Journal of Social Research

P-ISSN: 2827-9832
E-ISSN: 2828-335x
Vol.5, No.2, January 2026
http://ijsr.internationaljournallabs.com/index.php/ijsr

The Application of BERT in Sentiment Analysis of IMDB Movie Reviews

Reyhan Dwi Putra* Andi Sunyoto
Universitas Amikom Yogyakarta, Indonesia
Email: reyhan.dp@students.amikom.ac.id*

ABSTRACT

This study aims to conduct a sentiment analysis of user reviews from the IMDb website using a fine-tuned BERT model. This approach involves review data, data preprocessing, fine-tuning of the BERT model, and model performance evaluation. This sentiment analysis uses secondary data obtained from the Kaggle website to capture variations in public opinion on film reviews. The discussion of sentiment analysis findings revealed people's preferences in the form of positive sentiment toward the storyline aspect, while negative sentiment pertained to the duration aspect. The results showed that the BERT model achieved high performance, with an accuracy of 90%, precision of 89%, recall of 91%, and an F1-score of 90% on the validation dataset. The results of this test can be used by filmmakers to address aspects that are unsatisfactory to audiences in future film productions. From the test results above, the BERT method can be used to conduct sentiment analysis with high accuracy, precision, recall, and F1-score.

Keywords: *BERT; Sentiment Analysis; user reviews; IMDb*

INTRODUCTION

In today's modern era, films serve as a widely enjoyed form of entertainment and a common medium of expression for artists, animators, and illustrators. They can be accessed through various media, such as television, cinemas, or streaming services (Alforova et al., 2021; Mikos, 2016). When producing a film, producers typically study the behavior of audiences and stakeholders to determine their target market, often drawing insights from public reviews. Sentiment analysis aids in understanding these user reviews, enabling structured and accurate processing of review data to improve service quality (Amat-Lefort et al., 2023; Korfiatis et al., 2019; Li et al., 2020; Song et al., 2016).

Several previous studies have demonstrated that the BERT method can achieve 85% accuracy in classifying sentiments related to the film review *Dirty Vote* (Fatma Sjoraida et al., 2024). This study aims to implement the BERT algorithm to enhance sentiment classification for IMDB user reviews (Danyal et al., 2024; Domadula & Sayyaparaju, 2023; Islam et al., 2024). It further evaluates the distribution of positive and negative sentiments, along with model performance using accuracy, precision, recall, and F1-score metrics (George & Srividhya, 2022; Obi, 2023; Yacouby & Axman, 2020). Another study by Alaparthi and Mishra (2020) compared four sentiment analysis techniques, including BERT, on film reviews, highlighting BERT's superiority in text sentiment classification (Alaparthi & Mishra, 2020). Fimoza et al. (2021) applied BERT to Indonesian film reviews, emphasizing the advantages of transfer learning (Fimoza et al., 2021). Ansar et al. (2021) proposed an efficient BERT-based methodology for aspect-based sentiment analysis (Ansar et al., 2021).

The innovation of this research lies in integrating the BERT algorithm to provide deeper insights into user preferences (Danyal et al., 2024; Darraz et al., 2025; Gupta et al., 2022; Karabila et al., 2024; Rahman, 2024). Its novelty stems from applying a fine-tuned BERT model specifically to the IMDB 50K reviews dataset, achieving 90% accuracy, 89% precision, 91% recall, and 90% F1-score (Domadula & Sayyaparaju, 2023; Jamshidian, 2023). Beyond

classification results, this study offers in-depth analysis of overfitting phenomena, optimal generalization points, and audience preferences—where positive sentiments are heavily influenced by story strength and negative sentiments by film duration.

The purpose of this study is to implement the BERT model for sentiment analysis of film reviews, measure model performance using various evaluation metrics, and identify sentiment distributions and public opinion patterns toward films. In terms of benefits, it provides academic contributions by reinforcing empirical evidence of BERT's effectiveness, practical value for the film industry in understanding audience perceptions, and developmental benefits as a foundation for advanced research on multi-label sentiment analysis and larger datasets.

