Jurnal Teknik Informatika Vol. 18 No. 2, October 2025 (196-208)
ISSN: p-ISSN 1979-9160 (Print)| e-ISSN 2549-7901 (Online)
DOI: https://doi.org/10.15408/jti.v18i2.42459

JURNAL TEKNIK INFORMATIKA

Homepage : http://journal.uinjkt.ac.id/index.php/ti

Evaluating BiLSTM Performance with BERT, RoBERTa, and DistilBERT in Online Bullying News Detection

Moh. Rosidi Zamroni[1*], Miftahus Sholihin[2], Erna Hayati[3], Rahayu A Hamid[4], Nurul Aswa Omar[5]

[1,2]Informatics Engineering, Faculty of Science and Technology, Lamongan Islamic University
[3]Accounting (Statistics), Faculty of Economics and Business, Lamongan Islamic University
[4,5]Informatics Engineering, Faculty of Science and Technology, Tun Hussein Onn University Malaysia
[1,2,3] Jl. Veteran No. 53A Lamongan-East Java 62214, Indonesia
[4,5] Parit Raja, 86400, Malaysia

Article:
Accepted: August 22, 2025
Revised: February 13, 2025
Issued: October 30, 2025

***Correspondence Address:**
rosidizamroni@unisla.ac.id

ABSTRACT

This study examines the performance of BiLSTM combined with three transformer-based word embeddings—BERT, RoBERTa, and DistilBERT—in classifying bullying news in online media. BiLSTM was chosen for its significant advantages in processing text sequences compared to traditional RNN and LSTM models. The study used a dataset of 2,800 articles from three major Indonesian news portals, with 2,000 articles for training and 800 for testing, labeled using the lexicon method. The testing results showed that the combination of BiLSTM and RoBERTa achieved the best performance, with an accuracy of 94% and a near-perfect precision of 99%. Statistical significance tests confirmed that BiLSTM with RoBERTa performs significantly better than with BERT or DistilBERT. These findings suggest that the BiLSTM and RoBERTa combination is the most effective for classifying bullying news, especially for new or unseen data. This research contributes to the development of automatic bullying content detection systems to enhance content moderation on news platforms.

Keywords : *bullying; news classification; word embedding; BiLSTM; NLP.*

Jurnal Teknik Informatika Vol. 18 No. 2, October 2025 (196-208)
ISSN: p-ISSN 1979-9160 (Print)| e-ISSN 2549-7901 (Online)
DOI: https://doi.org/10.15408/jti.v18i2.42459

1. INTRODUCTION

The increase in the use of online media has led to the widespread spread of bullying-related content, which can have a negative impact on individuals and society. Bullying, defined as intentional harassment, has the potential to damage mental and emotional health [1]. In addition to happening in the real world, this phenomenon is also often found in the digital space, including in online news that often contains explicit or implicit bullying narratives [2], [3], [4]. News that focuses on bullying can reinforce stereotypes, spark conflict, and cause psychological trauma to victims [5]. Uncontrolled content can even make harmful behavior look normal, encouraging the audience to adopt the mindset or action presented. Therefore, automatically detecting and classifying bullying news content is an important step to reduce its negative impact and support more ethical content moderation [6].

Various previous studies have explored Natural Language Processing (NLP) techniques for detecting bullying content, including transformer-based embedding methods such as BERT, RoBERTa, and DistilBERT, which show promising potential in understanding the context of text in depth. For example, research by Nemkul [7] showed that RoBERTa achieved an accuracy of 95.3%, higher than BERT, which reached 93.3% in the Nepali-language news classification. In addition, Pratima's research [8] showed that the combination of BERT and BiLSTM was able to achieve up to 96.8% accuracy on social and political news classifications, indicating that BiLSTM can capture sequential information that is important in understanding complex sentence structures.

BiLSTM (Bidirectional Long Short-Term Memory) is a development of LSTM that allows the model to understand the context of words from two directions, both forward and backward [9]. This is especially useful in the analysis of news texts that have complex sentence structures and interrelated inter-word contexts.

Although several studies have explored the use of transformer-based embeddings for sentiment analysis and news classification, the effectiveness of the combination of BERT, RoBERTa, and DistilBERT with BiLSTM in detecting bullying news content is still rarely studied. Most studies focus on only one type of embedding without considering how this combination can improve understanding of context and word order in the text [10]. In addition, An's research [11], which uses the ALBERT-BiLSTM model, only focuses on entity extraction without reviewing the specific classification of bullying content.

This study aims to compare the performance of BERT, RoBERTa, and DistilBERT combined with BiLSTM in the classification of bullying news. Key contributions from this study include:

1. Provides an in-depth analysis of the effectiveness of transformer-based embedding when combined with BiLSTM in handling complex news texts.
2. Identify the advantages and limitations of each model in terms of classification accuracy and computational efficiency.
3. Provides new insights for the development of an automatic detection system for news platforms, so that it can strengthen the moderation of negative content more efficiently and ethically.

