Iğdır Üniversitesi Fen Bilimleri Enstitüsü Dergisi, 13(3), 1496-1510, 2023
Journal of the Institute of Science and Technology, 13(3), 1496-1510, 2023

ISSN: 2146-0574, eISSN: 2536-4618

Computer Engineering **DOI: 10.21597/jist.1292050**

Research Article

Received: 05.05.2023 **Accepted: 29.05.2023**

To Cite: Bayat, S. & Işık, G. (2023). Evaluating the Effectiveness of Different Machine Learning Approaches for Sentiment Classification . *Journal of the Institute of Science and Technology*, 13(3), 1496-1510.

Evaluating the Effectiveness of Different Machine Learning Approaches for Sentiment Classification

Seda BAYAT[*1], Gültekin IŞIK[2]

Highlights:

- Machine learning algorithms used for sentiment analysis on AR-P dataset
- Transformer-based distilbert performed the best
- This study can be used for future works

Keywords:

- Deep Learning
- Distilbert
- Sentiment Analysis
- Text Classification
- Transformer

ABSTRACT:

This study presents a comparison of four different machine learning algorithms for sentiment analysis on a small subset of the AR-P (Amazon Reviews - Polarity) dataset. The algorithms evaluated are multilayer perceptron (MLP), Naive Bayes, Decision Tree, and Transformer architectures. The results show that the Transformer-based DistilBERT model performed the best with an accuracy rate of 96.10%, while MLP had a better performance than the other remaining methods. Confusion matrices and ROC curves are provided to illustrate the results, and a comparison with previous studies is presented. The study concludes that the results can serve as a basis for future work, such as using larger datasets or comparing the performance of algorithms on different tasks. Overall, this study provides insights into the use of traditional machine learning and modern deep learning methods for sentiment analysis and their potential applications in real-world scenarios.

[1]Seda BAYAT (**Orcid ID:** 0000-0002-8427-9971), Iğdır University, Faculty of Engineering, Department of Mechatronics Engineering, Iğdır, Türkiye
[2]Gültekin IŞIK (**Orcid ID:** 0000-0003-3037-5586), Iğdır University, Faculty of Engineering, Department of Computer Engineering, Iğdır, Türkiye

***Corresponding Author:** Seda BAYAT, e-mail: bayatseda@gmail.com

INTRODUCTION

Natural Language Processing (NLP) has emerged as a prominent field of research in the realm of artificial intelligence, focusing on enabling computers to understand, analyze, and generate human language. One crucial task in NLP is text classification, which involves assigning predefined categories or labels to text data based on their content. Sentiment analysis, a popular form of text classification, specifically aims to determine the sentiment or emotional tone of a given text, such as positive, negative, or neutral (Gupta et al., 2021).

Sentiment analysis is a text analysis technique that involves determining the emotional or subjective tone of a given text (Wankhade et al., 2022). It aims to understand the underlying sentiment or emotional state of the text, which can be broadly classified into positive, negative, or neutral categories. However, sentiment analysis can also go beyond simple polarity detection and delve into detecting specific feelings, emotions, urgency, intentions, or other nuanced aspects of the text. For instance, sentiment analysis can be used to detect emotions such as anger, happiness, sadness, fear, or surprise in text data (Kim, 2020). This can be useful in understanding the emotional tone of customer feedback, social media posts, or product reviews, and can provide insights into customer satisfaction, brand perception, and overall sentiment towards a particular topic or product (Nandwani & Verma, 2021).

Sentiment analysis can also be applied to detect urgency in text data, such as identifying if a customer inquiry or complaint requires immediate attention or can be addressed later. This can be valuable in customer service or support scenarios where timely responses are crucial. Furthermore, sentiment analysis can be used to infer intentions from text data, such as identifying if a user is interested or not interested in a particular topic, product, or service. This can be useful in market research, customer profiling, or personalized recommendations, where understanding user intentions can drive business strategies and decision-making. Sentiment analysis holds significant importance in various domains, including marketing, social media analysis, customer feedback analysis, and reputation management, as it allows businesses and organizations to gain insights from large amounts of textual data and make informed decisions (Nandwani & Verma, 2021).

Traditional machine learning algorithms, such as Multilayer Perceptron (MLP), Naive Bayes, and Decision Tree, have been widely used for sentiment analysis. However, the recent advancements in transformer-based models, such as DistilBERT (Sanh et al., 2019), have shown remarkable performance in various NLP benchmarks, leading to increased interest in their application for sentiment analysis.

DistilBERT, a compact and computationally efficient variant of the original BERT (Bidirectional Encoder Representations from Transformers) model (Joshi et al., 2020), has gained attention due to its state-of-the-art performance on several NLP benchmarks. However, it is essential to evaluate its performance in sentiment analysis tasks and compare it with other traditional machine learning algorithms. In this paper, we present a comparative study of DistilBERT with Multilayer Perceptron (MLP), Naive Bayes, and Decision Tree for sentiment analysis of Amazon reviews.

In this paper, we focused on coarse-grained sentiment analysis, where the sentiment of a given text is broadly categorized into positive or negative sentiments. Coarse-grained sentiment analysis is a commonly used approach that provides a high-level overview of the overall emotional tone of a text, without delving into finer details of specific emotions or intentions expressed in the text. Coarse-grained sentiment analysis can be useful in various applications, such as brand monitoring, market research, customer feedback analysis, and sentiment tracking in social media. By categorizing text data into positive or negative sentiments, it provides a quick and effective way to gauge the overall sentiment

towards a particular topic or product. For our study, we utilized a subset of the Amazon Reviews – Polarity (AR-P) dataset, comprised a total of 200,000 examples, randomly selected from both positive and negative classes. The dataset was split into three subsets, with 70% used for training, 15% for validation, and the remaining data allocated for testing purposes. We preprocessed the dataset and conducted experiments using DistilBERT, MLP, Naive Bayes, and Decision Tree classifiers. We compare the performance of these models in terms of *accuracy*, *precision*, *recall*, and F_1*-score*. Our study aims to provide insights into the suitability and effectiveness of DistilBERT for sentiment analysis tasks, and its performance compared to traditional machine learning algorithms.

Related Works

Deep learning has proven to be effective in numerous domains, as evidenced by its success in various areas (Bayat & Işık, 2022; Gündüz & Işık, 2023), including sentiment analysis, where it has demonstrated promising outcomes. Sentiment analysis has become a popular research area in artificial intelligence and machine learning in recent years (Chen et al.,2020). Several techniques have been proposed for sentiment analysis, including rule-based methods, machine learning algorithms, and deep learning models (Ray & Chakrabarti, 2022).