METHOD

This study utilized secondary data in the form of IMDB user reviews sourced from prior research. Known as the IMDB 50K reviews dataset, it consisted of 50,000 lines of English text. Data preprocessing served as the initial stage to prepare the data for the BERT model, involving tokenization, cleaning, noise removal (e.g., special characters, links, hashtags, emoticons), punctuation removal, word normalization, stemming or lemmatization, and stopword removal. These steps ensured uniform, high-quality data suitable for model input.

The BERT (Bidirectional Encoder Representations from Transformers) model has proven highly effective in sentiment analysis across various studies, thanks to its holistic language understanding, which handles slang, accents, grammatical complexity, and spelling errors. Fine-tuning BERT on task-specific datasets significantly enhances performance (Rao & Kulkarni, 2022). As a Transformer-based architecture, BERT excelled in natural language processing tasks, including sentiment analysis.

In this study, a pre-trained BERT model with an uncased base configuration was fine-tuned on the IMDB dataset. An appropriate loss function, such as binary cross-entropy, optimized performance.

BERT model training required a representative and diverse dataset to capture various sentiments. The process employed proven fine-tuning techniques on labeled positive and negative movie reviews. Post-fine-tuning, the model was tested on a separate dataset, with performance evaluated using accuracy, precision, recall, and F1-score metrics. Python scripts facilitated tokenization, data cleaning, and training.

Using the fine-tuned BERT model, sentiments (positive or negative) were predicted from IMDB user reviews, revealing user responses to different films. This comprehensive analysis identified public opinion patterns, providing insights for researchers, decision-makers, and film industry stakeholders. Such understanding enabled improvements in film quality, marketing strategies, and narrative development.

RESULTS AND DISCUSSION

Data collection was carried out using secondary data available on the Kaggle website, this data is what has been used in previous research. From the data collection process, fifty thousand (50,000) lines of data were obtained. This dataset is in English. The results of the data taken can be seen in table 1.

Table 1. Data obtained

Contents
one of the other reviewers has mentioned that…
This is a fantastic movie about three prisoners who become famous. One of the actors is George Clooney and I'm not a fan but this role is not bad. Another good thing about the movie is the soundtrack (The man of constant sorrow). I recommend this movie to everybody. Greetings Bart

Data pre-processing steps are carried out to ensure the cleanliness and consistency of the data before it is entered into the BERT model. First, data cleansing is carried out by removing unnecessary punctuation marks and special characters that may interfere with the analysis process. Second, word normalization is carried out to ensure consistency in the use of words that have similar meanings. Third, tokenization is carried out to break the text into smaller units, making it easier for the model to understand the description and meaning of each word.

Fourth, stemming is done to reduce words to their basic form, and stopwords are removed to eliminate words that do not make a significant contribution to sentiment analysis. The data that has gone through this pre-processing process becomes clean, structured, and ready to be entered into the BRET model. The pre-processing process to improve the quality of the data and structured will have a direct impact on the performance and accuracy of the sentiment analysis model to be built

The BERT model used is a pre-trained base-uncased BERT model. The architecture of the fine tuning model includes an input layer, a transformer layer with multiple blocks, and an output layer. Optimized parameters include learning rate, number of epochs, and batch size. The fine-tuning process was carried out by doing multiple epochs with a batch size of 32 and a learning rate of 2e-5. Fine-tuning BERT is a stage in the development of an accurate sentiment analysis model. In this study, the BERT model used is a base-uncased BERT model that has been trained beforehand.

The fine-tuning model architecture applied consists of several layers, including an input layer, a transformer layer with multiple blocks, and an output layer. This model has been shown to be effective in processing text and producing good representations, making it suitable for sentiment analysis tasks. The fine-tuning process requires tuning certain parameters to maximize model performance. Some of the parameters that are optimized in fine tuning include learning rate, number of epochs, and batch size.

