Sentiment Analysis on IMDB Movie Reviews

Alan Wong (ajw2252), Anna Micros (am6529), Emma Corbett (ec3745),
Nuneke Kwetey (nfk2108), Xiqian Yuan (xy2655)

Introduction

IMDb is an online platform where users can provide ratings and reviews for various forms of visual entertainment, particularly movies. Our analysis aims to examine the connection between the review sentiment and the textual content, with the ultimate goal of predicting binary sentiment ratings from the text. We aim to develop an optimal model capable of generating highly accurate sentiment predictions and offering insights into the factors that influence these emotions in the reviews. This will be achieved by employing and comparing a variety of machine learning and language models, as detailed below. By identifying the underlying emotions in movie reviews, we aim to tackle the challenge of translating subjective judgment into quantitative metrics.

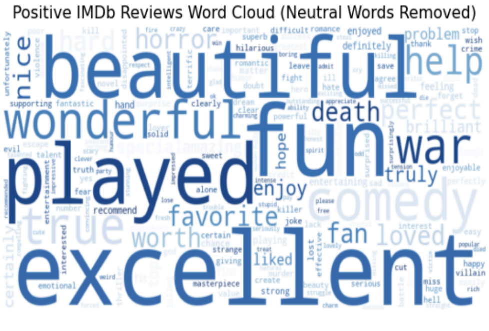
Data Processing

The dataset contains 50,000 balanced movie reviews, evenly split between positive and negative sentiments. We cleaned the text by removing special characters, URLs, and HTML tags, filtering out stopwords with NLTK, and converting text to lowercase for consistency. Tokenization split reviews into words, while unigrams, bigrams, and trigrams were generated to capture word and phrase patterns. START and STOP markers were added to structure the data, ensuring it was clean and ready for analysis. These steps reduced noise and highlighted linguistic elements relevant to sentiment.

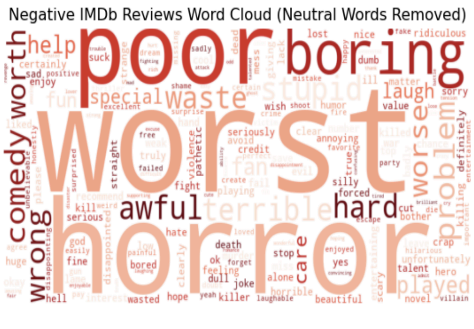
Exploratory Data Analysis

With the dataset preprocessed, we explored its structure and patterns using textual and visual analyses. This phase aimed to uncover insights into word usage, sentiment trends, and linguistic characteristics in the reviews.

Word clouds revealed frequent terms, with positive reviews using words like "excellent" and "fun," while negative reviews featured "worst" and "poor," highlighting clear sentiment distinctions.



Caption



Caption

Methodology

We employed four embedding techniques - Bag of words (BOW), TF-IDF, GloVe, and BERT - to convert textual reviews into numerical formats of analysis. These embeddings were then used as inputs for a range of models, including Logistic Regression, K-Nearest Neighbors (KNN), random forest, and deep learning approaches like BERT transformer and deep neural network. The subsequent sections detail the implementation, evaluation metrics, and performance of each model.

Logistic Regression

A logistic regression model lends itself well to the binary nature of this classification problem. Four models were trained, one for each word embedding type. Grid search was used to tune the optimal hyperparameters for each model, focusing on the penalty, C, and solver parameters. Interestingly enough, the L2 or Ridge penalty was chosen as most optimal for all embedding types out of the search space of L1, L2, elastic net, or no regularization penalty. This may indicate that less sparse solutions are more optimal for sentiment analysis as it allows all/most words to contribute to the prediction, rather than a select few as would be the case with Lasso regression.

The TF-IDF model performed best with C = 10 with search space of C = 0.01, 0.1, 1.0, 10, 100, and with solver as liblinear in a search space of lbfgs, liblinear, and saga. As for BOW, BERT, and GloVe, all had C = 0.1 and solver as the default lbfgs. With these hyperparameters, we see that TF-IDF performed the best on the metric of accuracy (89.83% accuracy, AUC 0.96), followed by BOW (89.4% accuracy, AUC 0.96), BERT (82.85% accuracy, AUC 0.91), and finally GloVe (72.2% accuracy, AUC 0.79).

K-Nearest Neighbors

Four K-Nearest Neighbors (KNN) classifiers, one for each embedding type, were evaluated for their ability to leverage word embeddings to identify proximity-based patterns. Grid search optimized hyperparameters, including the number of neighbors (k, odd values from 1 to 21), the distance metric (Euclidean, Manhattan, and Minkowski), and the weighting scheme (uniform or distance-based). Distance-based weighting consistently outperformed uniform weighting, and the Euclidean distance metric was optimal for all models except BERT, which preferred the Manhattan metric. The maximum k value of 21 was consistently optimal across all models.