Rule-based methods rely on manually crafted rules to identify emotions in text, while machine learning algorithms learn from a labeled dataset to classify text into positive or negative emotions. (Sudhir et al., 2021). Deep learning models such as neural networks have shown superior performance in sentiment analysis due to their ability to capture complex patterns in data (Abdi et al., 2019).

In recent years, transformer-based models have emerged as a promising approach for sentiment analysis (Al- Garadi et al., 2021). Transformer utilize self-attention mechanisms to capture contextual relationships between words in a sentence and have achieved the best performance on several benchmark datasets for sentiment analysis. For example, BERT and RoBERTa are two popular transformer-based models that achieve high accuracy in sentiment analysis tasks (Devlin et al., 2019; Liu et al., 2019). Recent studies have shown that transformer-based models achieve the best results in various NLP tasks, including sentiment analysis (Devlin et al., 2018; Liu et al., 2019). BERT is a pre-trained deep duplex transformer model with impressive performance on a variety of NLP measurements. BERT has proven highly effective in a wide variety of natural language processing tasks, including sentiment analysis (Devlin et al., 2018). BERT is built on top of RoBERTa, a robustly optimized version that achieves even better performance by taking a more comprehensive pre-training approach and fine-tuning learning hyperparameters. RoBERTa has shown superior results in sensitivity analysis and other NLP tasks (Liu et al., 2019). Further research has been conducted on BERT-based language models, exploring their applications and advancements in various fields (Delobelle et al., 2020). These studies have made significant contributions to the development and advancement of sentiment analysis methods using transformer models.

Researchers have made significant advancements in developing smaller and faster transformer models for sentiment analysis tasks. Sanh et al. (2019) introduced DistilBERT, which is a more compact, efficient, and cost-effective variant of BERT. This model retains a high level of performance while reducing its size and computational requirements (Sanh et al., 2019). Furthermore, another study proposed ALBERT, a lightweight BERT model with fewer parameters that exhibits enhanced learning efficiency compared to BERT (Lan et al., 2020). These transformative models provide valuable alternatives for sentiment analysis applications, enabling faster inference and deployment while maintaining high accuracy (Sanh et al., 2019; Lan et al., 2020). The utilization of these models can contribute to the development of efficient and effective sentiment analysis systems.

In conclusion, transformer-based models have proven to be highly effective in sentiment analysis tasks, with BERT and its variants being the most important models in recent years. More research is needed to explore the potential of these models in different domains and languages (Balahur et al., 2013; Pang & Lee, 2008; Devlin et al., 2018; Liu et al., 2019; Sanh et al., 2019; Lan et al., 2020; Sun et al., 2021).

MATERIALS AND METHODS

In this study, traditional machine learning and deep learning algorithms were used: Multi-Layer Perceptron, Naive Bayes, Decision Tree and Transformers. These 4 different architectures were chosen to examine their advantages and disadvantages as they are the most popular techniques in today's classifications.

Dataset

The Amazon Reviews for Sentiment Analysis dataset is a widely used benchmark dataset for sentiment analysis tasks in natural language processing. It contains a large collection of customer reviews from Amazon, covering multiple product categories, and includes sentiment labels indicating whether each review is positive, negative, or neutral. One commonly used version of this dataset, known as the "Amazon Reviews - Polarity" dataset, comprises reviews collected from Amazon.com, with an equal number of positive and negative reviews. The dataset is split into 3.6 million training documents and 400,000 test documents for model evaluation. Each document in the dataset contains an average of 91 words. For our study, we randomly selected a subset of 200,000 examples from the dataset, both positive and negative classes. The dataset was split into three subsets, with 70% used for training, 15% for validation, and the remaining data allocated for testing purposes.

This sampling strategy allowed us to create a balanced dataset with an equal number of positive and negative examples, ensuring that both sentiment categories were adequately represented in our analysis. By carefully curating the dataset in this manner, we aimed to obtain a representative sample that would allow us to conduct a robust and reliable analysis of sentiment using the selected dataset.

Sentiment Analysis

Sentiment analysis is a method used to analyze people's expressed emotional states (Cambria & White, 2014). It is usually applied on written or spoken texts and is performed using various techniques such as *word analysis*, *structural analysis* and *machine learning* (Gao & Wong, 2014). *Word analysis* analyzes the meaning of words in a text and the emotions they reflect (Poria et al., 2017). *Structural analysis* analyzes sentiment by examining the grammatical structure of a text (Hutto & Gilbert, 2014). *Machine learning*, on the other hand, allows an algorithm trained on a predetermined dataset to analyze sentiment (Mohammad & Bravo-Marquez, 2017). Sentiment analysis has many different uses such as customer service, social media analysis, marketing, political analysis and health analysis (Cambria & White, 2014; Gao & Wong, 2014). For example, in customer service, sentiment analysis is used to measure customer satisfaction and address customer complaints (Bollen et al., 2011). In the field of social media analytics, sentiment analysis measures the success of social media campaigns by analyzing the sentiment of posts people share on social media.

Deep Learning-Based Sentiment Analysis

Nowadays, sentiment analysis is a frequently used area of machine learning. Sentiment analysis is a subset of natural language processing (NLP) techniques and is used to detect emotional content in text data. Deep learning methods are often used to work on high-dimensional and complex datasets (Ain et al., 2017). The most popular deep learning methods for sentiment analysis are convolutional neural

networks (CNN), long short-term memory (LSTM) neural networks and Transformers (Othan et al., 2019). As a result, deep learning methods in sentiment analysis can have higher accuracy rates than traditional machine learning techniques. However, it is important to note that these methods require large amounts of data to implement and train, and the training process can be lengthy.

Multi-Layer Perceptron (MLP)

In this study, Multi-Layer Perceptron (MLP) one of the deep learning methods, is used as the first architecture. MLP is an artificial neural network with multiple hidden layers between input and output. The basic principle of its operation is to obtain the output by multiplying the input data by weights in the layers and applying an activation function (Goodfellow et al., 2016). The advantages of MLP are that it provides an effective solution for nonlinear problems, achieves high accuracy rates and is compatible with different types of data. However, it is also prone to overfitting and memorization (Goodfellow et al., 2016). The training of MLP is performed by a method called backpropagation. This method is used to update the weights and threshold values. The training data allows the network to adjust the weights and thresholds to minimize the error function.