The findings of this study are expected to be the basis for the development of an AI-based content moderation system that is more effective in detecting and reducing the spread of bullying news on digital platforms.

2. METHODS

The method carried out in this study includes several stages, as seen in the following flowchart:

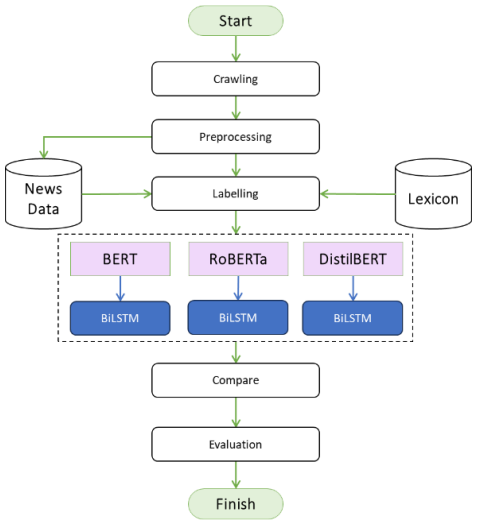

Figure 1. *Flowchart of the research process*

The stages of the flowchart in Figure 1. It will be described in the following explanation:

2.1. Data Collection (crawling)

The data for this study was taken from three online news portals, namely Kompas.com, iNews, and Detik.com. Articles from all three sources were collected through a crawling process, resulting in a total of 2800 articles. Of these, 2000 articles were used to train the model, while another 800 articles were allocated for testing.

2.2. Pre-Processing

Processing online news data before it is applied to the model. This stage aims to convert the raw data into a cleaner and more structured format so that it can be processed effectively [12]. The preprocessing process implemented consists of the following steps [13]:

a. Punctuation and Symbol Removal: performs the removal of punctuation, numbers, and non-alphabetic characters.
b. Case Folding: The entire text is converted to lowercase letters to ensure consistency in data processing.
c. Tokenization: This process separates sentences into individual words (tokens).
d. Stopword Removal: Stopwords such as *"dan", "yang", "di",* and *"ke"* often do not have significant meaning in the analysis. Therefore, the stopword is removed to reduce the data dimension and focus on more relevant words.

2.3. Data Labeling

Each article is labeled using the lexicon method, which is by recognizing words related to the domain (bullying) [14], [15]. The lexicon method was chosen for text labeling because of several advantages it offers. One of the advantages of lexicon is its ability to reduce subjectivity in labeling, since the marking process is carried out automatically by the system based on the available dictionary of opinion words, so it does not depend on individual interpretations[16]. In addition, this method allows for handling large amounts of data, such as data from social media, without the need for time-consuming manual labeling [17]. Using a predefined dictionary and its weights, the lexicon method can automatically identify and label text based on the occurrence of certain words[18]. This approach also makes it easier to interpret the labeling results because decisions are made based on specific words contained in the dictionary, making it suitable for use in research that requires accuracy and consistency in data tagging, such as the detection of bullying content in online news. Helps shorten labeling time when compared to manually labeling large amounts of data. Some of the keywords used for identification include: *'bullying', 'perundungan', 'intimidasi', 'penghinaan', 'pelecehan', 'mengejek', 'menghina', 'merendahkan', 'mengolok-olok', 'menindas', 'ancaman', 'fitnah', 'kekerasan verbal', 'kekerasan fisik', 'pengucilan', 'mengasingkan', 'penganiayaan', 'mencaci maki', 'memfitnah', 'mem-bully', 'perlakuan kasar', 'suka mengancam', 'pencemaran nama baik', 'mengintimidasi', 'perlakuan diskriminatif', 'menyebarkan gosip', 'menyingkirkan teman', ' membentak', 'memaksa', 'menghujat'*. Based on this method, as many as 852 articles were marked with the label of bullying, while the other 1243 articles were marked as non-bullying.

2.4. Modelling (BiLSTM)

BiLSTM was chosen because it has a significant advantage in handling text sequences compared to traditional RNNs and LSTMs. Although RNNs are suitable for processing text of variable length, the model faces serious problems such as gradient explosion, gradient disappearance, and difficulty handling remote dependencies. LSTM has successfully overcome this problem with a gate control mechanism, but it can only process information in one direction, thus limiting its ability to understand the context thoroughly. To overcome this limitation, BiLSTM was introduced by adding a second hidden layer that allows the flow of information in both directions, both forward and backward. This approach allows BiLSTM to understand contexts from the past and the future simultaneously, making it highly effective in sequence modeling tasks, especially in understanding complex contexts in texts [19]. Therefore, the use of BiLSTM is superior in providing better accuracy and deeper insights compared to ordinary LSTM.