The TF-IDF model delivered the best performance among the embeddings, with an accuracy of 79.92% and an AUC of 0.88. The BERT model followed closely, achieving 74.91% accuracy and an AUC of 0.85. The BoW model and the GloVe model performed less effectively, with the BoW model achieving 64.92% accuracy and an AUC of 0.72, and the GloVe model struggling significantly with 54.66% accuracy and an AUC of 0.57.

Random Forest

A Random Forest classifier was also employed to evaluate the effectiveness of the various text embedding methods. Leveraging its ensemble-based learning allowed us to capture complex relationships in the data. Four models were trained, one for each embedding method. Hyperparameter tuning was conducted via grid search, focusing on the number of estimators (50, 100, 150), the maximum depth of trees (ranging from 10 to 50 in increments of 10), and the minimum samples required for a split (2, 5, 10).

The results demonstrated that the TF-IDF embedding method delivered the highest performance, achieving an accuracy of 84.02%, closely followed by BERT, which achieved 83.62% accuracy. The Bag of Words model showed moderate performance, achieving 69.44% accuracy, while the GloVe embedding method underperformed relative to the others, with an accuracy of 62.92%.

Pre-trained BERT Transformer Model

We also used a pre-trained BERT transformer model to perform sentiment analysis on the IMDb dataset. BERT (Bidirectional Encoder Representations from Transformers) is a deep learning model that processes input text bidirectionally, allowing it to capture contextual relationships between words in a sentence more effectively than traditional unidirectional models. Unlike simpler embeddings, BERT incorporates both the left and right context of a word to generate a dynamic representation that adapts to its use in different contexts.

For this task, the pretrained BERT model was fine-tuned on the sentiment analysis dataset.The model achieved an impressive AUC score of 0.97 and an accuracy of 91.74%, demonstrating its effectiveness in distinguishing positive and negative sentiments. The best performance was obtained with the following hyperparameters: 2 epochs, a learning rate of 0.00002, and a batch size of 16. The optimizer used was Adam, which facilitated efficient training. The model relied solely on BERT embeddings, leveraging the powerful contextual representation capabilities of the transformer architecture.

Deep Neural Network

The deep neural network (DNN) used for this analysis comprised an input layer, two hidden layers with ReLU activation, dropout layers to prevent overfitting, and a sigmoid output layer for binary classification. Hyperparameters, including the number of neurons, learning rate, and dropout rate, were optimized through random search. The best performance for the DNN was achieved with TF-IDF embeddings, using 64 neurons in the first hidden layer, 32 neurons in the second, 16 neurons in the third, a dropout rate of 0.3, and a learning rate of 0.001. This configuration yielded an accuracy of 87.23%, precision/recall of 0.87, and an AUC of 0.95. BoW closely followed with 85.45% accuracy and an AUC of 0.92. BERT achieved moderate performance (82.05% accuracy, AUC 0.91), while GloVe significantly underperformed (54.64% accuracy, AUC 0.55).

Result

We compared the accuracy of four machine learning models, across the four text embedding techniques. The results are summarized in the table below.

	BERT	GloVe	BOW	TF-IDF
KNN	74.91%	54.66%	64.92%	79.92%
Random Forest	83.62%	62.92%	69.44%	84.02%
Logistic Regression	82.85%	72.2%	89.4%	89.83%
BERT Transformer	91.74%	-	-	-
Deep Neural Network	82.05%	54.64%	85.45%	87.23%

Table 1. Accuracy comparison across models and embedding methods. See model section for implementation details.

Conclusion & Future Improvements

TF-IDF consistently outperformed other embeddings across models, demonstrating its effectiveness in capturing dataset-specific patterns. Pre-trained BERT showed strong performance, achieving the best overall accuracy of 91.74% with an AUC of 0.97. BoW delivered competitive results, while GloVe consistently underperformed, likely due to its static nature and inability to adapt to the nuanced sentiment context in movie reviews. These results highlight the advantages of contextual or dataset-specific embeddings for sentiment analysis tasks.

Future improvements include experimenting with advanced transformer models like XLNet or GPT to compare with BERT, using data augmentation techniques like synonym replacement and paraphrasing to enhance generalization, and integrating ensemble methods to boost robustness and accuracy. These enhancements could refine sentiment analysis models and expand their applicability. Overall, this analysis underscores the critical role of embedding techniques and model selection in sentiment analysis, laying a foundation for further advancements in accurately interpreting and predicting textual sentiments.