Naive Bayes Algorithm

The second architecture used in this study is the Multinomial Naive Bayes algorithm. Naive Bayes is a probabilistic machine learning algorithm commonly used for text classification tasks, including sentiment analysis (Boyko & Boksho, 2020). It assumes that features in a text are conditionally independent, given the class label, allowing for efficient computations. The algorithm estimates the probabilities of a sample belonging to each class based on the occurrence or frequency of features in the text data, and assigns the class label with the highest probability as the predicted sentiment label (Raschka & Mirjalili, 2021). Naive Bayes is known for its simplicity and efficiency, making it suitable for large datasets and scenarios with relatively independent features, such as bag-of-words representations. Despite its simplicity, Naive Bayes has been shown to be effective in various NLP tasks, including sentiment analysis (Alexandridis et al., 2021).

Decision Tree Algorithm

The third architecture used in this study is the decision tree architecture. Decision Tree algorithm is a classification and regression method that models the decision-making process by creating a tree structure based on certain characteristics of the data (Kumar, 2022). The algorithm can be used to classify data or perform regression analysis. Basically, the algorithm creates a decision tree and generates tree branches and leaves based on certain characteristics of the data (Han & Kamber, 2011). The Decision Tree algorithm uses a so-called tree structure. This structure consists of a root node, branches and leaves. The root node represents an overall decision-making process that encompasses all the data. Branches are lines leading from the root node to lower-level nodes. Lower-level nodes enable classification or regression analysis of data according to certain characteristics. Leaves are the lowest level nodes and represent the results or classifications.

Transformer Architecture

The transformer model is a type of deep learning architecture introduced by Vaswani et al. in the paper "Attention is All You Need" in 2017. The transformer model utilizes self-attention mechanisms, where the model can weigh the importance of different words in a sentence based on their contextual relevance. This allows the transformer to capture long-range dependencies and relationships between words, making it highly effective for tasks such as language translation, text generation, and sentiment analysis. The transformer model has been further improved with various variants, including BERT,

GPT-2, and DistilBERT, which have achieved state-of-the-art results on many NLP benchmarks. The transformer model's ability to capture complex contextual information and its versatility for various NLP tasks have made it a groundbreaking innovation in the field of deep learning for natural language processing (Al-Qurishi, Khalid, & Souissi, 2021). The architecture of the transformer model is illustrated in Figure 1.

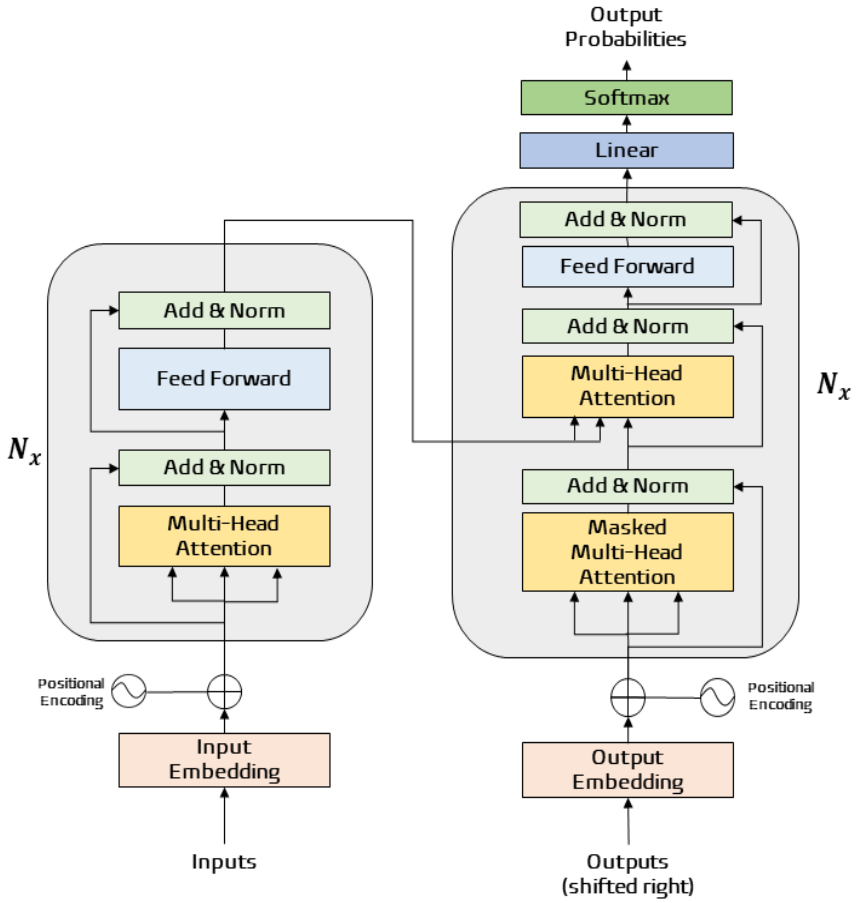

Figure 1. Transformer architecture

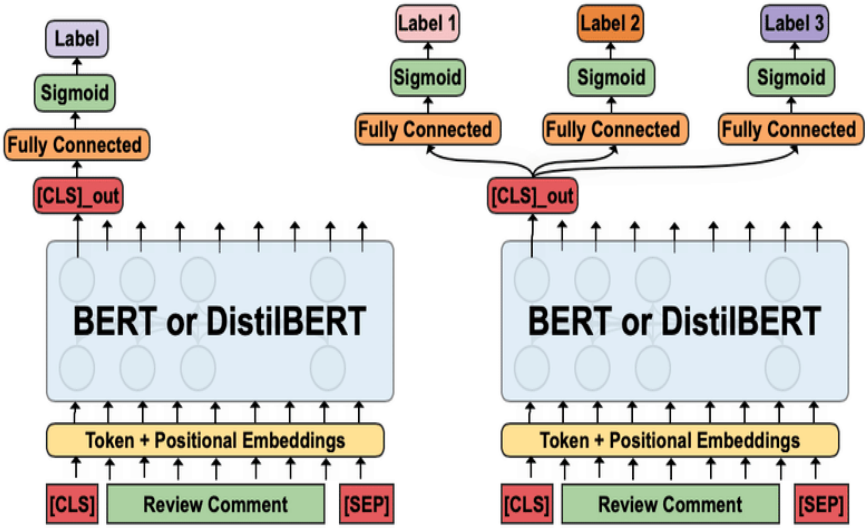

Figure 2. BERT and DistilBERT-Based single-tasking and multitasking learning

The BERT model is part of Transformers and is designed to perform classification. This model takes a given text input and produces output to predict which class this text belongs to and can be used especially in natural language processing tasks such as sentiment analysis, spam filtering and language recognition (Hugging Face, 2023b; Dogra et al., 2022)

DistilBERT is a lighter and faster version of the BERT model. This architecture retains the basic coding of BERT while making some optimizations to have fewer parameters and run faster (Sanh et al., 2019). DistilBERT uses a combination of BERT's Bidirectional Attention Mechanism and output layers. However, unlike BERT, DistilBERT offers a lighter model by reducing the number of layers.