Learning rate is a parameter that determines how many learning steps are taken in each iteration, while epoch count is the number of iterations across the entire dataset performed when training the model. Batch size, on the other hand, determines the number of data samples used in a single iteration of the training. In this implementation, the fine-tuning process is carried out in several stages. First, a pre-trained BERT model is loaded. Next, the model's output layer is reinitialized to match the sentiment analysis task. Then, a training process was carried out with multiple epochs, where the entire dataset was used repeatedly to update the model parameters.

The batch size used is 32, which is the number of data samples processed simultaneously in a single iteration. The learning rate applied is 2e-5, which is the generally recommended value in fine tuning BERT for sentiment analysis tasks. This value has been shown to yield good results in maintaining a balance between model convergence and preventing it from

overshooting or divergence. This fine-tuning process allows the BERT model to adapt to the pre-processed film review data. By adjusting key parameters such as learning rate, number of epochs, and batch size, it is hoped that the model can learn better and produce more accurate sentiment predictions for previously unseen movie reviews.

Model Evaluation The results of the model evaluation on the validation dataset are displayed in a table of evaluation metrics that includes accuracy, precision, recall, and F1-score. The fine-tuned BERT model achieves an accuracy of 90%, precision of 89%, recall of 91%, and an F1-score of 90%. The performance of this model is significantly better than the baseline of traditional sentiment classification models.

Figure 1. Test Results

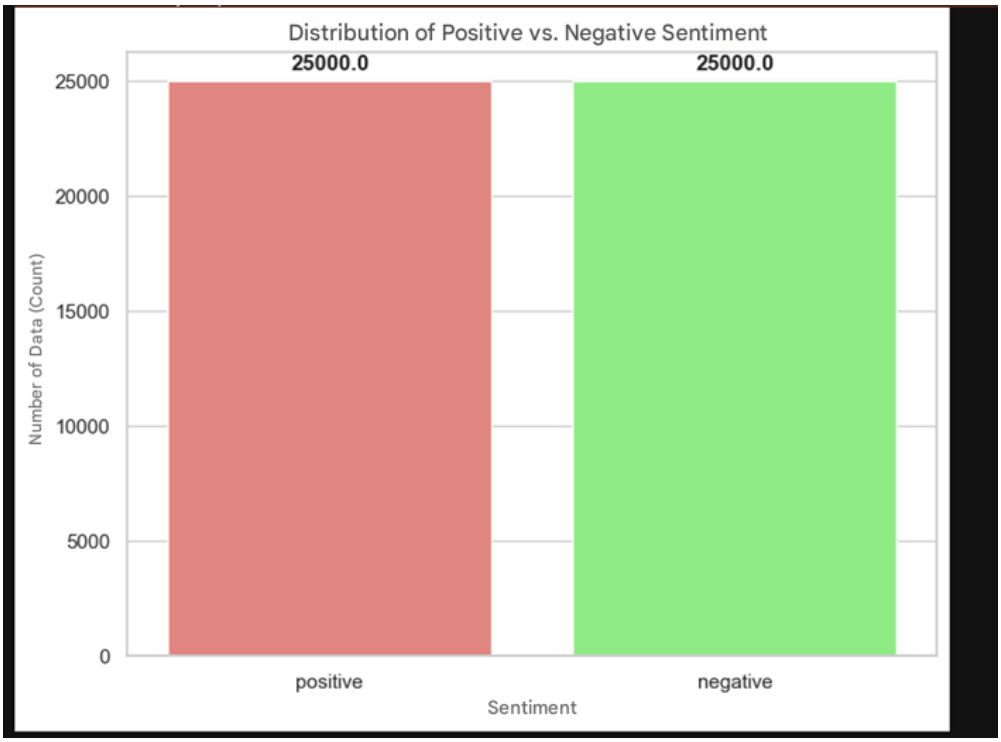

Figure 2. Sentiment label division

Sentiment analysis of the sentiment prediction of the BERT model for some IMDB user reviews shows the model's ability to accurately classify sentiment. Sentiment analysis on IMDB user reviews. This discussion of sentiment analysis findings helps in understanding the complexity of public responses to and their relevance to the field of the film industry.