Jurnal Teknik Informatika Vol. 18 No. 2, October 2025 (196-208)
ISSN: p-ISSN 1979-9160 (Print)| e-ISSN 2549-7901 (Online)
DOI: https://doi.org/10.15408/jti.v18i2.42459

Before the classification process with BiLSTM, data needs to be transformed using Transformer-based techniques such as BERT, RoBERTa, and DistilBERT. These techniques include embedding layers to represent each word numerically and vectorally by converting the root word into a vector form, which is part of feature extraction [20]. Transformer-based embeddings play an important role in sentiment analysis by strengthening the representation of the text to improve classification [21]. There are many types of embeddings to support text classification models.

Three transformer models are used in this process, namely:

a. BERT, introduced in 2018 by Jacob Devlin and the Google team [22], is a bidirectional language representation model that understands words based on the context from both sides [23], [24]. Trained on a vast set of freely available texts, BERT has shown exceptional performance in a wide range of natural language processing tasks, including text classification [25].
b. RoBERTa, introduced by Facebook AI in 2019, is an optimized version of BERT. It improves performance by modifying the training procedure, including longer training, more data, larger batches, and removing the next sentence prediction objective [26], [27].
c. DistilBERT is a lighter and more efficient version of BERT, designed to reduce training time without significantly compromising performance. Studies by Barbon and Akabane versatile [28], [29] highlight its effectiveness in text classification and efficiency in handling large datasets.

Each of these transformer models is followed by a BiLSTM layer that allows bidirectional text processing (forward and backward) to better understand the context. The architecture of this model can be seen in figure 2.

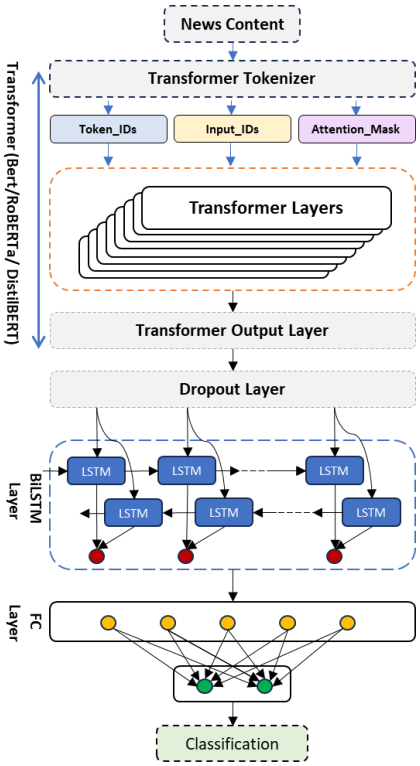

Figure 2. *BiLSTM models with each transformer layer*

2.5. Compare

After the three models (BERT + BiLSTM, RoBERTa + BiLSTM, and DistilBERT + BiLSTM) were trained and tested, the results of each model were compared to see how they performed.

2.6. Evaluation

At this stage, evaluations are performed to measure the performance of each model using evaluation metrics such as accuracy, precision, and recall to determine which model provides the best results [30].

$$Accuracy = \frac{TP+TN}{TP+TN+FP+FN} \quad (1)$$

$$Precision = \frac{TP}{TP+FP} \quad (2)$$

$$Recall = \frac{TP}{TP+FN} \quad (3)$$

In addition, the model will be evaluated with the Receiver Operating Characteristic (ROC) curve. The ROC curve serves as a metric to evaluate the performance of the classification model. The sharpness of the curve is associated with the model's classification ability and error rate. A sharper ROC curve usually represents a model that performs well. In this case, the model can demonstrate high accuracy, a low error rate, and proficient classification ability.

This implies that the model's predictions provide a clearer distinction between the actual classes. Conversely, a less sharp ROC curve can indicate lower classification performance for the model or a higher number of classification errors in the confusion matrix [31].

To test the significance of accuracy, a one-way ANOVA was conducted to determine whether there was a significant difference between the average accuracy of the three models tested[32]. In this test, the hypothesis used is:

a. H_0 (Null hypothesis): There is no significant difference between the accuracy of the three models.
b. H_1 (Alternative hypothesis): There is at least one model that has a significant difference in accuracy compared to the others.

The test was carried out by taking accuracy results from each model in 5 experiments. F-statistic and p-value values are obtained to evaluate hypotheses. If the p-value < 0.05, then the null hypothesis is rejected, which means that there is a significant difference between the models tested.

If the ANOVA results show significant differences, further tests such as Tukey HSD or post-hoc carried out to determine which model pairs have significant accuracy differences. The results of this test help in selecting the best model based on statistical significance, not just based on the average accuracy value.

With this approach, the model evaluation becomes more statistically valid, so it can provide a strong justification in selecting the best model for the classification of bullying news in online media [33].

3. RESULTS AND DISCUSSION

In this discussion, the performance of the model used for each combination will be explained. This stage will be divided into two sub-discussions, namely taining and testing.

3.1. Training

This training process uses 2095 data with an equalized model structure, namely using two layers of BiLSTM with memory units of 128 and 46, two layers of Dropout with a rate of 0.5 each, and the Adam Optimizer with a learning rate of 0.0001. The training was carried out with 50 epochs and a batch size of 64.

