heavily rely on recommendation systems to suggest movies and TV shows to users. Sentiment analysis can be used to analyze user reviews, ratings, and viewing history to make personalized recommendations. For example, if a user consistently rates and watches action movies positively, the recommendation system can prioritize action movies in their suggestions.

2-Content Curation: Sentiment analysis can assist in curating content libraries. By analyzing user reviews and sentiments, streaming platforms can identify popular and well-received movies and prominently feature them on their platform. Conversely, poorly rated or negatively reviewed content can be de-prioritized or removed from the catalog.

3-Genre Preferences: Understanding user sentiment towards different movie genres is crucial. Sentiment analysis can help in identifying which genres are trending or preferred by specific user segments. Platforms can then invest in producing or licensing more content in these genres.

4-Viewer Engagement: Sentiment analysis can be used to gauge viewer engagement during a movie or TV show. By analyzing real-time sentiments expressed on social media while users are watching, platforms can identify key moments that generate strong positive or negative reactions. This information can be used to improve content or marketing strategies.

5-Quality Assessment: Sentiment analysis can assist in assessing the quality of original content produced by the platform. By analyzing user reviews and sentiments for Netflix Originals, for example, the platform can determine which shows or movies are resonating with viewers and make data-driven decisions about future productions.

6-Predicting Movie Success: Sentiment analysis can be used to predict the success of upcoming movies or TV shows. By analyzing pre-release sentiment from trailers, promotional materials, and early reviews, platforms can estimate audience interest and adjust marketing strategies accordingly.

2 Related work

A paper have reported that SVM performed better classification with 82.9 % accuracy compared with the Naive Bayes that achieved 81% on positive and negative movie reviews from the IMDB movie reviews dataset that consists of 752 negative and 1301 positive reviews [9]. The researchers of another paper have proposed Recursive Neural Tensor Network (RNTN) model for identifying sentences as positive or negative by using fully labelled parse trees. They used a dataset based on a corpus of 11,855 movie reviews. The RNTN have obtained 80.7% accuracy on fine-grained sentiment prediction across all phrases and captures different sentiments and scope more accurately than previous models [8]. A third paper has focused on customer reviews. Products reviews were collected from Amazon.com (11,754 sentences). It proposed a novel deep learning framework named Weakly supervised Deep Embedding for reviewing sentence sentiment classification with accuracy 87% [3] Another work that was published in 2018 performed several deep learning models for a binary sentiment classification problem. They used movie reviews in Turkish from the website www.beyazperde.com to train and test the deep learning models [5]. The whole dataset consists of 44,617 samples including positive and negative reviews. 4000 samples from the dataset were used for testing the models. Two major deep learning architectures were applied, CNN and LSTM. Word embedding were created by applying the word2vec algorithm with a skip-gram model on the used dataset. Experimental results have shown that the use of word embedding with deep neural networks effectively yields performance improvements in terms of run time and accuracy. Lastly, [1] have used the same IMDB dataset that we are using with MLP, CNN, LSTM, and a hybrid model CNN-LSTM. Their results are shown later in the comparative evaluation section.

3 Datasets and benchmarks

1- IMDB dataset: it consists of 50,000 movie reviews,it is a valuable resource for natural language processing and text analytics tasks. Unlike earlier benchmark datasets, this collection stands out due to its significant volume. It is labeled with positive or negative sentiment for each review [6]. The Amazon reviews dataset: it consists of reviews from amazon. The data span a period of 18 years, including 35 million reviews up to March 2013. Reviews include product and user information, ratings, and a plaintext review.[7]

4 Existing tools, libraries, papers with code

Scikit-Learn: I used Scikit-Learn, a popular machine learning library in Python, for traditional machine learning tasks such as TF-IDF vectorization and training classifiers like SVM, Naive Bayes, and MLP. It provides a wide range of tools for data preprocessing, feature extraction, and model evaluation.

Hugging Face Transformers: Hugging Face provides pre-trained transformer-based models like BERT, GPT, and more, along with easy access to their embeddings. I used the Hugging Face Transformers library to load pre-trained BERT model and extract embeddings for sentiment analysis.

TensorFlow and Keras: They are used in deep learning tasks, for building and training neural networks. These libraries provide the flexibility to create custom deep learning models.

NumPy and Pandas: NumPy and Pandas are used for data manipulation, including loading and preprocessing the dataset, as well as for converting data to compatible formats for machine learning models.

Papers:1-BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding" by Jacob Devlin et al. This paper introduces the BERT model, which has become a standard for natural language understanding tasks, including sentiment analysis. [2]

2-ULMFiT: Universal Language Model Finetuning for Text Classification" by Jeremy Howard and Sebastian Ruder. This paper presents the ULMFiT approach, which is an alternative to BERT for transfer learning in text classification tasks. Code and pre-trained models are available in the fastai library. [4]

5 State-of-the-art evaluation

Most systems are examined by splitting their labeled data to training and testing sets and use the accuracy as the main performance measure in case of a balanced dataset or other measures including precision, recall, and F1-score when there is data imbalance. Accuracy: Accuracy measures the proportion of correctly classified instances (positive or negative sentiments) out of the total number of instances. While accuracy is a straightforward metric, it may not be sufficient for imbalanced datasets, where one sentiment class significantly outweighs the other.