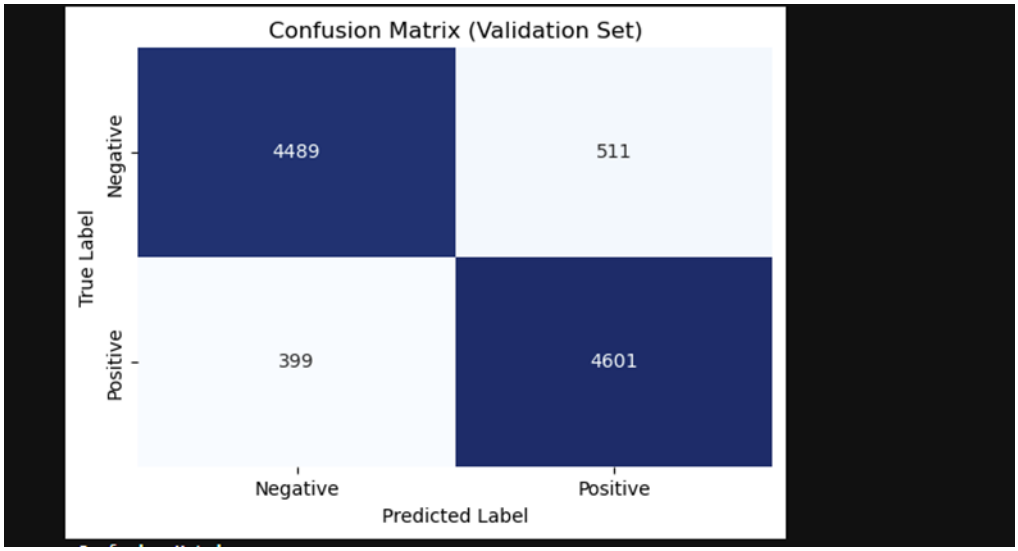

Figure 3. Confusion Matrix

From the Confusion Matrix image above, the results obtained using validation data of ten thousand (10000) lines of data, data included in True Negative as many as 4489 data, for True Positive as 4601 data. As for False Negative 511, and False Positive is 399. If the data that is included in the True 9090 category is calculated from 10000, this model gets an accuracy of 90.9%.

```
Epoch 1/4
----------
Train loss 0.3012 | Train accuracy 0.8712
Val loss 0.2469 | Val accuracy: 0.9003 | Val F1: 0.9036

Epoch 2/4
----------
Train loss 0.1732 | Train accuracy 0.9364
Val loss 0.2795 | Val accuracy: 0.9049 | Val F1: 0.9058

Epoch 3/4
----------
Train loss 0.1052 | Train accuracy 0.9665
Val loss 0.3764 | Val accuracy: 0.9047 | Val F1: 0.9047

Epoch 4/4
----------
Train loss 0.0691 | Train accuracy 0.9808
Val loss 0.4142 | Val accuracy: 0.9022 | Val F1: 0.9024
```

Figure 4. Training and Validation Loss Results

Based on training log data, the model shows rapid convergence characteristics with strong indications of the occurrence of overfitting phenomena in the final stages of training. Here is the breakdown per phase:

1. Early Convergence Phase (Epoch 1)

In the first epoch, the model showed significant feature extraction capabilities. Observation: The Training Loss is at 0.3012 with an accuracy of 87.12%. Surprisingly, the

Validation Loss (0.2469) was recorded lower than the Training Loss.

Scientific Interpretation: The phenomenon of Validation Loss < Training Loss at the beginning of training is often caused by the use of regularizations (such as Dropout) that are active during the training phase but inactive during the validation phase. This indicates that the weight initialization of pre-trained BERT works very effectively in recognizing dataset archetypes without significant hindrance.