3.1.1. BiLSTM+ BERT Combination

The results of the model are shown in Figure 3. using the Bidirectional LSTM architecture (BiLSTM), which utilizes the embedding of BERT as the representation of the input text. BERT generates an embedding that considers the context of the word in depth, and the model uses a representation token [CLS] from BERT that represents the overall meaning of the text. This embedding is fed to the first layer of bidirectional LSTM, which has 256 units. The use of BiLSTM allows the model to better understand the context in the text sequence, as it processes data in both directions, both forward and backward. After the first BiLSTM, there is a Dropout layer that reduces the chance of overfitting by disabling a random portion of nodes during training.

Model: "sequential_1"

Layer (type)	Output Shape	Param #
bidirectional (Bidirectional)	(None, 1, 256)	918,528
dropout (Dropout)	(None, 1, 256)	0
bidirectional_1 (Bidirectional)	(None, 92)	111,504
dropout_1 (Dropout)	(None, 92)	0
dense (Dense)	(None, 1)	93

Total params: 1,030,125 (3.93 MB)
Trainable params: 1,030,125 (3.93 MB)
Non-trainable params: 0 (0.00 B)

Figure 3. *Results of the BiLSTM Model with BERT Embeding*

Then, the results from the first BiLSTM layer were passed on to the second Bidirectional LSTM layer with 92 units. This second layer helps to enrich the representation of the text more deeply. After that, a second Dropout layer is applied to increase the generalization. Finally, the Dense layer with a single unit and a sigmoid activation function produces a final output in the form of probabilities for binary classification. The model has approximately 1,030,125 trained parameters, with all parameters being updated during training to optimize classification performance. This structure allows the model to leverage the context in the text as well as the dynamic features of the BERT to achieve accurate results in the classification of news or other text.

Jurnal Teknik Informatika Vol. 18 No. 2, October 2025 (196-208)
ISSN: p-ISSN 1979-9160 (Print)| e-ISSN 2549-7901 (Online)
DOI: https://doi.org/10.15408/jti.v18i2.42459

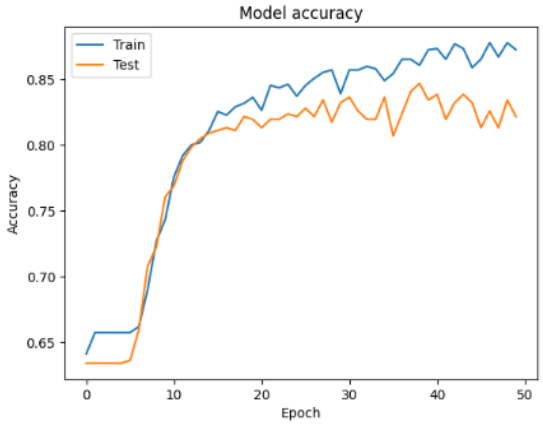

Figure 4. *BiLSTM accuracy graph with BERT Embedding*

The trend of increasing accuracy in Figure 4 for the data train increases steadily and reaches a high value (85%). The accuracy of the test data has also increased, but not as high as the train data (80%).

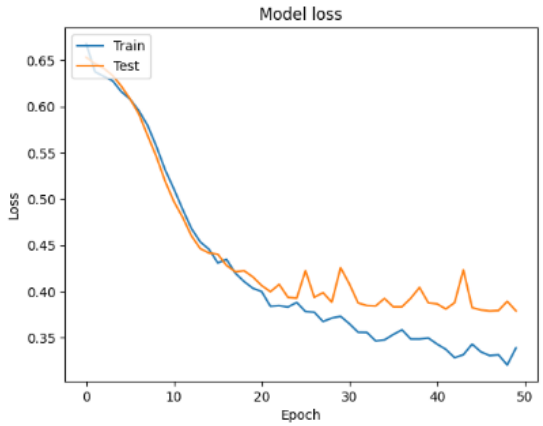

Figure 5. *BiLSTM accuracy graph with BERT Embedding*

The Trend Loss on the Chart shown in Figure 5. shows that the loss for the data train (blue line) decreases consistently, while the loss for the test data (orange line) also shows a decrease but with greater fluctuations. This decrease in loss shows that the model learns well from the training data provided.

If the loss test remains undiminished in proportion to the loss train, this could be a sign that the model is starting to overfit. This can be seen from the greater fluctuations in the loss test even though the loss train is decreasing.

The training process with this model resulted in an accuracy of 82%, a recall of 65%, and a precision of 83%.