In this study, the Transformer architecture is used to perform sentiment analysis on Amazon review data. This architecture has a similar working logic with other architectures. The model used in this study is the DistilBERT, shown in Figure 2.

RESULTS AND DISCUSSION

In this section, we will provide implementation details and present the results of the conducted experiments. Subsequently, we will engage in discussions to further analyze the findings.

Implementation Details

Figure 3 depicts the preprocessing and other steps undertaken in our study. These steps were implemented to prepare the text data for analysis and modeling. The preprocessing tasks included tokenization, stopword removal, and stemming/lemmatization to standardize and normalize the text data. Furthermore, other processing steps, such as feature extraction, feature engineering, and data splitting, were carried out to extract meaningful features from the text data and create appropriate inputs for subsequent analyses. The details of these steps are visually depicted in Figure 3 for a comprehensive understanding of our data preparation process across all four methods. Table 1 displays the hyperparameters utilized in the four methods employed in our study. The hyperparameters are tested and evaluated on a validation set to find the best performing model.

Figure 4 presents the confusion matrices for the four methods: MLP, Naive Bayes, Decision Tree, and DistilBERT. A confusion matrix provides a detailed breakdown of the predicted and actual class labels for a classification model. It helps assess the model's performance in terms of correctly classifying instances into true positives, true negatives, false positives, and false negatives. By examining Figure 4, we can see the distribution of these predictions for each method. It appears that DistilBERT achieves the highest number of correct predictions (true positives and true negatives), indicating its superior performance in correctly classifying instances compared to the other methods.

Figure 5 shows the Receiver Operating Characteristic (ROC) curves for the four methods: MLP, Naive Bayes, Decision Tree, and DistilBERT. The ROC curve is a graphical representation of the trade-off between the true positive rate and the false positive rate for different classification thresholds. A method with a higher ROC curve that is closer to the top-left corner indicates better performance in distinguishing between positive and negative classes. By analyzing Figure 5, it can be observed that DistilBERT demonstrates the highest performance among the four methods, as its ROC curve is closer to the top-left corner.

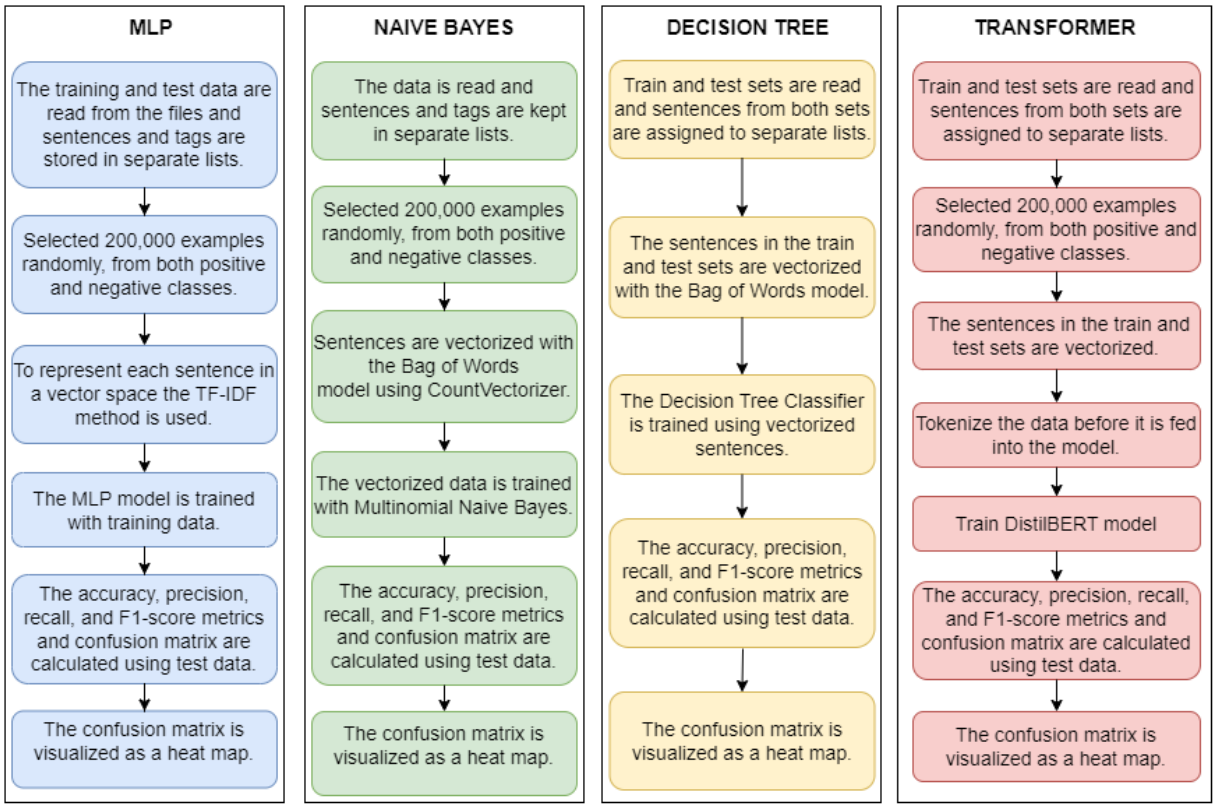

Figure 3. Flowcharts of the four methods used in this study

The methods include a multilayer perceptron (MLP), the Naive Bayes algorithm, the Decision Tree architecture, and the Transformer architecture with the DistilBERT model. The input to each method is the preprocessed text data, and the output is the predicted sentiment label (positive or negative). The performance of each method is evaluated based on accuracy, precision, recall, and F1-score metrics.

Table 1. Hyperparameters used in the four methods

Architecture	Hyperparameter	Values
MLP	Number of Layers	3
	Hidden Layer Size	128
	Activation Function	ReLU
	Output Activation	Softmax
	Loss Function	Cross-Entropy
Naive Bayes	Smoothing	0.7
Decision Tree	Maximum Depth	10
	Minimum Samples Split	7
Transformer	Number of Layers	10
	Embedding Size	128
	Attention Heads	7
	Feedforward Hidden Size	1024
	Dropout	0.4
	Learning Rate	5e-5

Results

In this study, sentiment analysis is performed on Amazon reviews using different machine learning algorithms. All reviews from the same dataset were analyzed in 4 different architectures and the results were compared according to evaluation metrics. Based on the obtained results, the BERT-base model of the Transformer architecture demonstrated the highest success rate, achieving an accuracy of 96.10%.