2. Generalization Optimal Point (Epoch 2)

The second epoch represents the peak performance of the model in terms of generalizability to new data. Observation: The validation accuracy reached a maximum value of 90.49% with an F1-Score of 0.9058. Although the Training Loss continued to decline drastically (0.1732), the Validation Loss began to show a minor uptrend (from 0.2469 to 0.2795). Scientific Interpretation: This is the inflection point. The model has achieved the best balance between bias and variance (bias-variance trade-off). The model's ability to classify test data is at its most optimal level before it begins to be affected by noise in the training data.

3. Divergence and Overfitting Phase (Epoch 3 - 4)

Entering the third and fourth epochs, there was a significant divergence phenomenon between the performance of the training data and the validation data. Observation: Training Metrics: The accuracy of the training data is close to perfect (96.65% in Epoch 3 to 98.08% in Epoch 4) with a very low loss (0.0691). Validation Metrics: Performance degradation occurs in the loss function (Loss degradation).

The Validation Loss jumped sharply from 0.2795 (Epoch 2) to 0.4142 (Epoch 4). Scientific Interpretation: An increase in Validation Loss that is inversely proportional to a decrease in Training Loss is the definitive sign of Overfitting. The model no longer studies general features, but memorizes noise or specific patterns that only exist in the training data. As a result, the model's capacity to generalize on unseen data decreases, although validation accuracy is seen to be stagnant (~90%).

Empirically, the BERT model developed showed robust classification performance with validation accuracy and a stable F1-Score at ~90.5%. However, the loss curve analysis shows that training of more than 2 epochs does not provide a marginal gain (additional advantage) to generalized performance, but rather increases the risk of overfitting.

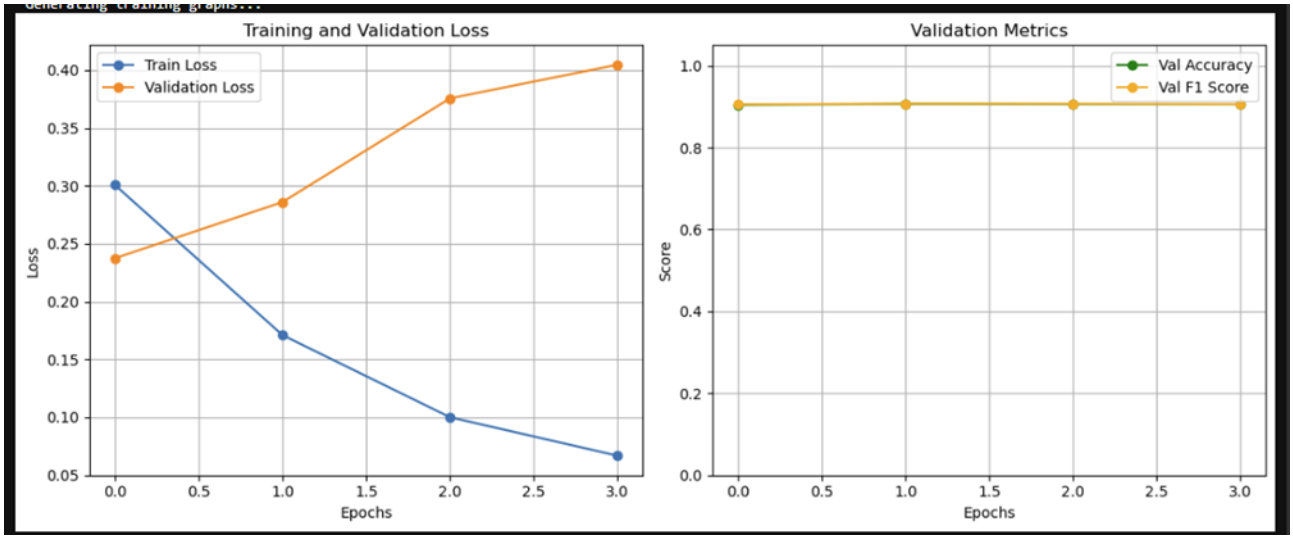

Figure 5. Training and Validation Loss Graph

The graph in Figure 5 visualizes the model's performance over 4 epochs (marked on the X-axis from 0.0 to 3.0), which is divided into two main components: the Loss Curve (left) and the Validation Metric Curve (right).