3.1.2. BiLSTM+RoBERTa Combination

Model: "sequential_1"

Layer (type)	Output Shape	Param #
bidirectional_2 (Bidirectional)	(None, 1, 256)	918,528
dropout_2 (Dropout)	(None, 1, 256)	0
bidirectional_3 (Bidirectional)	(None, 92)	111,504
dropout_3 (Dropout)	(None, 92)	0
dense_1 (Dense)	(None, 1)	93

Total params: 1,030,125 (3.93 MB)
Trainable params: 1,030,125 (3.93 MB)
Non-trainable params: 0 (0.00 B)

Figure 6. *Results of the BiLSTM Model with RoBERTa Embedding*

The model results in Figure 6. is a Bidirectional LSTM (BiLSTM) architecture that uses embedding from RoBERTa as its input. RoBERTa is a BERT-based transformer model that is known for its ability to capture word representations with better context. In this model, the input data that has gone through the embedding process using RoBERTa is then processed by two layers of bidirectional LSTM (Bidirectional LSTM). The first layer has the ability to process information from both forward and backward directions, which helps the model understand the more complete context of a given text.

After the first Bidirectional LSTM layer, the model is equipped with a Dropout layer that serves to reduce the risk of overfitting. Dropouts randomly disable a subset of neurons during training, which helps the model be more robust and able to better generalize on data that has never been seen before. Then, the data that has been processed by the first Bidirectional LSTM is further processed through the second LSTM layer to deepen the understanding of data representation. The output of this layer is more concise, representing the core of the previously processed information.

Eventually, the model completes the process with a Dense layer that produces an output in the form of probabilities for the desired target class, in this case using the sigmoid activation function for binary classification. The total parameters that can be learned in this model reach 1,030,125, which covers all layers and allows the model to learn from the given data. The model is expected to capture complex relationships in texts, such as those often encountered in text classification tasks, including in the case of news or opinions.

Jurnal Teknik Informatika Vol. 18 No. 2, October 2025 (196-208)
ISSN: p-ISSN 1979-9160 (Print)| e-ISSN 2549-7901 (Online)
DOI: https://doi.org/10.15408/jti.v18i2.42459

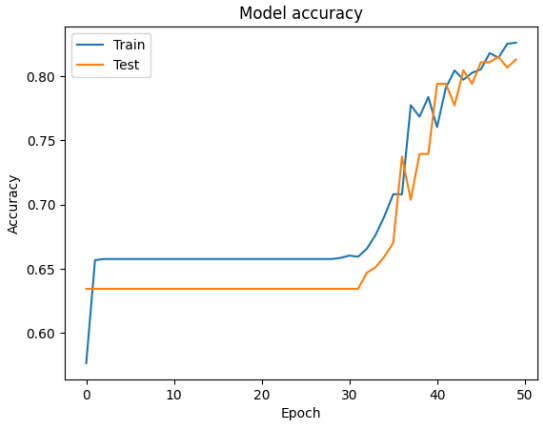

Figure 7. *Results of the BiLSTM Model with RoBERTa Embedding*

Figure 7 shows the accuracy trend increased significantly from the beginning of the training. It shows that the model learns well from the training data and gets a better understanding of the patterns in the data. It peaked at 82% in the final epoch. Meanwhile, the test line shows a slower improvement compared to the accuracy of the train. It fluctuates frequently, especially in the final epoch, but eventually achieves stability at 80%.

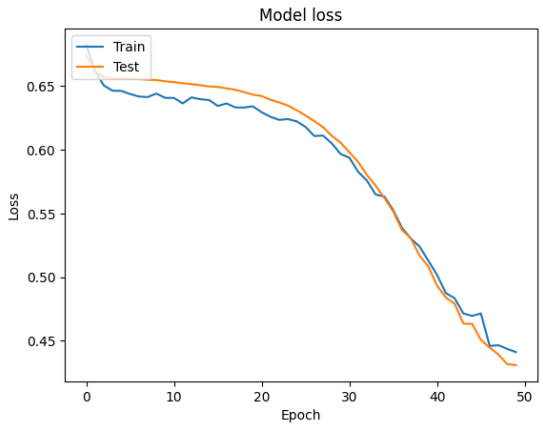

Figure 8. *Results of the BiLSTM Model with RoBERTa Embedding*

In figure 8. The loss for the training data in the Figure shows a consistent downward trend from the beginning to the end of the epoch. Reflects that the model successfully learns from the training data, with predicted errors decreasing over time.

This model results in an accuracy of 81%, a recall of 57%, and a precision of 86%.

3.1.3. BiLSTM+DistilBERT Combination

Model: "sequential"

Layer (type)	Output Shape	Param #
bidirectional (Bidirectional)	(None, 1, 256)	918,528
dropout (Dropout)	(None, 1, 256)	0
bidirectional_1 (Bidirectional)	(None, 92)	111,504
dropout_1 (Dropout)	(None, 92)	0
dense (Dense)	(None, 1)	93

Total params: 1,030,125 (3.93 MB)
Trainable params: 1,030,125 (3.93 MB)
Non-trainable params: 0 (0.00 B)

Figure 9. *Results of the BiLSTM model with DistilBERT embedding*

The results of the BiLSTM model with DistilBERT embedding in Figure 9 are designed for efficient text classification. The model begins by incorporating the DistilBERT embedding into a bidirectional LSTM (BiLSTM) layer with 128 units in each direction, resulting in a combined output with dimensions of 256. This structure allows the model to capture contextual information from the beginning and end of each text sequence, which is very useful in understanding complex language patterns.

