BACA: Jurnal Dokumentasi dan Informasi, Vol. 46, No. 1, 2025
ISSN 0125-9008 (Print); ISSN 2301-8593 (Online)
DOI: 10.55981/j.baca.2025.8240
SK Dirjen Dikti Ristek - Kemdikbudristek No. 105/E/KPT/2022 (Peringkat 2 SINTA)

Optimizing Multi-Layer Perceptron performance in sentiment classification through neural network feature extraction

Muhammad Fikri Alam[1*]; Aang Nuryaman[1]; Purnomo Husnul Khotimah[2*]; Anne Parlina[2]; Andre Sihombing[2]

[1]Universitas Lampung, Lampung, Indonesia
[2]Research Center for Data Science and Information, National Research and Innovation Agency

*Corresponding author: purn005@brin.go.id

Diajukan: 22-10-2024; **Direview**: 04-01-2025; **Diterima**: 12-03-2025; **Direvisi**: 18-02-2025

ABSTRACT

There are some problems with using the Multi-Layer Perceptron (MLP) model for complex tasks because it can be hard to understand hierarchical relationships and tends to overfit data with a lot of dimensions. This research proposes an enhanced MLP model for sentiment classification by integrating feature extraction layers from advanced neural networks, specifically the Convolutional Neural Network (CNN), Long Short-Term Memory (LSTM), and Bidirectional LSTM (Bi-LSTM). These layers aim to improve the model's representation capabilities by capturing more nuanced features. To evaluate the performance improvements of this augmented MLP model, metrics such as accuracy, precision, recall, F1-score, and the Area Under the Curve for Receiver Operating Characteristics (ROC-AUC) were employed. A key metric focus is the delta value, representing changes in the ROC-AUC, to assess the significance of these enhancements. The integration of CNN as a feature extraction layer yielded optimal ROC-AUC results, achieving values of 93.30% and 93.00%, which reflect an improvement of 0.51% and 4.46% over the baseline model. These findings indicate that adding feature extraction layers significantly enhances MLP performance in sentiment classification tasks. Future research may explore the potential of using alternative neural networks as feature extractors to continue advancing MLP capabilities in complex NLP applications.

ABSTRAK

Model Multi-Layer Perceptron (MLP) adalah arsitektur deep learning yang memiliki keterbatasan dalam tugas-tugas kompleks karena tantangan dalam menangkap hubungan hierarkis serta kecenderungannya untuk overfitting, terutama pada data berdimensi tinggi. Penelitian ini mengusulkan model MLP yang ditingkatkan untuk klasifikasi sentimen dengan mengintegrasikan lapisan ekstraksi fitur dari neural networks, khususnya Convolutional Neural Network (CNN), Long Short-Term Memory (LSTM), dan Bidirectional LSTM (Bi-LSTM). Lapisan-lapisan ini bertujuan untuk meningkatkan kemampuan representasi model dengan menangkap fitur yang lebih halus. Untuk mengevaluasi peningkatan kinerja model MLP dengan penambahan layer ini, digunakan metrik seperti akurasi, presisi, recall, F1-score, dan Area Under the Curve for Receiver Operating Characteristics (ROC-AUC). Fokus utama metrik adalah nilai delta yang mewakili perubahan pada ROC-AUC untuk menilai signifikansi peningkatan tersebut. Integrasi CNN sebagai lapisan ekstraksi fitur menghasilkan nilai ROC-AUC optimal, mencapai 93,30% dan 93,00%, yang mencerminkan peningkatan sebesar 0,51% dan 4,46% jika dibandingkan dengan model dasar. Temuan ini menunjukkan bahwa penambahan layer ekstraksi fitur secara signifikan meningkatkan kinerja MLP dalam tugas klasifikasi sentimen. Penelitian selanjutnya dapat lebih mengeksplorasi potensi penggunaan jaringan saraf alternatif sebagai ekstraktor fitur untuk terus mengembangkan kemampuan MLP dalam aplikasi NLP yang kompleks.

Keywords: Multi-layer perceptron, Deep learning, Neural net, Performance improvement

1. INTRODUCTION

Sentiment classification is crucial in Natural Language Processing (NLP) and machine learning. It involves using computational techniques to detect and categorize the emotions or attitudes conveyed by a speaker or writer (Tan et al., 2023; Nandwani and Verma, 2022). This task focuses on extracting subjective information from text data to interpret the content's sentiments, opinions, evaluations, judgments, and emotions. Sentiment classification can be binary (negative, positive) (Khotimah et al., 2023) or multiclass (negative, positive, neutral) (Cahyo et al., 2024) and applied at various levels, including document, sentence, or sub-sentence levels (The NYU School of Professional Studies, 2024). Sentiment classification, also known as sentiment analysis, opinion analysis, or opinion mining, has gained widespread acceptance in recent years. It has drawn attention not only from researchers but also from organizations, businesses, and governments. Since the Internet has become the primary global information source, countless users are relying on various online platforms to express their views and opinions. Automated analysis of user-generated data is essential to monitor public sentiment effectively and support decision-making. As a result, sentiment classification has become increasingly prominent within research communities (Wankhade et al., 2022).

In recent years, deep learning has revolutionized the field of natural language processing, especially in sentiment classification tasks. Many tools, including machine learning and deep learning algorithms, are available for sentiment classification. Choosing the appropriate method is crucial, as it directly affects classification accuracy and ensures reliable results. Consequently, numerous studies have aimed to identify and refine the most effective techniques, continually improving existing methods to enhance performance (Socher et al., 2013; Vaswani et al., 2017).

Advanced models like the Transformer, introduced by Vaswani et al. (2017), have shown a remarkable ability to capture complex semantic and syntactic structures within text. These models effectively use attention mechanisms to pinpoint critical text components determining sentiment. On the other hand, the Multi-Layer Perceptron (MLP), one of the earliest neural network architectures, is often regarded as less competitive for complex tasks like sentiment classification. This is primarily due to MLP's limitations in capturing hierarchical relationships and long-term dependencies within textual data. Furthermore, MLP models tend to overfit on high-dimensional data, often requiring sophisticated regularization techniques to improve generalization Bayat and Işık (2023).

Several studies have explored the hybridization of MLP with other neural network models. For instance, Akhtar et al. (2017) utilized an MLP model enhanced with CNN, LSTM, and GRU to improve performance. Similarly, Munandar et al. (2021), implemented a combination of CNN, LSTM, and MLP models, integrating their learning processes before performing classification. In contrast, the present research distinguishes itself by focusing on MLP, which employs neural networks as a preliminary step before undergoing additional learning and classification specifically within the MLP framework. This approach addresses a key research gap, as previous studies primarily integrated multiple neural network architectures, whereas the current study emphasizes MLP's role as a standalone classifier following neural network utilization.