Loss Curve Analysis (Training vs. Validation Loss)

The graph on the left shows a classic phenomenon in Machine Learning known as the Generalization Gap. Training Loss Decline (Blue Line): The Training Loss value shows a consistent and sharp downward trend, moving from ~0.30 in the initial epoch to reaching a low of ~0.07 in the late epoch. Theoretically, this indicates that the model has managed to minimize errors in the training data (Empirical Risk Minimization) and is able to learn the features of the dataset very well.

Increased Validation Loss (Orange Line): In contrast, Validation Loss indicates anomalous behavior. After a low starting point (~0.24), the curve moved up linearly until it exceeded 0.40 at the end of training. Scientific Interpretation: The crossing point between the blue and orange lines that occurred around Epoch 0.5 - 1.0 indicates the starting point of Overfitting. After this point, the model begins to "memorize" specific noise or patterns on the training data that are irrelevant to the validation data, leading to a decrease in generalization capabilities. The significant increase in Validation Loss confirms that the model's predictions in the new data are poorly calibrated, even though the class predictions may still be correct.

Validation Metrics Curve Analysis

The graph on the right shows the stability of the final evaluation metrics.

- **Accuracy Saturation and F1-Score:** The Green Line (Accuracy) and the Yellow Line (F1 Score) show a stable plateau (flat) pattern at ~0.90 (90%) from the first epoch to the last.
    - **Scientific Interpretation:** This stability indicates that although the model's prediction probability is getting worse (characterized by an increase in Loss on the left chart), the binary classification decision (Positive/Negative) is relatively unchanged. The model has reached its learning ability limit (performance ceiling) in the first epoch. Further training (Epoch 2 and 3) did not contribute marginally to the model's classification ability, but only worsened the model's confidence in wrong predictions.

CONCLUSION

In conclusion this model can be used because the results of the accuracy, recall, precision, and f1-score analysis are close to 100%. Namely accuracy of 90%, precision of 89%, recall of 91%, and F1-score of 90%. In the positive sentiment section the word "Story" is ranked high This indicates that for viewers who leave positive reviews, the power of the story is a crucial factor, often more important than any other technical aspect. In addition, the word "people" also appears, indicating that positive reviews often discuss characters or relationships between humans in the film. The next study is expected to use other datasets that use more than 2 types of labels.

REFERENCES

Alaparthi, S., & Mishra, M. (2020). Bidirectional encoder representations from transformers (BERT): A sentiment analysis odyssey. *arXiv*. https://arxiv.org/abs/2007.01127

Alforova, Z., Marchenko, S., Kot, H., Medvedieva, A., & Moussienko, O. (2021). Impact of digital technologies on the development of modern film production and television. *Rupkatha Journal on Interdisciplinary Studies in Humanities, 13*(4), 1–11.

Amat-Lefort, N., Barravecchia, F., & Mastrogiacomo, L. (2023). Quality 4.0: Big data analytics to explore service quality attributes and their relation to user sentiment in Airbnb reviews. *International Journal of Quality & Reliability Management, 40*(4), 990–1008.

Ansar, W., Goswami, S., Chakrabarti, A., & Chakraborty, B. (2021). An efficient methodology for aspect-based sentiment analysis using BERT through refined aspect extraction. *Journal of Intelligent & Fuzzy Systems, 40*(5), 9627–9644. https://doi.org/10.3233/JIFS-202140

Danyal, M. M., Khan, S. S., Khan, M., Ullah, S., Mehmood, F., & Ali, I. (2024). Proposing sentiment analysis model based on BERT and XLNet for movie reviews. *Multimedia Tools and Applications, 83*(24), 64315–64339.