Furthermore, a dropout layer with a level of 0.5 was applied to reduce overfitting by randomly disabling half of the neurons during training. The result then passes through a second layer of BiLSTM with 46 units in each direction, resulting in an output with dimensions of 92. After that, a second layer of dropout is applied, further reducing the risk of overfitting. The final layer is a dense layer with a sigmoid activation function used for binary classification, resulting in a probability score between 0 and 1 for each input.

The model has a total of 1,030,125 trainable parameters, making it relatively lightweight mainly thanks to the efficiency of the DistilBERT architecture. This combination allows the model to work effectively in text classification tasks with compute needs maintained.

The accuracy of the training data in Figure 10 showed a consistent improvement, reaching a high score of 85% at the end of the training. The model learns well from the training data, reflecting a strong understanding of patterns in the data. Meanwhile, the accuracy of the test data also shows an increasing trend, although the fluctuations are more visible compared to the training data.

Jurnal Teknik Informatika Vol. 18 No. 2, October 2025 (196-208)
ISSN: p-ISSN 1979-9160 (Print)| e-ISSN 2549-7901 (Online)
DOI: https://doi.org/10.15408/jti.v18i2.42459

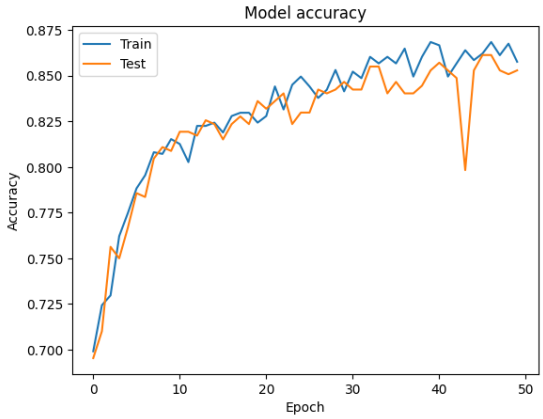

Figure 10. *Results of the BiLSTM Model with DistilBERT embedding*

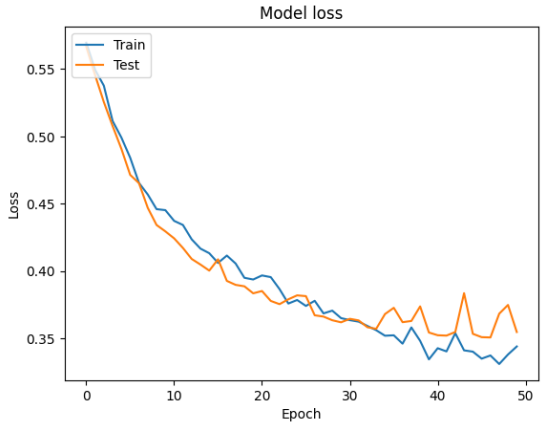

Figure 11. *Results of BiLSTM Model with DistilBERT Embedding*

The loss for the training data in figure 11 shows a consistent decline from the beginning to the end of the training. A significant decline occurred in the first few epochs and then stabilized at a lower level, which is 0.35.

This model gets an accuracy score of 85%, recall 70%, and precision 86%.

The evaluation results of the three combination models show that the combination of BiLSTM with various word embedding models (BERT, RoBERTa, and DistilBERT) provides varying performance on the classification task. DistilBERT gave the best results with an accuracy of 85%, recall of 70%, and accuracy of 86%, indicating that the model is not only accurate in identifying classes but also quite consistent in capturing patterns from text efficiently. BERT achieves a fairly good accuracy of 82%, with the highest accuracy at 83%, but a recall of 65% indicates that the model is slightly weaker at identifying all classes in complete than DistilBERT. The RoBERTa has the highest precision at 86%, but the recall is lower at 57%, resulting in an accuracy of 81%, which shows that the model is more conservative in its classification. Overall, the DistilBERT and BiLSTM are the best combination among the three models in terms of the balance between precision, recall, and accuracy.

Tabel 1. *Training performa model*

Type	Recall	Precission	Accuracy
Bert	65	83	82
RoBERTa	57	86	81
DistilBERT	70	86	85

3.2. Prediction Testing

The testing process is carried out after each model is saved. Testing is used to predict the labeling of new documents of 500 data.