This study aims to address the limitations of MLP in sentiment classification tasks by proposing a hybrid approach that combines the feature extraction strengths of advanced neural networks with the generalization capabilities of MLP. By incorporating more representative features, the objective of this research is to enable MLP's to better capture complex patterns in textual data. This approach is inspired by previous research that demonstrated how low-level feature extraction can enhance the performance of classification models (Socher et al., 2013). It is hypothesized that integrating the feature representation strengths of neural networks, such as the Bidirectional Encoder Representations from Transformers (BERT) model proposed by Devlin et al. (2018), with the generalization capabilities of MLP, will result in a hybrid model capable of capturing richer information from the text and achieving higher classification accuracy.

The findings of this study are expected to contribute to the development of more efficient and effective text classification models, particularly in handling increasingly complex and large-scale data. Additionally, this research provides valuable insights into the interaction between traditional MLP models and their enhanced versions. This work's broader implications span various applications, including social media sentiment analysis, product reviews, and public opinion surveys. By improving sentiment classification performance, the proposed hybrid model offers the potential for more accurate and nuanced insights into public sentiment on a wide range of topics. Moreover, this study may pave the way for future advancements in hybrid models that combine the strengths of multiple neural network architectures, further enhancing text classification capabilities.

The structure of this paper is organized as follows. It begins with an introduction, which overviews the research topic and outlines the study's objectives. Following the introduction is a literature review, where previous research and relevant theories are discussed to contextualize the current work. The following section presents a detailed explanation of the data and methods employed in the study. This section outlines the data sources and the methodologies used for analysis and evaluation. Subsequently, the results section presents the findings of the study, supported by relevant data and analysis. This is followed by a discussion, where the implications of the results are explored, potential limitations are acknowledged, and suggestions for future research are provided. Finally, the paper concludes with a conclusion section, summarizing the essential findings and their significance and reiterating the contribution of the study to the field.

2. LITERATURE REVIEW

The Multi-Layer Perceptron (MLP) is a deep learning model based on the concept of artificial neural networks. As its name suggests, the architecture of this model consists of multiple layers arranged hierarchically and interconnected, including an input layer, hidden layers, and an output layer. Each of these layers contains perceptrons, also known as neurons, which process data from neurons in the previous layer using weights, biases, and activation functions. The number of neurons in the input layer corresponds to the number of features in the dataset, while the number of neurons in the output layer matches the number of target classes (Ramchoun et al., 2016).

Sentiment analysis, or opinion mining, utilizes Natural Language Processing (NLP) and text mining techniques to extract and interpret opinions, emotions, and sentiments from textual data, categorizing sentiment polarity as positive, negative, or neutral (Wankhade et al., 2022). This approach has gained significant importance in areas such as social media monitoring, customer feedback analysis, and product reviews, where it supports informed decision-making. Organizations, governments, and individuals leverage sentiment analysis to understand public opinions, enhance products and services, and monitor market trends effectively. The objectives of sentiment analysis revolve around understanding and leveraging insights from textual data to analyze public emotions, opinions, and attitudes (Wankhade et al., 2022). A key objective is monitoring public sentiment by analyzing user-generated content, including social media posts and online reviews, to identify trends, preferences, and areas of concern. By extracting subjective information and identifying sentiment polarity, sentiment analysis contributes to better decision-making, product development, and strategic planning, offering a deeper understanding of public attitudes.

Sentiment classification, a fundamental component of sentiment analysis, focuses on assigning sentiment labels, such as positive, negative, or neutral, to textual data. This classification is conducted at various levels, including document-level (assigning a single sentiment to the entire text), sentence-level (analyzing the sentiment of individual sentences), and aspect-level (evaluating specific aspects such as "battery life" or "camera performance" in product reviews) (Wankhade et al., 2022).

Sentiment classification using deep learning has gained substantial traction due to its ability to identify emotions in text accurately. Researchers have developed various models, including hybrid architectures, to enhance performance across different languages and contexts. For instance, Salur and Aydin (2020) combined several word embedding algorithms with four deep learning models: CNN, LSTM, Bi-LSTM, and Gated Recurrent Unit (GRU), to boost sentiment classification accuracy. Their study showed that the proposed model surpassed previous sentiment classification models in performance.

Munandar et al. (2021) conducted a study that combined the MLP model with CNN and LSTM for sentiment classification tasks. Using two different datasets, they demonstrated that combining these three neural network models yielded better results in terms of accuracy compared to a basic MLP model. On the first dataset, an improvement of 27.58% was observed, while the second dataset showed an increase of 16.93%. This improvement occurred because the combination of these models leverages the strengths of each model to achieve optimal results. Each model independently learns from the data, and the outputs from each learning process are then integrated before performing classification.

Similarly, Dang et al. (2020) conducted research comparing the performance of Deep Neural Networks (DNN), CNN, and Recurrent Neural Networks (RNN) across eight sentiment analysis datasets with data sizes ranging from thousands to millions of records. They evaluated and compared the accuracy and processing times of each model. The experiments show that CNN provides the best balance between processing time and accuracy among the three deep learning models (DNN, CNN, and RNN) used for sentiment analysis. In contrast, RNN, despite achieving the highest accuracy with word embedding, takes ten times longer than CNN. RNN performs poorly with TF-IDF, which leads to less effective results, whereas DNN has average processing times and results. Word embedding generally outperforms TF-IDF in converting text to numeric vectors, especially with RNN. Additionally, sentiment classification accuracy was higher with tweets and IMDB movie reviews datasets, while topic-specific datasets, such as Tweets Airline, performed better than general-topic ones.

Tan et al. (2022) also explored deep learning for sentiment analysis, implementing eleven different methods, including CNN-LSTM, CNN-Bi-LSTM, and their novel approach, RoBERTa-LSTM. Across three datasets, their proposed model demonstrated the highest performance, showing a marked improvement from 79.10% to 89.70% over the base model, RoBERTa. RoBERTa's word embeddings, in particular, contributed to the LSTM's ability to capture temporal information effectively. In another study, Khotimah et al. (2020) sought to detect dengue fever events using online news with three deep learning models. They found that, compared to MLP, CNN required almost twice the computational time but achieved superior performance, with a 3-4% improvement.