Precision and Recall: Precision and recall are often used in binary sentiment classification. Precision measures the proportion of true positive predictions (correctly identified positive sentiments) out of all positive predictions made by the model. Recall, also known as sensitivity, measures the proportion of true positive predictions out of all actual positive instances in the dataset. The F1 score is a measure of a model's accuracy that considers both the precision and recall of the model. It is particularly useful when dealing with imbalanced datasets. The F1 score is calculated using the following formula: F1= 2.(P.R)/(P+R)

6 Comparative evaluation

- Dataset: The dataset used for this study is the IMDB dataset, which consists of 50,000 movie reviews for sentiment analysis. This dataset is widely recognized in the field of natural language processing (NLP) and serves as a benchmark for sentiment classification tasks. It contains an equal number of positive and negative movie reviews, making it suitable for binary sentiment analysis.

- Systems and Baselines: In this comparative evaluation, we compare our results with the performance of several sentiment analysis systems that were summarized in this paper [1], including the following:

  RNTN (Recursive Neural Tensor Network): A deep learning model designed for sentiment analysis tasks that can capture complex relationships within the text.

  NB (Naive Bayes): A classic machine learning algorithm that is often used as a baseline for text classification tasks, including sentiment analysis.

  SVM (Support Vector Machine): Another widely-used baseline algorithm for text classification tasks.

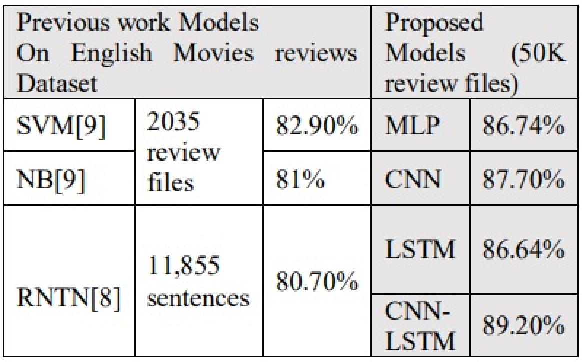
  CNN-LSTM (Convolutional Neural Network - Long Short-Term Memory): A hybrid model that combines convolutional layers for feature extraction and LSTM layers for sequential analysis.

  LSTM (Long Short-Term Memory): A recurrent neural network architecture known for its ability to capture sequential dependencies.

  CNN (Convolutional Neural Network): A deep learning architecture originally designed for image analysis but adapted for text classification tasks.

  MLP (Multi-Layer Perceptron): A feedforward neural network with multiple hidden layers used for text classification. Figure 1 shows the results of other systems:

- Measures and Evaluation Protocol: The primary measure used for evaluating the performance of these systems is accuracy. This comparative evaluation aims to provide insights into the effectiveness of various sentiment analysis methods.



Previous work Models On English Movies reviews Dataset			Proposed Models (50K review files)	
SVM[9]	2035 review files	82.90%	MLP	86.74%
NB[9]		81%	CNN	87.70%
RNTN[8]	11,855 sentences	80.70%	LSTM	86.64%
			CNN-LSTM	89.20%

Figure 1: Performance of other systems in the literature

6.1 Results

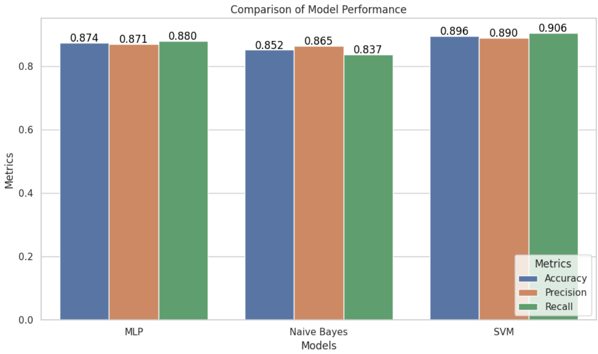


Figure 2: Performance on Tf-IDF feature

The above bar chart illustrates the results of MLP, Naive Bayes, and SVM algorithms using TF-IDF vectors as features.

Naive Bayes (NB):

Metric	Value
Accuracy	0.852
Precision	0.865
Recall	0.837

Table 2: Naive Bayes Performance

Naive Bayes is a probabilistic classifier that is known for its simplicity and speed. However, it makes the ”naive” assumption that features are independent, which may not hold in real-world text data. The test accuracy of 0.852 indicates that it correctly predicted the sentiment of approximately 85.2% of the test data. Precision of 0.865 means that when it predicted a positive sentiment, it was correct about 86.5% of the time. Recall of 0.837 suggests that it identified about 83.7% of actual positive sentiments in the dataset.