The MLP architecture ranked second with an accuracy of 85.06%, followed by the Naive Bayes algorithm with an accuracy rate of 82.63%. The Decision Tree architecture yielded the lowest accuracy rate of 70.70%.

The MLP classifier exhibits the highest precision, recall, and F1-score values compared to Naive Bayes and Decision Tree architectures. According to the evaluation results in Table 2, the MLP classifier has the ability to correctly predict 84% of the classified examples. 85% of the predictions for the Negative class and 86% of the predictions for the Positive class are correct. Furthermore, the F1-score value of the classifier is calculated as 86%, which is a measure of its classification performance. These results indicate that the MLP classifier is an effective classification tool.

Table 2. Evaluation results of the MLP

	Precision	Recall	F1-Score	Accuracy
Negative	0.84	0.85	0.85	0.84
Positive	0.86	0.85	0.85	0.88

According to the evaluation results reported in Table 3, the Naive Bayes classifier demonstrated moderate performance in sentiment classification, achieving 80% precision and 87% recall for the Negative class, and 86% precision and 78% recall for the Positive class.

Table 3. Evaluation results of the Naive Bayes

	Precision	Recall	F1-Score	Accuracy
Negative	0.80	0.87	0.83	0.85
Positive	0.86	0.78	0.81	0.79

The performance of the Decision Tree architecture model in classifying data labeled as "Negative" and "Positive" is presented in Table 4. The model achieved an accuracy of 0.67, a recall value of 0.59, and an F1-score of 0.66 for the data labeled as "Negative". In contrast, the model achieved an accuracy of 0.75, a recall value of 0.82, and an F1-score of 0.74 for the data labeled as "Positive". These results suggest that the model performs better in classifying the data labeled as "Positive" than the data labeled as "Negative".

Table 4. Evaluation results of the Decision Tree

	Precision	Recall	F1-Score	Accuracy
Negative	0.76	0.59	0.66	0.67
Positive	0.68	0.82	0.74	0.75

The BERT model achieved high accuracy rates of 96% in sentiment analysis, with precision, recall, and F1-Score of 96% in both Negative and Positive classes. These findings suggest that the BERT model is effective in sentiment analysis tasks. Table 5 summarizes the evaluation results for the BERT model, indicating high precision, recall, and F1-Score values for both classes. The results can be used by researchers and practitioners to develop sentiment analysis models with high accuracy rates, aiding businesses and organizations in making informed decisions based on customer feedback and opinions.

Table 5. Evaluation results of the Transformer

	Precision	Recall	F1-Score	Accuracy
Negative	0.96	0.96	0.96	0.96
Positive	0.96	0.96	0.96	0.97

Table 5 demonstrates the high performance of the transformer model in sentiment analysis. The precision, recall, and F1-Score values for the Negative and Positive classes are 0.96, suggesting that the model effectively predicts these classes. The high overall accuracy of 96% also indicates the model's

effectiveness in sentiment analysis tasks. Thus, the findings suggest that the model is successful in sentiment analysis tasks. These results can be used by researchers and practitioners in developing and refining sentiment analysis models, which can assist organizations in making informed decisions based on customer feedback and opinions.

Overall, the Transformer architecture demonstrates superior performance in sentiment analysis, followed by the MLP classifier as a close second. The Naive Bayes classifier also exhibits commendable performance in this task. However, the Decision Tree classifier lags behind the other three models in terms of accuracy, which may be attributed to its simple structure and limited capacity to handle the complexity of the data. These findings provide insights into the effectiveness of various machine learning models in sentiment analysis tasks.