The Multi-Layer Perceptron (MLP) serves as the base architecture in this research. MLP is a widely utilized neural network that operates on a supervised learning framework characterized by information flowing in a single direction without any loops. It consists of three layers: an input, an output, and one or more hidden layers. The input layer gathers the features that will be processed, while the hidden layers, which can be of any number, serve as the computational units within the MLP. The output layer is responsible for tasks like prediction and classification. (Naskath et al., 2023)

MLP has emerged as a significant method for sentiment analysis, demonstrating competitive performance compared to other machine learning algorithms. Studies indicate that MLP can effectively classify sentiments in various datasets, achieving notable accuracy rates. Bayat and Işık (2023) compared four machine learning algorithms: MLP, Naive Bayes, Decision Tree, and Transformer architectures, specifically the DistilBERT model. MLP achieved an accuracy of 85.06%, which was higher than both Naive Bayes (82.63%) and Decision Tree (70.70%), but significantly

lower than the Transformer-based Distil-BERT model, which reached an impressive accuracy of 96.10%. Nevertheless, MLP remains a viable option for sentiment analysis, particularly in scenarios where computational resources are limited or simpler models are preferred.

Shaker (2022) shows that the MLP classifier can achieve impressive accuracy rates, reaching 85% on the Twitter dataset and 99% on the movie review dataset. By using two different datasets—one from Twitter and the other from movie reviews—the study highlights the MLP's adaptability. The classifier is tested on short, informal texts and longer, structured reviews, demonstrating its versatility in handling different types of sentiment data. These findings suggest that the MLP effectively distinguishes positive and negative sentiments across diverse datasets.

Building on prior research, this study aims to evaluate the performance of a base MLP model and compare it with an MLP model enhanced by feature extraction layers from CNN, LSTM, and Bi-LSTM. The evaluation will focus on identifying the model with the highest performance for sentiment classification tasks.

The dataset is labelled to categorize the entries into three sentiment categories based on the COVID-19 conditions presented in the news (Shamrat et al., 2021): (1) Positive (1): News that conveys a positive sentiment towards the COVID-19 situation, indicating signs of improvement, such as a decrease in COVID-19 cases or an increase in recovered patients; (2) Neutral (0): News that does not indicate whether the situation is improving or worsening, including health advisories or news unrelated to COVID-19; and (3) Negative (-1): News that conveys a negative sentiment towards the COVID-19 situation, indicating signs of deterioration, such as an increase in positive cases or a rise in the COVID-19 death toll.

The other deep learning models utilized for feature extraction include: (1) Convolutional Neural Network (CNN): This algorithm consists of multiple layers of interconnected neurons. The first layer extracts features from images or text, while subsequent layers process these extracted features to produce classifications (Lecun et al., 1998); (2) Long Short-Term Memory (LSTM): This algorithm was designed to analyze sequential data, such as text and audio, by incorporating a structure that retains information from previously encountered data (Hochreiter and Schmidhuber, 1997); and (3) Bidirectional Long Short-Term Memory (Bi-LSTM): A variant of LSTM, Bi-LSTM, was designed to process sequential data in both directions, left to right and right to left. This capability allows Bi-LSTM to capture contextual information from both past and future elements within a sequence (Graves et al., 2013).

The evaluation metrics used for evaluation include accuracy, precision, recall, F1-score, ROC-AUC as follows : (1) Accuracy refers to the proportion of correctly classified data points out of the total observations (Vakili et al., 2020), (2) Precision represented the proportion of predicted positive cases that were actually true positives (Powers, 2020), (3) Recall was the percentage of actual positive cases that were accurately identified as positive (Powers, 2020), (4) F1-Score represented the harmonic average of precision and recall (Sasaki, 2007) , and (5) ROC-AUC was determined from the ROC curve, which illustrated the relationship between the true positive rate and the false positive rate. The Area Under the ROC Curve (AUC) was utilized in binary classification to assess how effectively a model distinguished between positive and negative target classes (Vakili et al., 2020). In general, for machine learning modeling, a performance of ≥80% in terms of accuracy, precision, recall, F1-score, and ROC-AUC is considered acceptable (good). In one article, it is mentioned that a performance above 70% already meets industry standards. In general, for machine learning modeling, a performance of ≥80% in terms of accuracy, precision, recall, F1-score, and ROC-AUC is considered acceptable (good). In one article, it is mentioned that a performance above 70% already meets industry standards. (Fiddler, 2024; Zams, 2022)

3. METHODS

The research framework comprises several stages: data preprocessing, data splitting, text encoding, model training, and performance evaluation, as shown in Figure 1 (Khotimah et al., 2020). The data and experimental steps in this study are explained below.

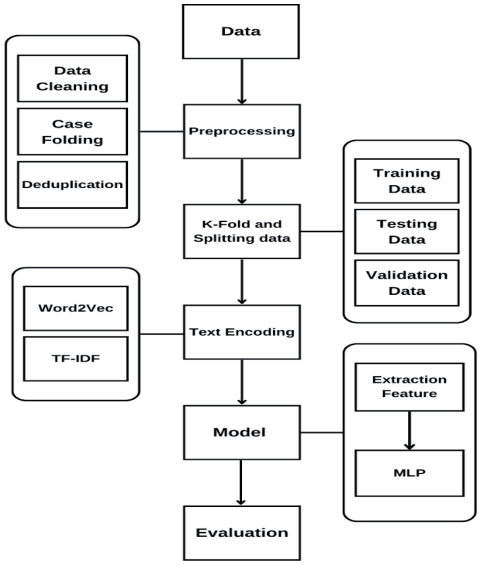

Figure 1. Research Framework
Source: Khotimah et al., (2020)

3.1 InaCOVED Data

This study utilized the Indonesian Corpus for COVID-19 Event Detection (InaCOVED) dataset from the National Research and Innovation Agency (BRIN) to conduct sentiment analysis on online news headlines reporting infectious disease events, specifically COVID-19. This dataset comprises 16,796 online news headlines collected from seven Indonesian news portals between January 26 and May 24, 2020. Online news headlines effectively represent the news for the sentiment classification process, as they typically capture the essence of the content.

As shown in Table 1, news articles on Indonesian news portals predominantly exhibit neutral sentiment, with 12,668 instances. Negative sentiment is represented by 3,136 instances, while positive sentiment accounts for 992 instances. The distribution of news categories from each portal, illustrated in Table 1, indicates that neutral news is more common than news with positive or negative sentiments.

Table 1. Online news distribution

News Portal	Sentiment		
	Positive	**Neutral**	**Negative**
Tirto	6	99	22
Republika	114	686	294
Tempo	96	1982	352
Merdeka	154	1993	407
Kompas	249	2029	793
Antara	177	2506	561
Detik	196	3373	707

Source: Data Processing by author (2024)

3.2 Data Preprocessing

The initial step before applying deep learning algorithms to text is text preprocessing. This process involves cleaning the data by removing tabs and punctuation marks, separating alphanumeric characters from numeric characters, and replacing repeated characters with a single unique character. Case folding is applied to convert all text to lowercase. Tokenization is then used to break sentences into individual words (tokens). Additionally, duplicate entries are removed to ensure a clean and ready-to-use dataset (Agustiningsih et al., 2021).