3.2.1. BiLSTM -BERT Combination

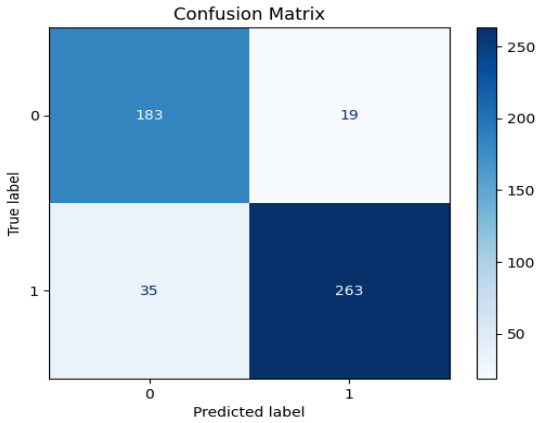

Figure 12. *CM BiLSTM- BERT*

Figure 12 of the confusion matrix shown shows the performance of the classification model with two classes, namely class 0 and class 1. In this matrix, the number 183 in position (0.0) indicates the correct number of predictions for class 0, while 263 in position (1.1) indicates the correct number of predictions for class 1. The number 19 in the position (0.1) is the number of false positives for class 1, and 35 in the position (1.0) is the wrong negative prediction for class 0 (false negatives). Overall, the model performed well, with a high level of accuracy in both classes, despite some errors in predictions.

The ROC curve graph in figure 13 shows the performance of the classification model in distinguishing between Bullying and Non-Bullying classes. The Y axis (True Positive Rate) represents the model's ability to detect positive classes, while the X axis (False Positive Rate) shows errors in predicting negative classes as positive. The orange curve is close to

Jurnal Teknik Informatika Vol. 18 No. 2, October 2025 (196-208)
ISSN: p-ISSN 1979-9160 (Print)| e-ISSN 2549-7901 (Online)
DOI: https://doi.org/10.15408/jti.v18i2.42459

the top left corner, showing excellent performance, with an AUC (Area Under the Curve) of 0.96, which indicates that the model can distinguish between the two classes very accurately. The diagonal blue line represents random predictions, and the distance of the curve from this line indicates the model's advantage over random guesses.

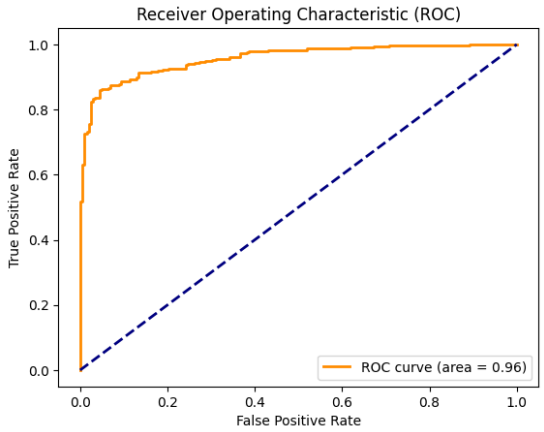

Figure 13. *Grafik ROC curve BiLSTM- BERT*

3.2.2. RoBERTA-BiLSTM

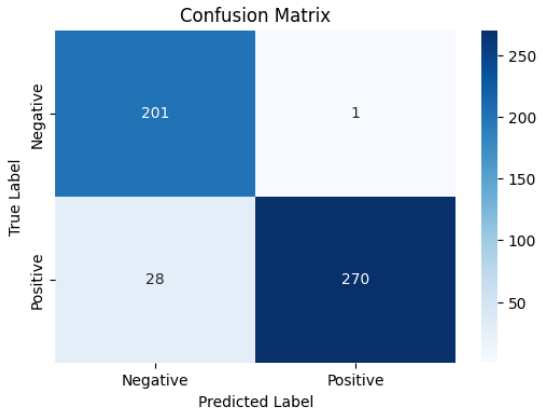

Figure 14. *CM BiLSTM -RoBERTa*

This confusion matrix shows the results of the evaluation of the classification model on the test data. The model successfully predicted 201 negative samples (True Negative) and 270 positive samples (True Positive) correctly. However, there was 1 negative sample that was incorrectly predicted as positive (False Positive) and 28 positive samples that were incorrectly predicted as negative (False Negative). These results show that the model has high accuracy in detecting both classes, with more prediction errors occurring in positive classes that are incorrectly classified as negative.

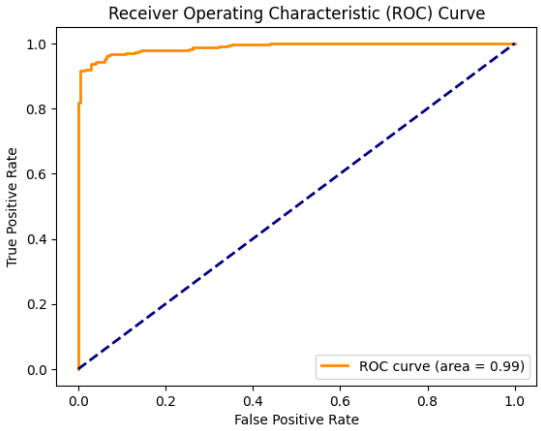

Figure 15. *ROC curve BiLSTM-RoBERTa graph*

The ROC curve graph in Figure 15 shows the performance of the classification model with excellent ability to distinguish between positive and negative classes. The Y-axis (True Positive Rate) indicates the success rate of the model in detecting positive classes, while the X-axis (False Positive Rate) indicates the error rate of prediction of negative classes as positive. The orange curve near the top left corner shows high accuracy, with an AUC (Area Under the Curve) of 0.99, indicating that the model is almost perfect at distinguishing the two classes. The blue diagonal line as the baseline of the random prediction confirms the model's superiority significantly.