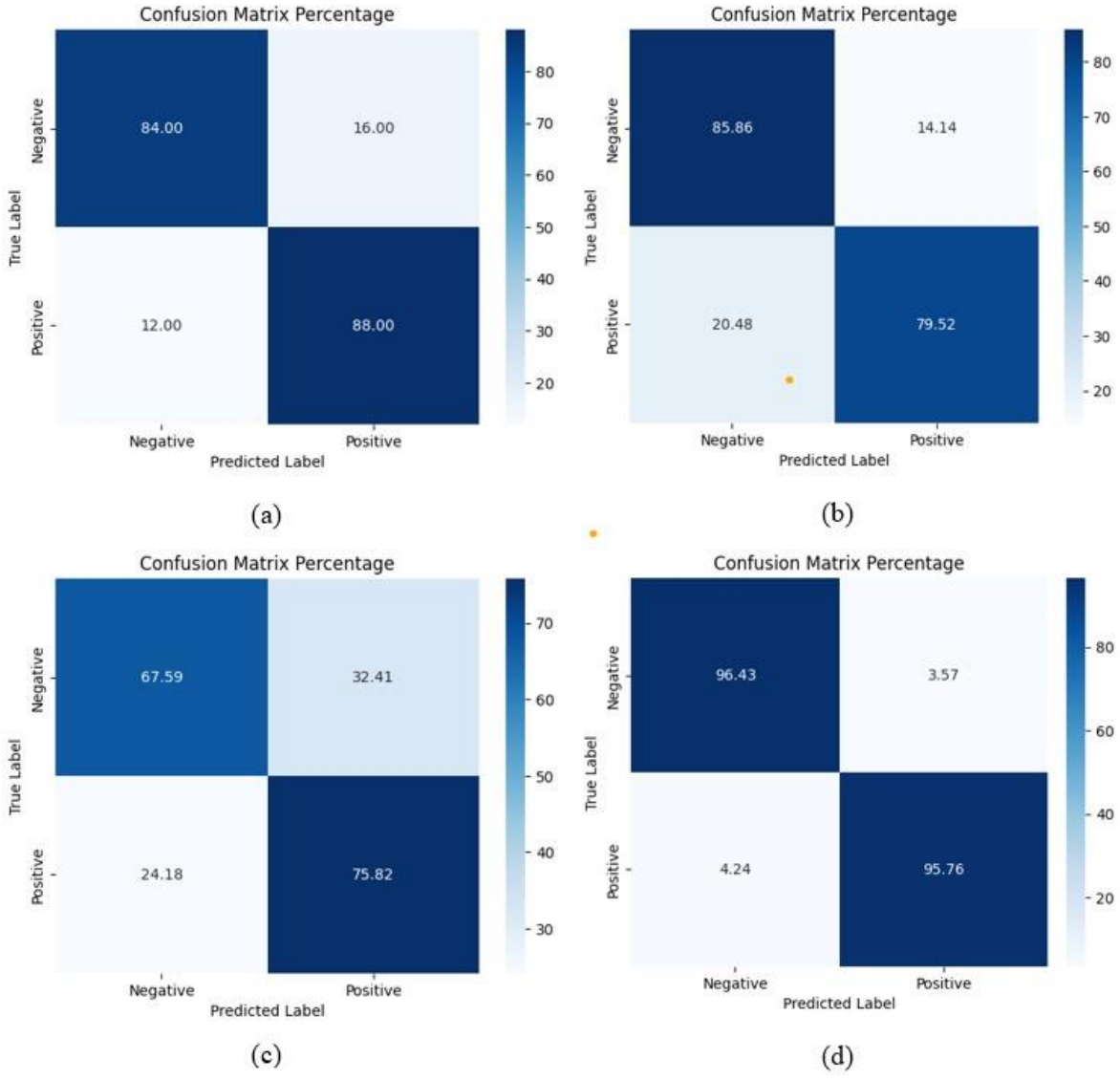

Figure 4. Confusion matrices of the methods a) MLP, b) Naïve Bayes c) Decision Tree d) DistilBERT

In conclusion, machine learning algorithms have been widely used in natural language processing problems such as sentiment analysis. The performance of different algorithms can vary depending on factors such as data set size and characteristics. Therefore, when deciding which algorithm to use, factors such as dataset size and characteristics should also be taken into account. Figure 4 and Figure 5 illustrate the confusion matrices and ROC curves, respectively, in our study.

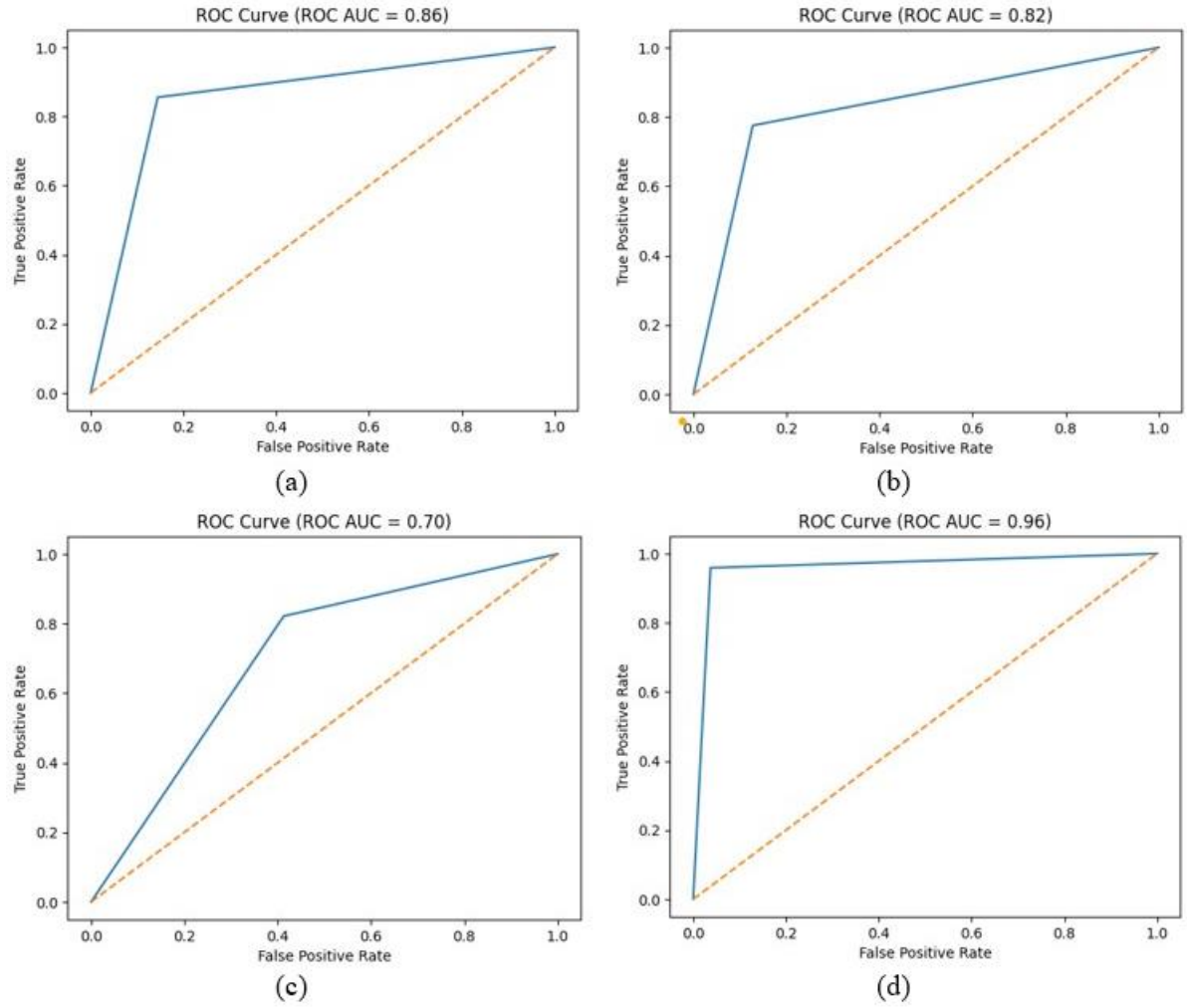

Figure 5. ROC curves of the mothods a) MLP b) Naive Bayes c) Decision Tree d) DistilBERT

We conducted a comprehensive comparison with previous studies in the literature, as depicted in Table 6, along with the corresponding outcomes. Based on the literature review presented, it is evident that BERT models have been highly effective in achieving good results. This study also confirms that the use of the Transformers architecture has resulted in the best performance rates, while MLP algorithms have outperformed other algorithms.