3.3 Data Splitting

The data was divided into training and testing sets using the Stratified K-Fold Cross-Validation (SKCV) method, ensuring that class distributions in each fold mirrored those of the original dataset. Stratified k-fold cross-validation, which uses a parameter 'k' to specify the number of folds, allowed each data point to be included in the training set (k-1) times, providing an unbiased assessment of model performance (Prusty et al., 2022). In k-Fold cross-validation, the commonly used values for k are 5 or 10 (Nti et al., 2021). In this study, k was set to 5, considering the distribution of training and testing data in each fold. This choice aims to ensure that the number of training and testing samples in each fold is not too small. During each iteration, one fold served as the testing set, while the remaining folds were used to train the model.

3.4 Text Encoding

Term Frequency-Inverse Document Frequency (TF-IDF) and Word Embedding (WE) methods were applied for text encoding to quantify word significance. TF-IDF and WE were used to convert text into numerical representations (Dessì et al., 2021). TF-IDF assigned weights based on each term's relevance within a set of documents (Hirsch and Hofer, 2022). WE, a language modelling technique, also maps words into a vector space, positioning words with similar meanings in closer proximity (Kilimci and Duvar, 2020). Word2Vec (Mikolov et al., 2013) often provided more informative features than TF-IDF (Sohrabi and Hemmatian, 2019) and generated more robust vector representations for neural network models (Jatnika et al., 2019).

This research used TF-IDF and WE with 17-dimensional vectors to independently train and test the MLP model. The vector dimensions of TF-IDF and WE were also increased to 17, 24, 50, and 100 for training and testing the MLP model with feature extraction from CNN, LSTM, and Bi-LSTM. Dimensions 17 and 24 were selected based on frequency analysis and word distribution within the dataset. In contrast, dimensions 50 and 100 were chosen to analyze trends and assess whether increasing the vector dimensions impacts performance scores.

3.5 Classifier Model

The model employed in this study for the sentiment classification task was a deep learning model, the Multi-Layer Perceptron (MLP), enhanced with feature extraction from other deep learning models. The MLP is a neural network of perceptrons, with neurons arranged in a hierarchical structure across interconnected layers. The architecture of the MLP begins with an input layer, passes through one or more hidden layers, and culminates in an output layer (Alboaneen et al., 2017). The MLP architecture is shown in Figure 2. The other deep learning models utilized for feature extraction (Jogin et al., 2018; Wu et al., 2019; Hameed and Garcia-Zapirain, 2020): CNN architecture is presented in Figure 3, the LSTM architecture is illustrated in Figure 4, and the BiLSTM architecture is illustrated in Figure 5.

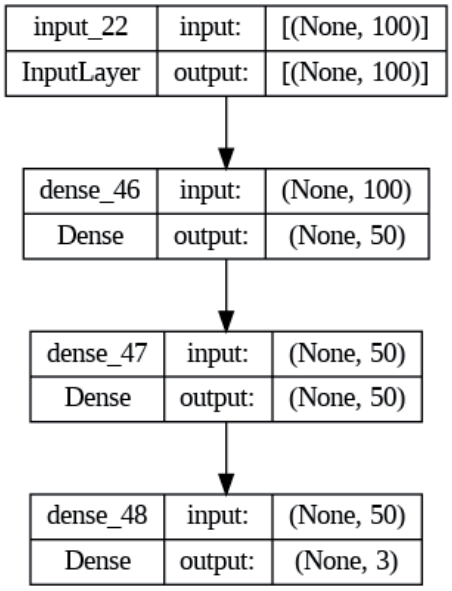

Figure 2. MLP Architecture
Source: Data Processing by author (2024)

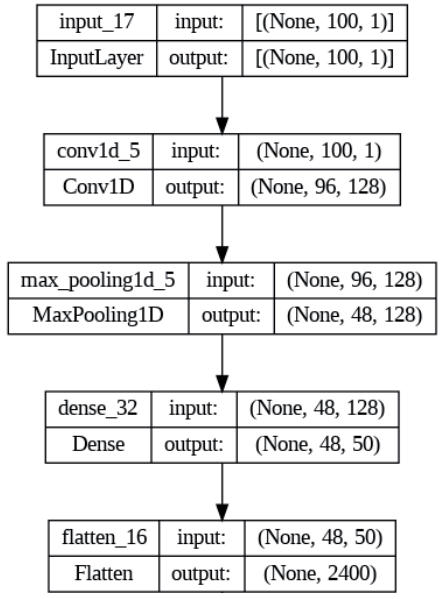

Figure 3. CNN Architecture
Source: Data Processing by author (2024)

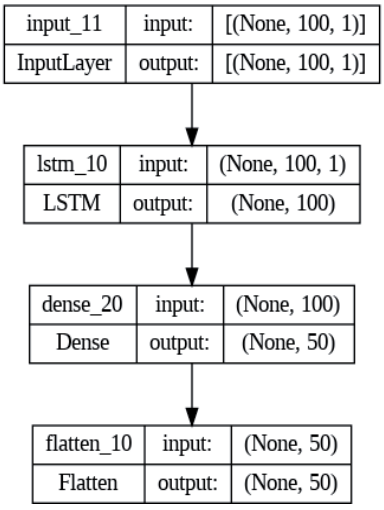

Figure 4. LSTM Architecture
Source: Data Processing by author (2024)

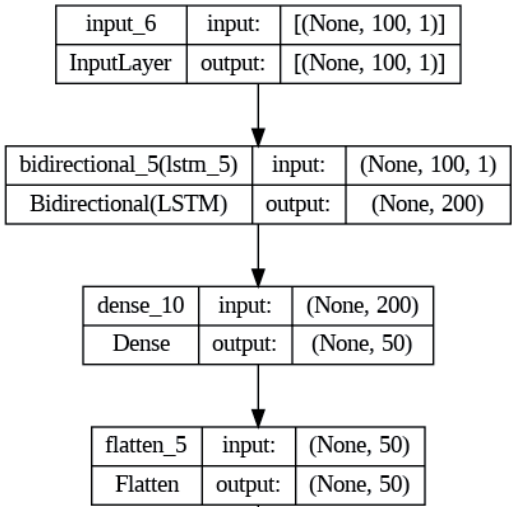

Figure 5. Bi-LSTM Architecture
Source: Data Processing by author (2024)

3.5 Evaluation

In this study, the performance of the MLP model combined with CNN, LSTM, and Bi-LSTM models for feature extraction was evaluated and compared to the performance of the MLP model alone. This combination is expected to enhance the performance of the MLP model by adding a feature extraction layer. The evaluation metrics used for evaluation include accuracy, precision, recall, F1-score, ROC-AUC (Naidu et al., 2023). The primary focus of the metrics was the delta value to assess the significance of the improvements. The delta value reflects the change in ROC-AUC before and after adding feature extraction. Performance improvement is generally considered significant at a value of 0.1%, though this threshold may vary depending on the context.