Multilayer Perceptron (MLP):

Metric	Value
Accuracy	0.874
Precision	0.871
Recall	0.880

Table 3: MLP Performance

MLP is a type of neural network that can capture complex relationships in data. It is known for its ability to model intricate patterns. The higher test accuracy of 0.874 suggests that the MLP model performed slightly better than Naive Bayes in terms of overall sentiment prediction. The precision of 0.871 indicates that it had a slightly lower false positive rate compared to Naive Bayes, meaning it was a bit more accurate when predicting positive sentiments. Recall of 0.880 implies that it identified a higher proportion of actual positive sentiments. Support Vector Machine (SVM):

Metric	Value
Accuracy	0.896
Precision	0.890
Recall	0.906

Table 4: Performance Metrics

SVM is a powerful linear classifier that can also be used for text classification when combined with appropriate text representations. The highest test accuracy of 0.896 indicates that SVM outperformed both Naive Bayes and MLP in terms of overall sentiment prediction accuracy and outperformed CNN and LSTM as well. The precision of 0.890 suggests that SVM had a low false positive rate, indicating that it was accurate when predicting positive sentiments. The high recall of 0.906 indicates that SVM was very effective at identifying actual positive sentiments in the dataset.

2- Results of the BERT-based model for sequence classification: It achieved a 83.53% of accuracy with the following settings: learning-rate=3e-5, epsilon=1e-08, and clipnorm=1.0 , knowing that I wasn’t able to train it enough due to computational resources limitations of Google Colab.

6.2 Discussion

Limits: One of the significant limitations of BERT-based models is their high computational demands. Training a large BERT model from scratch or fine-tuning it on a specific task can require substantial GPU resources and memory.

This limitation can be a barrier for individuals or organizations with limited access to powerful hardware. in Naive Bayes: The ”naive” assumption that features (words in this case) are conditionally independent can be overly simplistic for text data. In reality, word dependencies and correlations exist, and NB may not capture them effectively. Sensitivity to Out-of-Vocabulary Words: NB models can struggle when encountering words or phrases not present in the training data. Lack of Interpretability: Deep MLPs are often treated as ”black-box” models, making it challenging to interpret why a particular prediction was made. This lack of transparency can be a limitation, especially in applications where interpretability is crucial.

7 Conclusions

In this work, I conducted a comprehensive study of sentiment analysis using different machine learning models. I employed three traditional models—Naive Bayes (NB), Support Vector Machines (SVM), and Multi-Layer Perceptrons (MLP)—with TF-IDF vectorization. Additionally, I explored a state-of-the-art approach by utilizing a BERT-based text classifier for sentiment analysis.

The results of the study revealed that the BERT-based model achieved an accuracy of 83.53% with the specified hyperparameters, despite being constrained by the limited resources of the Google Colab environment. This demonstrated the potential of deep learning models in capturing complex sentiment patterns in text data.

Future work:

Hyperparameter Tuning: Fine-tuning hyperparameters for both traditional models and deep learning models could potentially yield better performance, like optimizing regularization parameters, network architectures for MLP, learning rate, and batch sizes.

Ensemble Methods: Ensemble techniques, such as stacking or blending, could be explored to combine the strengths of different models for improved sentiment analysis.

References

[1] Nehal Mohamed Ali, Marwa Mostafa Abd El Hamid, and Aliaa Youssif. Sentiment analysis for movies reviews dataset using deep learning models. *International Journal of Data Mining & Knowledge Management Process (IJDKP)*, 9(2/3), Jun 2019.

[2] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. Bert: Pre-training of deep bidirectional transformers for language understanding, 2019.

[3] Z. Guan, L. Chen, W. Zhao, Y. Zheng, S. Tan, and D. Cai. Weakly-supervised deep learning for customer review sentiment classification. In *Proceedings of the International Joint Conference on Artificial Intelligence (IJCAI)*, 2016.

[4] Jeremy Howard and Sebastian Ruder. Universal language model fine-tuning for text classification, 2018.

[5] B. Ay Karakuş, M. Talo, İ. R. Hallaç, and G. Aydin. Evaluating deep learning models for sentiment classification. *Concurrency and Computation: Practice and Experience*, 30(21):E4783, November 2018.

[6] Andrew L. Maas, Raymond E. Daly, Peter T. Pham, Dan Huang, Andrew Y. Ng, and Christopher Potts. Learning word vectors for sentiment analysis. In *The 49th Annual Meeting of the Association for Computational Linguistics (ACL 2011)*, 2011.

[7] Julian McAuley and Jure Leskovec. Hidden factors and hidden topics: Understanding rating dimensions with review text. In *RecSys*, 2013.

[8] A. Y. Ng and C. P. Richard Socher. Recursive deep models for semantic compositionality over a sentiment treebank. *PLOS ONE*, 8(11):e80455, 2013.

[9] Bo Pang, Lillian Lee, and Shivakumar Vaithyanathan. Thumbs up? sentiment classification using machine learning techniques. In *Proceedings of the 2002 Conference on Empirical Methods in Natural Language Processing (EMNLP 2002)*, pages 79–86. Association for Computational Linguistics, July 2002.