Table 6. Accuracy rates of the studied in the literature

Study	Algorithm	Dataset	Accuracy
Balahur et al. (2013)	SVM	Movie review dataset	88.2%
Devlin et al. (2018)	BERT	GLUE benchmark (MRPC task)	93.2%
Liu et al. (2019)	RoBERTa	GLUE benchmark (RTE task)	95.2%
Sanh et al. (2019)	DistilBERT	GLUE benchmark (MRPC task)	91.4%
Lan et al. (2020)	ALBERT	GLUE benchmark (RTE task)	93.0%
Sun et al. (2021)	Multi-tasking	CMRC 2018 and CMRC 2019	92.6%
Xie et al. (2021)	BERT	Amazon Reviews dataset	92.5%
Huang et al. (2021)	XLNet	Amazon Reviews dataset	93.6%

Table 7. Our evaluation results

Architecture	Accuracy Rate (%)	F1-Score
Transformer	96.10	0.96
MLP	86.06	0.85
Naive Bayes	82.63	0.82
Decision Tree	71.70	0.71

It is important to note that these findings are specific to the dataset used in this study, and further research using different datasets will be necessary to assess the generalizability of these results and inform algorithm selection. Table 7 presents the evaluation results of four sentiment classification methods: Transformer, MLP, Naive Bayes, and Decision Tree. The evaluation metrics used are Accuracy Rate (%) and F1-Score.These findings provide valuable insights into the relative performance of the different methods in sentiment classification, based on the chosen evaluation metrics. They contribute to the understanding of the strengths and limitations of each method, offering guidance for their practical application in sentiment analysis tasks.

Discussion

The evaluation results of different machine learning algorithms for sentiment analysis on Amazon reviews reveal important insights into their performance and effectiveness. Among the methods examined, the Transformer architecture, specifically the BERT model, emerges as the most successful, achieving an impressive accuracy rate of 96.10%. This aligns with prior research demonstrating the prowess of Transformer models in capturing contextual relationships and attaining state-of-the-art results in natural language processing tasks.

The MLP architecture also demonstrates promising performance, achieving an accuracy rate of 85.06%. This underscores the capability of MLP models to capture complex non-linear relationships, making them a strong candidate for sentiment classification tasks. However, it is essential to note that the MLP model's accuracy falls slightly short of the Transformer model, underscoring the significance of leveraging contextual information through self-attention mechanisms.

The Naive Bayes algorithm, despite its simplicity, exhibits moderate performance with an accuracy rate of 82.63%. This indicates that Naive Bayes remains a viable option for sentiment analysis, particularly in scenarios where computational resources are constrained. Nevertheless, it is important to acknowledge that Naive Bayes models may struggle to capture the intricate nuances of sentiment compared to more advanced models such as Transformers and MLPs.

In contrast, the Decision Tree architecture displays the lowest accuracy rate of 71.70%. This can be attributed to the Decision Tree's limited capacity to handle the intricacies of sentiment analysis and effectively capture the underlying patterns. Decision Trees rely on simplistic hierarchical rules, which may prove inadequate for accurately classifying sentiment.

Further analysis of the evaluation metrics, including precision, recall, and F1-score, reinforces the performance disparities between the methods. The MLP classifier outperforms the others, demonstrating the highest precision, recall, and F1-score values. This attests to its effectiveness in correctly predicting both positive and negative sentiment labels. The Naive Bayes classifier exhibits moderate precision and recall values, indicating its ability to make reasonably accurate predictions. However, the Decision Tree classifier lags behind, displaying lower precision and recall values and thus revealing its limitations in accurate sentiment classification.

The superior performance of the Transformer architecture, particularly the BERT model, can be attributed to its capacity to capture contextual relationships and leverage large-scale pretraining. The self-attention mechanisms employed in Transformers facilitate the modeling of long-range dependencies and semantic relationships, contributing to their remarkable success in sentiment analysis tasks.

It is crucial to acknowledge the limitations of this study. The evaluation was conducted on a small subset of the AR-P dataset, potentially restricting the generalizability of the results. Additionally, the hyperparameters employed in each method were optimized based on the validation set, introducing a potential source of bias that may influence the performance outcomes.

In conclusion, the evaluation outcomes highlight the superior performance of the Transformer architecture, particularly the BERT model, in sentiment analysis tasks. The MLP architecture also exhibits promising results, while the Naive Bayes algorithm demonstrates moderate performance. The Decision Tree architecture lags behind the other methods in terms of accuracy. These findings contribute to the understanding of the efficacy of different machine learning models in sentiment analysis and provide valuable insights for their practical application in real-world scenarios. Future research can build upon these findings by investigating larger datasets, incorporating more advanced models, and considering additional evaluation metrics to gain a comprehensive understanding of sentiment analysis techniques.

CONCLUSION

The aim of this study was to compare the performance of different classification methods using the AR-P dataset. MLP, Naive Bayes, Decision Tree and Transformer architectures were introduced to the training dataset under the same conditions and their success rates were compared. The results support the proven success of the Transformer architecture and show that MLP algorithms give better results than the other algorithms. However, these results only depend on the dataset used in this study. Further studies on different datasets will provide a better basis for generalization of the results and algorithm selection. The results of this study can serve as a basis for future work. For example, comparing the performance of algorithms using a larger dataset, adding more classification methods, or comparing the performance of algorithms on a different task (e.g., topic classification instead of sentiment analysis). This study could be an important step towards using machine learning methods in real-world applications such as sentiment analysis.

Conflict of Interest

The article authors declare that there is no conflict of interest between them.

Author's Contributions

The authors declare that they have contributed equally to the article.

REFERENCES

Abdi, A., Shamsuddin, S. M., Hasan, S., & Piran, J. (2019). Deep learning-based sentiment classification of evaluative text based on Multi-feature fusion. Information Processing & Management, 56(4), 1245-1259.

Ain, Q. T., Ali, M., Riaz, A., Noureen, A., Kamran, M., Hayat, B., & Rehman, A. (2017). Sentiment analysis using deep learning techniques: a review. International Journal of Advanced Computer Science and Applications, 8(6).

Alexandridis, G., Varlamis, I., Korovesis, K., Caridakis, G., & Tsantilas, P. (2021). A survey on sentiment analysis and opinion mining in greek social media. Information, 12(8), 331.

Al-Garadi, M. A., Yang, Y. C., Cai, H., Ruan, Y., O'Connor, K., Graciela, G. H., ... & Sarker, A. (2021). Text classification models for the automatic detection of nonmedical prescription medication use from social media. *BMC Medical Informatics and Decision Making*, 21(1), 1-13. DOI: 10.1186/s12911-021-01488-1

Balahur, A., Turchi, M., & Steinberger, R. (2013). Multilingual sentiment analysis using machine translation–based techniques. ACM Transactions on Intelligent Systems and Technology (TIST), 4(1), 1-26. DOI: 10.1145/2444776.2444777

Bayat, S., & Işık, G. (2022). Recognition of Aras Bird Species From Their Voices With Deep Learning Methods. Journal of the Institute of Science and Technology, 12(3), 1250-1263.