4. RESULT AND DISCUSSION

4.1 Results

Table 2 presents the evaluation results of the MLP performance without adding feature extraction layers, using TF-IDF and WE encoding with a vector dimension of 17. As explained in the previous section, the dataset used in this experiment was imbalanced, inevitably making the classification results less optimal. Nonetheless, classification with the MLP model using both TF-IDF and word embedding achieved a relatively high accuracy rate, exceeding 88%.

Table 2. MLP classification performance

Model	Sentiment				
	Accuracy	**Precision**	**Recall**	**F1-Score**	**ROC-AUC**
TF-IDF+MLP	88,63%	88,17%	88,63%	88,11%	88,54%
WE+MLP	88,53%	87,94%	88,53and	88,10%	92,79%

Source: Data Processing by author (2024)

Due to the imbalanced data, the performance improvement of the MLP model with added extraction features was assessed using the delta value. The delta represents the change in ROC-AUC between the MLP base model and the MLP model with added extraction features. This approach aimed to identify the best model based on its balance rather than solely on its accuracy. The vector dimensions of TF-IDF and WE were increased to 17, 24, 50, and 100 for each MLP model with feature extraction layers using CNN, LSTM, and Bi-LSTM.

The performance of the MLP model with added extraction features was detailed in Tables 3, 4, 5, 6, 7, and 8. Table 3 presents the performance evaluation of the MLP model combined with WE as the text encoder and CNN as an additional layer for feature extraction. Table 4 provides the evaluation results for the classification performance using the combined model of WE, LSTM, and MLP. Table 5 displays the performance evaluation for the model that combines word embedding, Bi-LSTM, and MLP. The classification performance for the TF-IDF, CNN, and MLP combined models is shown in Table 6. Meanwhile, Table 7 presents the evaluation results for the TF-IDF, LSTM, and MLP models. Lastly, Table 8 shows the evaluation results for the combined TF-IDF, Bi-LSTM, and MLP models.

Table 3. WE+CNN-MLP classification performance

Vector	Metrics Evaluation					
	Accuracy	**Precision**	**Recall**	**F1-Score**	**ROC-AUC**	**Delta**
17	88,90%	88,31%	88,90%	88,39%	92,91%	0,12%
24	89,11%	88,72%	89,11%	88,69%	93,11%	0,32%
50	88,86%	88,34%	88,86%	88,44%	92,97%	0,18%
100	89,78%	89,33%	89,78%	89,36%	93,30%	0,51%

Source: Data Processing by author (2024)

Table 4. WE+LSTM-MLP classification performance

Vector	Metrics Evaluation					
	Accuracy	***Precision***	***Recall***	***F1-Score***	***ROC-AUC***	***Delta***
17	88,35%	87,89%	88,35%	87,89%	92,41%	-0,38%
24	88,86%	88,21%	88,86%	88,26%	92,48%	-0,31%
50	89,02%	88,66%	89,02%	88,58%	92,38%	-0,41%
100	88,58%	88,10%	88,58%	88,09%	91,66%	-1,13%

Source: Data Processing by author (2024)

Table 5. WE+BilLSTM-MLP classification performance

Vector	Metrics Evaluation					
	Accuracy	**Precision**	**Recall**	**F1-Score**	**ROC-AUC**	**Delta**
17	88,17%	87,54%	88,17%	87,75%	92,36%	-0,43%
24	88,78%	88,31%	88,78%	88,14%	92,47%	-0,32%
50	88,42%	87,98%	88,42%	87,84%	92,24%	-0,55%
100	88,97%	88,69%	88,97%	88,61%	92,18%	-0,61%

Source: Data Processing by author (2024)

Table 6. TF-IDF+CNN-MLP classification performance

Vector	Metrics Evaluation					
	Accuracy	**Precision**	**Recall**	**F1-Score**	**ROC-AUC**	**Delta**
17	88,49%	88,07%	88,49%	87,97%	88,84%	0,30%
24	88,54%	88,03%	88,54%	87,89%	89,58%	1,04%
50	89,72%	89,39%	89,72%	89,37%	92,48%	3,94%
100	89,57%	89,12%	89,57%	89,19%	93,00%	4,46%

Source: Data Processing by author (2024)

Table 7. TF-IDF+-LSTM-MLP classification performance

Vector	Metrics Evaluation					
	Accuracy	**Precision**	**Recall**	**F1-Score**	**ROC-AUC**	**Delta**
17	88,41%	87,93%	88,41%	87,84%	88,45%	-0,09%
24	88,40%	87,90%	88,40%	87,81%	89,07%	0,53%
50	88,82%	88,37%	88,82%	88,27%	90,18%	1,64%
100	84,22%	83,33%	84,22%	82,61%	84,23%	-4,31%

Source: Data Processing by author (2024)

Table 8. TF-IDF+Bi-LSTM-MLP classification performance

Vector	Metrics Evaluation					
	Accuracy	**Precision**	**Recall**	**F1-Score**	**ROC-AUC**	**Delta**
17	88,42%	87,97%	88,42%	87,80%	88,61%	0,07%
24	88,50%	88,03%	88,50%	87,91%	89,13%	0,59%
50	88,28%	87,69%	88,28%	87,68%	89,99%	1,45%
100	87,13%	86,45%	87,13%	86,23%	88,17%	-0,37%

Source: Data Processing by author (2024)

4.2 Discussion

In sentiment analysis, the ideal values for accuracy, precision, recall, F1-score, and ROC-AUC depend on the context and dataset, requiring nuanced interpretation rather than absolute benchmarks. Accuracy can approach 95% under optimal conditions but may fall below 50% depending on the model and data characteristics, with contextual factors accounting for over 75% of variance (Hartmann et al., 2023). Precision and recall are vital for assessing the balance between true positives and false negatives, with advanced frameworks like NPSC leveraging techniques such as BERT and BiLSTM to enhance these metrics (Ramana et al., 2025). The F1-score, critical for imbalanced datasets, balances precision and recall, while ROC-AUC assesses the model's class distinction capabilities. These metrics collectively offer a robust framework for evaluation, yet their ideal values are inherently application-specific and data-dependent, emphasizing the need for contextualized assessment (Sokolova et al., 2006).