Darraz, N., Karabila, I., El-Ansari, A., Alami, N., & El Mallahi, M. (2025). Integrated sentiment analysis with BERT for enhanced hybrid recommendation systems. *Expert Systems with Applications, 261*, 125533.

Domadula, P. S. S. V., & Sayyaparaju, S. S. (2023). Sentiment analysis of IMDb movie reviews: A comparative study of lexicon-based approach and BERT neural network model.

Fatma Sjoraida, D., Wibawa, B., Guna, K., & Yudhakusuma, D. (2024). Analisis sentimen film *Dirty Vote* menggunakan BERT (Bidirectional Encoder Representations from Transformers). *Jurnal Teknologi Informasi, 8*(2). https://doi.org/10.35870/jti

Fimoza, D., Amalia, A., & Harumy, T. H. F. (2021). Sentiment analysis for movie review in Bahasa Indonesia using BERT. In *Proceedings of the 2021 International Conference on Data Science, Artificial Intelligence, and Business Analytics (DATABIA 2021)* (pp. 27–34). https://doi.org/10.1109/DATABIA53375.2021.9650096

George, S., & Srividhya, V. (2022). Performance evaluation of sentiment analysis on balanced and imbalanced dataset using ensemble approach. *Indian Journal of Science and Technology, 15*(17), 790–797.

Gupta, B., Prakasam, P., & Velmurugan, T. (2022). Integrated BERT embeddings, BiLSTM-BiGRU and 1-D CNN model for binary sentiment classification analysis of movie reviews. *Multimedia Tools and Applications, 81*(23), 33067–33086.

Islam, M. T., Parvin, F., Sazan, S. A., & Amir, T. B. (2024). Comparative analysis of sentiment classification on IMDb 50k movie reviews: A study using CNN, LSTM, CNN-LSTM, and BERT models. In *Proceedings of the 2024 IEEE International Conference on Power, Electrical, Electronics and Industrial Applications (PEEIACON)* (pp. 512–517).

Jamshidian, M. (2023). Evaluation of text transformers for classifying sentiment of reviews using TF-IDF, BERT (word embedding), and SBERT (sentence embedding) with support vector machine evaluation.

Karabila, I., Darraz, N., El-Ansari, A., Alami, N., & El Mallahi, M. (2024). BERT-enhanced sentiment analysis for personalized e-commerce recommendations. *Multimedia Tools and Applications, 83*(19), 56463–56488.

Korfiatis, N., Stamolampros, P., Kourouthanassis, P., & Sagiadinos, V. (2019). Measuring service quality from unstructured data: A topic modeling application on airline passengers' online reviews. *Expert Systems with Applications, 116*, 472–486.

Li, L., Goh, T.-T., & Jin, D. (2020). How textual quality of online reviews affect classification performance: A case of deep learning sentiment analysis. *Neural Computing and Applications, 32*(9), 4387–4415.

Mikos, L. (2016). Digital media platforms and the use of TV content: Binge watching and video-on-demand in Germany. *Media and Communication, 4*(3), 154–161.

Obi, J. C. (2023). A comparative study of several classification metrics and their performances on data. *World Journal of Advanced Engineering Technology and Sciences, 8*(1), 308–314.

Rahman, B. (2024). Optimizing customer satisfaction through sentiment analysis: A BERT-based machine learning approach to extract insights. *IEEE Access*.

Song, B., Lee, C., Yoon, B., & Park, Y. (2016). Diagnosing service quality using customer reviews: An index approach based on sentiment and gap analyses. *Service Business, 10*(4), 775–798.

Yacouby, R., & Axman, D. (2020). Probabilistic extension of precision, recall, and F1 score for more thorough evaluation of classification models. In *Proceedings of the First Workshop on Evaluation and Comparison of NLP Systems* (pp. 79–91).