Bollen, J., Mao, H., &. Zeng, X. (2011). Twitter mood predicts the stock market. Journal of Computational Science, 2(1), 1-8. https://doi.org/10.1016/j.jocs.2010.12.007

Boyko, N., & Boksho, K. (2020, November). Application of the Naive Bayesian Classifier in Work on Sentimental Analysis of Medical Data. In Proceedings of the International Conference on Intelligent Data and Digital Medicine (IDDM) (pp. 230-239).

Cambria, E., & White, B. (2014). Jumping NLP curves: A review of natural language processing research. IEEE Journal of Computational Intelligence, 9(2), 48-57. https://doi.org/10.1109/MCI.2014.2307227

Chen, L. C., Lee, C. M., and Chen, M. Y. (2020) published a study in Soft Computing, in which they explored social media for sentiment analysis using deep learning techniques.

Delobelle, P., Winters, T., & Berendt, B. (2020). Robbert: a dutch roberta-based language model. arXiv preprint arXiv:2001.06286.

Devlin, J., Chang, M. W., Lee, K., & Toutanova, K. (2018). BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. arXiv preprint arXiv:1810.04805.

Dogra, V., Verma, S., Kavita, C., Chatterjee, P., Shafi, J., Choi, J., & Ijaz, M. F. (2022). A Complete Process of Text Classification System Using State-of-the-Art NLP Models. Computational Intelligence and Neuroscience, 2022, 1883698. https://doi.org/10.1155/2022/1883698

Gao, J., &. Wong, K.-F. (2014). A review of sentiment analysis research in Chinese language. Informatics, 1(3), 191-208. https://doi.org/10.3390/informatics1030191

Ghulam, H., Zeng, F., Li, W., & Xiao, Y. (2019). Deep learning-based sentiment analysis for roman urdu text. Procedia computer science, 147, 131-135.

Goodfellow, I., Bengio, Y., & Courville, A. (2016). Deep Learning. MIT Press.

Gupta, R., Sameer, S., Muppavarapu, H., Enduri, M. K., & Anamalamudi, S. (2021, September). Sentiment analysis on Zomato reviews. In 2021 13th International Conference on Computational Intelligence and Communication Networks (CICN) (pp. 34-38). IEEE.

Gündüz, M. Ş., & Işık, G. (2023). A new YOLO-based method for social distancing from real-time videos. Neural Computing and Applications, 1-11.

Han, J., & Kamber, M. (2011). Data mining: concepts and techniques. Elsevier.

Hugging Face. (2023b). AutoModelForSequenceClassification. https://huggingface.co/transformers/model_doc/auto.html#automodelforsequenceclassification

Hutto, C. J., & Gilbert, E. (2014). VADER: A parsimonious rule-based model for sentiment analysis of social media text. Eighth International AAAI Conference on Weblogs and Social Media. https://www.aaai.org/ocs/index.php/ICWSM/ICWSM14/paper/view/8109/8122

Joshi, A., Bhattacharyya, P., & Carman, M. J. (2020). A thorough examination of the DistilBERT model for sentence classification. arXiv preprint arXiv:2010.16061.

Kim, S. (2020). Sentiment analysis: A comprehensive guide to detecting emotions, opinions, and sentiments.

Kumar, V. (2022). A Review of Decision Tree Algorithms for Classification in Machine Learning. International Journal of Computer Applications, 182(40), 10-16.

Lan, Z., Chen, M., Goodman, S., Gimpel, K., Sharma, P., & Soricut, R. (2020). ALBERT: A Lite BERT for Self-supervised Learning of Language Representations. arXiv preprint arXiv:1909.11942.

Liu, Y., Ott, M., Goyal, N., Du, J., Joshi, M., Chen, D., ... & Stoyanov, V. (2019). RoBERTa: A Robustly Optimized BERT Pretraining Approach. arXiv preprint arXiv:1907.11692.

M. Al-Qurishi, T. Khalid and R. Souissi, "Deep Learning for Sign Language Recognition: Current Techniques, Benchmarks, and Open Issues," in *IEEE Access*, vol. 9, pp. 126917-126951, 2021, doi: 10.1109/ACCESS.2021.3110912.

Mohammad, S. M., & Bravo-Marquez, F. (2017). WASSA-2017 shared task on emotion intensity. Proceedings of the 8th Workshop on Computational Approaches to Subjectivity, Sentiment and Social Media Analysis, 1-10. https://doi.org/10.18653/v1/W17-5201

Nandwani, P., & Verma, R. (2021). A review on sentiment analysis and emotion detection from text. Social Network Analysis and Mining, 11(1), 81.

Othan, D., Kilimci, Z. H., & Uysal, M. (2019, December). Financial sentiment analysis for predicting direction of stocks using bidirectional encoder representations from transformers (BERT) and deep learning models. In Proc. Int. Conf. Innov. Intell. Technol (pp. 30-35).

Pang, B., & Lee, L. (2008). Opinion mining and sentiment analysis. Foundations and Trends in Information Retrieval, 2(1-2), 1-135. https://doi.org/10.1561/1500000011

Poria, S., Cambria, E., &. Bajpai, R. (2017). A review of affective computing: From unimodal analysis to multimodal fusion. Information Fusion, 37, 98-125. https://doi.org/10.1016/j.inffus.2017.02.003

Raschka, S., & Mirjalili, V. (2021). Naive Bayes and Text Classification. In Python Machine Learning, Third Edition (pp. 373-394). Packt Publishing.

Ray, P., & Chakrabarti, A. (2022). A mixed approach of deep learning method and rule-based method to improve aspect level sentiment analysis. Applied Computing and Informatics, 18(1/2), 163-178.

Sanh, V., Debut, L., Chaumond, J., & Wolf, T. (2019). DistilBERT, a distilled version of BERT: smaller, faster, cheaper and lighter. arXiv preprint arXiv:1910.01108.

Sudhir, P., & Suresh, V. D. (2021). Comparative study of various approaches, applications and classifiers for sentiment analysis. Global Transitions Proceedings, 2(2), 205-211.

Sun, C., Li, L., Wang, W., &. Jiang, B. (2021). Multi-task learning for sentiment analysis using transformer- Khalid based models. Neural Networks, 137, 181-190. https://doi.org/10.1016/j.neunet.2020.11.010

Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., ... & Polosukhin, I. (2017). Attention is all you need. Advances in Neural Information Processing Systems, 30, 5998-6008.

Wankhade, M., Rao, A. C. S., & Kulkarni, C. (2022). A survey on sentiment analysis methods, applications, and challenges. Artificial Intelligence Review, 55(7), 5731-5780.