Based on Table 2, the performance results of the MLP base model for the sentiment classification task were observed. It was evident that whether using TF-IDF or WE, the MLP base model achieved an average performance score of 88%. However, as anticipated, the MLP with WE attained a relatively high ROC-AUC score of 92.79%. Nonetheless, this was not the best result, as the data used were imbalanced. According to Ningsih et al. (2022), deep learning models provide better and more stable performance when utilizing balanced data compared to imbalanced data.

Based on the results presented in Tables 3, 4, and 5, which showed the performance metrics of the MLP model with extraction features, it was observed that CNN-MLP achieved the optimal performance for WE + MLP with extraction features with a vector dimension of 100. This combination attained an ROC-AUC score of 93.30%, indicating a performance improvement of 0.51%. These findings align with prior research (Tan et al., 2022; Munandar et al., 2021; Mohbhey et al., 2024) on sentiment analysis across various datasets, reinforcing that adding feature extraction layers boosts model performance. Although this increase may seem modest when viewed purely in numerical terms, it was likely due to the baseline MLP model with WE already producing optimal results compared to those using TF-IDF. These findings were consistent with previous research by Dang et al. (2020), Ningsih et al. (2022), and Khotimah et al. (2023), which also demonstrated that using WE yields higher performance compared to TF-IDF.

Based on Tables 6, 7, and 8, the CNN-MLP combination exhibited the most significant improvement for the TF-IDF + MLP model. With a vector size of 100, this model increased the ROC-AUC score by a substantial 4.46%, reaching 93.00%. Even with a vector size of 50, the CNN-MLP model increased by 3.94%, just 0.5% less than with a vector size 100. While the ROC-AUC score improved observably, the increase across the other performance metrics was not as significant. As shown in Tables 3, 4, and 5, the performance metrics did not surpass the 90% threshold. This could be attributed to the highly imbalanced nature of the dataset, where the disparity between the most frequent and least frequent classes was too large. When analyzed per class, the performance of the minority class lagged significantly behind that of the majority class.

Convolutional Neural Networks (CNNs) outperform Long Short-Term Memory (LSTM) and Bidirectional LSTM (BiLSTM) models in various tasks due to their computational efficiency and early predictive power. CNNs excel in applications requiring rapid processing, such as process outcome prediction, named entity recognition (NER), and time series forecasting. For instance, CNNs increase training time by only 25% in NER tasks compared to word-based models, whereas LSTMs more than double the training time (Zhai et al., 2018).

A negative ROC-AUC delta indicates a decline in the performance of a binary classifier when comparing two models or the same model under different conditions, reflecting reduced ability to distinguish between positive and negative instances. This metric is essential for assessing relative classifier performance, highlighting the potential inefficacy of a newer model or condition compared to the baseline due to factors such as overfitting, data imbalance, or inappropriate model complexity. Overfitting occurs when an overly complex model performs well on training data but poorly on unseen data (Zhou and Skiena, 2023). Data imbalance can distort ROC-AUC as it may not adequately represent classifier performance across all classes (Carrington et al., 2023). Additionally, threshold sensitivity can influence the delta since ROC-AUC evaluates all thresholds, and performance variations at specific thresholds can significantly affect results (Muschelli, 2020). In this study, the proposed method successfully increased the ROC-AUC value by up to 4.46% in TF-IDF+CNN-MLP hybridization (from 88.54% with TF-IDF+MLP to 93.00% with TF-IDF+CNN-MLP).

5. CONCLUSION

This study evaluated the performance of the MLP model combined with CNN, LSTM, and Bi-LSTM models for feature extraction, comparing them with the standalone MLP model. TF-IDF and WE were applied to convert text into numerical representations, and the data was divided into training and testing sets using the SKCV method. Experimental results indicated that enhancing the MLP model with neural network feature extraction layers significantly improved performance, as measured by ROC-AUC. Among the tested configurations, CNN emerged as the most compelling feature extraction layer for MLP. Specifically, the WE + CNN-MLP model achieved a ROC-AUC score of 93.30%, and the TF-IDF + CNN-MLP model reached 93.00%. These results indicate that incorporating feature extraction layers enhances model performance. Additionally, using WE as a text encoder yields relatively better results than TF-IDF.

This study contributes to developing a text classification model using deep learning, specifically the Multi-Layer Perceptron (MLP), to achieve greater accuracy, effectiveness, and efficiency in handling large and complex datasets. The limitation of this study was that it relied on a single dataset, while different types of data could influence classification results. Future work could build upon these results by experimenting with various deep learning models as base architectures, exploring alternative neural networks for feature extraction, and utilizing a more comprehensive range of datasets.

ACKNOWLEDGEMENT

We would like to express our sincere gratitude to everyone who has contributed to the completion of this work. Special thanks to our colleagues and collaborators for their insightful discussions and technical assistance. Finally, we are deeply grateful to our institutions for providing the necessary resources and support.

CREDIT (CONTRIBUTOR ROLES TAXONOMY)

Fikri: Writing-Original draft preparation, Software. **Nuryaman:** Supervision, Methodology. **Khotimah:** Conceptualization, Supervision. **Parlina:** Writing-Reviewing and Editing. **Sihombing:** Validation.

REFERENCES

Agustiningsih, K. K., Utami, E., & Fatta, H. Al. (2021). Sentiment Analysis of COVID-19 Vaccine on Twitter Social Media: Systematic Literature Review. *2021 IEEE 5th International Conference on Information Technology, Information Systems and Electrical Engineering (ICITISEE)*, 121–126. https://doi.org/10.1109/ICITISEE53823.2021.9655960

Akhtar, M. S., Kumar, A., Ghosal, D., Ekbal, A., & Bhattacharyya, P. (2017). A Multilayer perceptron based ensemble technique for fine-grained financial sentiment analysis. *EMNLP 2017 - Conference on Empirical Methods in Natural Language Processing, Proceedings*, 540–546. https://doi.org/10.18653/v1/d17-1057

Alboaneen, D. A., Tianfield, H., & Zhang, Y. (2017). Sentiment analysis via multi-layer perceptron trained by meta-heuristic optimisation. *Proceedings - 2017 IEEE International Conference on Big Data, Big Data 2017*, *2018-Janua*, 4630–4635. https://doi.org/10.1109/BigData.2017.8258507

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. https://doi.org/10.21597/jist.1292050

Cahyo, G. A. D., Khotimah, P. H., Rozie, A. F., Nugraheni, E., Arisal, A., & Nuryaman, A. (2024). CNN-Based Hybrid Performance Evaluation Towards Online News Sentiment Classification Task. *2024 International Conference on Computer, Control, Informatics and Its Applications (IC3INA)*, 349–354.

Carrington, A. M., Manuel, D. G., Fieguth, P. W., Ramsay, T., Osmani, V., Wernly, B., Bennett, C., Hawken, S., Magwood, O., Sheikh, Y., Mcinnes, M., & Holzinger, A. (2023). Deep ROC Analysis and AUC as Balanced Average Accuracy, for Improved Classifier Selection, Audit and Explanation. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, *45*(1), 329–341. https://doi.org/10.1109/TPAMI.2022.3145392

Dang, N. C., Moreno-Garcia, M. N., & Prieta, F. D. la. (2020). Sentiment Analysis Based on Deep Learning: A Comparative Study. *Electronics (Switzerland)*, *9*(3), 483. https://doi.org/10.3390/electronics9030483

Dessì, D., Helaoui, R., Kumar, V., Recupero, D. R., & Riboni, D. (2021). TF-IDF vs word embeddings for morbidity identification in clinical notes: An initial study. In *arXiv preprint arXiv:2105.09632*. https://doi.org/10.48550/arXiv.2105.09632

Devlin, J., Chang, M.-W., Lee, K., Google, K. T., & Language, A. I. (2018). BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. *Naacl-Hlt 2019*, *Mlm*, 4171–4186. https://aclanthology.org/N19-1423.pdf

Fiddler. (2024). *Which is more important: model performance or model accuracy?* https://www.fiddler.ai/model-accuracy-vs-model-performance/which-is-more-important-model-performance-or-model-accuracy

Graves, A., Jaitly, N., & Mohamed, A. (2013). Hybrid speech recognition with Deep Bidirectional LSTM. *2013 IEEE Workshop on Automatic Speech Recognition and Understanding*, 273–278. https://doi.org/10.1109/ASRU.2013.6707742

Hameed, Z., & Garcia-Zapirain, B. (2020). Sentiment Classification Using a Single-Layered BiLSTM Model. *IEEE Access*, *8*, 73992–74001. https://doi.org/10.1109/ACCESS.2020.2988550

Hartmann, J., Heitmann, M., Siebert, C., & Schamp, C. (2023). More than a Feeling: Accuracy and Application of Sentiment Analysis. *International Journal of Research in Marketing*, *40*(1), 75–87. https://doi.org/10.1016/j.ijresmar.2022.05.005

Hirsch, T., & Hofer, B. (2022). Using textual bug reports to predict the fault category of software bugs. *Array*, *15*, 100189.

Hochreiter, S., & Schmidhuber, J. (1997). Long Short-Term Memory. *Neural Computation*, *9*(8), 1735–1780. https://doi.org/10.1162/neco.1997.9.8.1735

Jatnika, D., Bijaksana, M. A., & Suryani, A. A. (2019). Word2vec model analysis for semantic similarities in English words. *Procedia Computer Science*, *157*, 160–167. https://doi.org/10.1016/j.procs.2019.08.153

Jogin, M., Mohana, Madhulika, M. S., Divya, G. D., Meghana, R. K., & Apoorva, S. (2018). Feature Extraction using Convolution Neural Networks (CNN) and Deep Learning. *2018 3rd IEEE International Conference on Recent Trends in Electronics, Information & Communication Technology (RTEICT)*, 2319–2323. https://doi.org/10.1109/RTEICT42901.2018.9012507

Khotimah, P. H., Arisal, A., Rozie, A. F., Nugraheni, E., Riswantini, D., Suwarningsih, W., Munandar, D., & Purwarianti, A. (2023). Monitoring Indonesian online news for COVID-19 event detection using deep learning. *International Journal of Electrical and Computer Engineering*, *13*(1), 957–971. https://doi.org/10.11591/ijece.v13i1.pp957-971

Khotimah, P. H., Fachrur Rozie, A., Nugraheni, E., Arisal, A., Suwarningsih, W., & Purwarianti, A. (2020). Deep Learning for Dengue Fever Event Detection Using Online News. *Proceeding - 2020 International Conference on Radar, Antenna, Microwave, Electronics and Telecommunications, ICRAMET 2020*, 261–266. https://doi.org/10.1109/ICRAMET51080.2020.9298630

Kilimci, Z. H., & Duvar, R. (2020). An efficient word embedding and deep learning based model to forecast the direction of stock exchange market using twitter and financial news sites: A case of istanbul stock exchange (BIST 100). *IEEE Access*, *8*(Bist 100), 188186–188198. https://doi.org/10.1109/ACCESS.2020.3029860

Lecun, Y., Bottou, L., Bengio, Y., & Haffner, P. (1998). Gradient-Based Learning Applied to Document Recognition. *Proceedings of the IEEE*, *86*(11), 2278–2324. https://doi.org/10.1109/5.726791

Mikolov, T., Chen, K., Corrado, G., & Dean, J. (2013). Efficient estimation of word representations in vector space. *1st International Conference on Learning Representations, ICLR 2013 - Workshop Track Proceedings*, 1–12.

Mohbey, K. K., Meena, G., Kumar, S., & Lokesh, K. (2024). A CNN-LSTM-Based Hybrid Deep Learning Approach for Sentiment Analysis on Monkeypox Tweets. *New Generation Computing*, *42*(1), 89–107. https://doi.org/10.1007/s00354-023-00227-0

Munandar, D., Rozie, A. F., & Arisal, A. (2021). A multi domains short message sentiment classification using hybrid neural network architecture. *Bulletin of Electrical Engineering and Informatics*, *10*(4), 2181–2191. https://doi.org/10.11591/EEI.V10I4.2790

Muschelli, J. (2020). ROC and AUC with a Binary Predictor: a Potentially Misleading Metric. *Journal of Classification*, *37*(3), 696–708. http://dx.doi.org/10.1007/s00357-019-09345-1

Naidu, G., Zuva, T., & Sibanda, E. M. (2023). *A Review of Evaluation Metrics in Machine Learning Algorithms BT - Artificial Intelligence Application in Networks and Systems* (R. Silhavy & P. Silhavy (eds.); pp. 15–25). Springer International Publishing.

Nandwani, P., & Verma, R. (2021). A review on sentiment analysis and emotion detection from text. *Social Network Analysis and Mining*, *11*(1). https://doi.org/10.1007/s13278-021-00776-6

Naskath, J., Sivakamasundari, G., & Begum, A. A. S. (2023). A Study on Different Deep Learning Algorithms Used in Deep Neural Nets: MLP SOM and DBN. *Wireless Personal Communications*, *128*(4), 2913–2936. https://doi.org/10.1007/s11277-022-10079-4

Ningsih, F. S. S., Khotimah, P. H., Arisal, A., Rozie, A. F., Munandar, D., Riswantini, D., Nugraheni, E., Suwarningsih, W., & Kurniasari, D. (2022). Synonym-based Text Generation in Restructuring Imbalanced Dataset for Deep Learning Models. *2022 5th International Conference on Networking, Information Systems and Security: Envisage Intelligent Systems in 5g//6G-Based Interconnected Digital Worlds (NISS)*, 1–6. https://doi.org/10.1109/NISS55057.2022.10085156

Nti, I. K., Nyarko-Boateng, O., & Aning, J. (2021). Performance of Machine Learning Algorithms with Different K Values in K-fold CrossValidation. *International Journal of Information Technology and Computer Science*, *13*(6), 61–71. https://doi.org/10.5815/ijitcs.2021.06.05

Powers, D. M. W. (2020). Evaluation: from precision, recall and F-measure to ROC, informedness, markedness and correlation. *ArXiv Preprint ArXiv:2010.16061*, 37–63. http://arxiv.org/abs/2010.16061

Prusty, S., Patnaik, S., & Dash, S. K. (2022). SKCV: Stratified K-fold cross-validation on ML classifiers for predicting cervical cancer. *Frontiers in Nanotechnology*, *4*(August), 1–12. https://doi.org/10.3389/fnano.2022.972421

Ramana, A. V., Prasad, K. S. N., & Rao, A. S. (2025). Nonlinear Deep Learning Framework for Precise Sentiment Classification in Textual Data. *Communications on Applied Nonlinear Analysis*, *32*(1S), 573–584. https://doi.org/10.52783/cana.v32.2345

Ramchoun, H., Amine, M., Idrissi, J., Ghanou, Y., & Ettaouil, M. (2016). Multilayer Perceptron: Architecture Optimization and Training. *International Journal of Interactive Multimedia and Artificial Intelligence*, *4*(1), 26. https://doi.org/10.9781/ijimai.2016.415

Salur, M. U., & Aydin, I. (2020). A Novel Hybrid Deep Learning Model for Sentiment Classification. *IEEE Access*, *8*, 58080–58093. https://doi.org/10.1109/ACCESS.2020.2982538

Sasaki, Y. (2007). *The truth of the F-measure*. http://www.cs.odu.edu/~mukka/cs795sum09dm/Lecturenotes/Day3/F-measure-YS-26Oct07.pdf

Shaker, K. (2022). Optimizing Sentiment Big Data Classification Using Multilayer Perceptron. *Anbar Journal of Engineering Sciences*, *13*(2), 14–21. https://doi.org/10.37649/aengs.2022.176353

Shamrat, F. M. J. M., Chakraborty, S., Imran, M. M., Muna, J. N., Billah, M. M., Das, P., & Rahman, M. O. (2021). Sentiment analysis on twitter tweets about COVID-19 vaccines using NLP and supervised KNN classification algorithm. *Indonesian Journal of Electrical Engineering and Computer Science*, *23*(1), 463–470. https://doi.org/10.11591/ijeecs.v23.i1.pp463-470

Socher, R., Perelygin, A., Wu, J. Y., Chuang, J., Manning, C. D., Ng, A. Y., & Potts, C. (2013). Recursive Deep Models for Semantic Compositionality Over a Sentiment Treebank. *Proceeding of the 2013 Conference on Empirical Methods in Natural Language Processing*, *October*, 1631–1642.

Sohrabi, M. K., & Hemmatian, F. (2019). An efficient preprocessing method for supervised sentiment analysis by converting sentences to numerical vectors: a twitter case study. *Multimedia Tools and Applications*, *78*(17), 24863–24882. https://doi.org/10.1007/s11042-019-7586-4

Sokolova, M., Japkowicz, N., & Szpakowicz, S. (2006). *Beyond Accuracy, F-Score and ROC: A Family of Discriminant Measures for Performance Evaluation BT - AI 2006: Advances in Artificial Intelligence* (A. Sattar & B. Kang (eds.); pp. 1015–1021). Springer Berlin Heidelberg.

Tan, K. L., Lee, C. P., Anbananthen, K. S. M., & Lim, K. M. (2022). RoBERTa-LSTM: A Hybrid Model for Sentiment Analysis With Transformer and Recurrent Neural Network. *IEEE Access*, *10*, 21517–21525. https://doi.org/10.1109/ACCESS.2022.3152828

Tan, K. L., Lee, C. P., & Lim, K. M. (2023). A Survey of Sentiment Analysis: Approaches, Datasets, and Future Research. *Applied Sciences (Switzerland)*, *13*(7). https://doi.org/10.3390/app13074550

The NYU School of Professional Studies. (2024). *What is Sentiment Classification?* https://www.aimasterclass.com/glossary/sentiment-classification

Vakili, M., Ghamsari, M., & Rezaei, M. (2020). Performance Analysis and Comparison of Machine and Deep Learning Algorithms for IoT Data Classification. *ArXiv Preprint*. http://arxiv.org/abs/2001.09636

Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., Kaiser, Ł., & Polosukhin, I. (2017). Attention is all you need. *Advances in Neural Information Processing Systems*, *2017-Decem*(Nips), 5999–6009.

Wankhade, M., Rao, A. C. S., & Kulkarni, C. (2022). A survey on sentiment analysis methods, applications, and challenges. In *Artificial Intelligence Review* (Vol. 55, Issue 7). Springer Netherlands. https://doi.org/10.1007/s10462-022-10144-1

Wu, Y. X., Wu, Q. B., & Zhu, J. Q. (2019). Data-driven wind speed forecasting using deep feature extraction and LSTM. *IET Renewable Power Generation*, *13*(12), 2062–2069. https://doi.org/10.1049/iet-rpg.2018.5917

Zams. (2022). *How To Know if Your Machine Learning Model Has Good Performance | Obviously AI*. Obviously. Ai. https://www.obviously.ai/post/machine-learning-model-performance

Zhai, Z., Nguyen, D. Q., & Verspoor, K. (2018). Comparing CNN and LSTM character-level embeddings in BiLSTM-CRF models for chemical and disease named entity recognition. *EMNLP 2018 - 9th International Workshop on Health Text Mining and Information Analysis, LOUHI 2018 - Proceedings of the Workshop*, 38–43. https://doi.org/10.18653/v1/w18-5605

Zhou, B., & Skiena, S. (2023). Does It Pay to Optimize AUC? *Proceedings of the 37th AAAI Conference on Artificial Intelligence, AAAI 2023*, *37*, 11408–11416. https://doi.org/10.1609/aaai.v37i9.26349