3.2.3. BiLSTM-DistlBERT

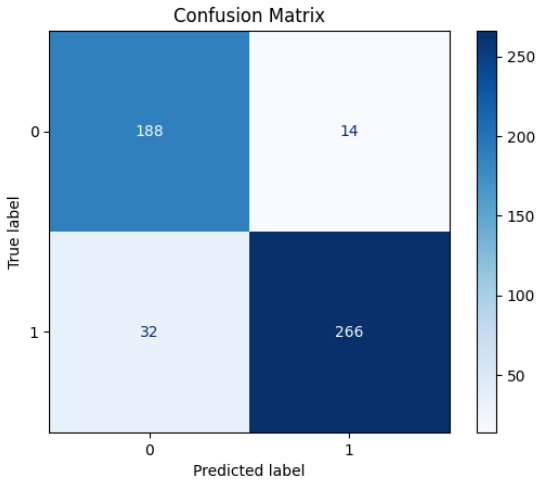

Figure 16. *CM BiLSTM-DistilBERT*

Figure 16 of the Confusion Matrix above shows the performance of the classification model with four main metrics: True Positive (TP) as many as 266, True Negative (TN) as many as 188, False Positive (FP) as many as 14, and False Negative (FN) as much as 32. This

Jurnal Teknik Informatika Vol. 18 No. 2, October 2025 (196-208)
ISSN: p-ISSN 1979-9160 (Print)| e-ISSN 2549-7901 (Online)
DOI: https://doi.org/10.15408/jti.v18i2.42459

means that the model successfully predicted 266 positive samples and 188 negative samples correctly. However, there were 14 negative samples that were incorrectly predicted as positive, and 32 positive samples that were incorrectly predicted as negative. Overall, this matrix provides a clear picture of the model's accuracy as well as its error distribution, which can be used to evaluate and improve the model's performance.

Figure 17. The Receiver Operating Characteristics Curve (ROC) is an evaluation tool used to assess the performance of a classification model. The graph above shows the relationship between the True Positive Rate and the False Positive Rate. The curve indicated by the orange line indicates that the model has an area under the curve (AUC) of 0.97, which indicates excellent classification capabilities. The closer the AUC value to 1, the better the model can distinguish between positive and negative classes. The blue dotted line represents the random guess effect, confirming that a better model should be well above the line.

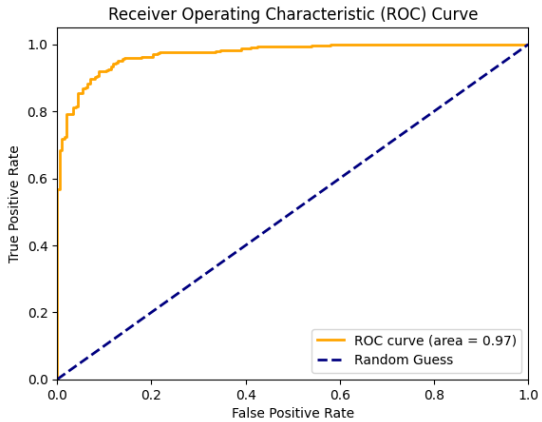

Figure 17. *BiLSTM-DistlBERT ROC curve graph*

The results of the testing can be seen in the following table:

Table 2. *Test model performance*

tim	Recall	Precission	Accuracy	ROC
Bert	89.5	86.5	89	96
RoBERTa	90	99	94	99
DistilBERT	89	95	90	97

The results of the test evaluation on the new data demonstrated improved performance across all three models, with the combination of BiLSTM and RoBERTa yielding the best results. RoBERTa achieved the highest accuracy at 94%, alongside an exceptional precision of 99% and a recall of 90%. This indicates that RoBERTa is highly effective in identifying classes with near-perfect precision while consistently capturing most patterns in the data. In terms of ROC, RoBERTa also excelled with a score of 99, reinforcing its superior discriminative ability. The combination of BiLSTM with DistilBERT delivered solid performance as well, achieving 90% accuracy, 95% precision, and 89% recall. DistilBERT strikes a good balance between accuracy and precision, and its lighter architecture makes it a suitable choice for applications requiring efficiency. Meanwhile, the combination of BiLSTM and BERT reached 89% accuracy, with a recall of 89.5% and a precision of 86.5%. Although its precision is lower than that of the other models, BERT still demonstrates its capability in pattern recognition and delivering reliable results. Overall, RoBERTa achieved the highest testing accuracy (94%) and precision (99%), indicating its strong generalization ability. This can be attributed to its improved pretraining methodology, which optimizes masked language modeling and dynamically changes learning objectives. In contrast, DistilBERT exhibited a balanced trade-off between speed and accuracy, making it a viable choice for resource-constrained applications.

3.3. Significance Test

In this sub-chapter, a significance test is carried out to find out if there is a significant difference in the average accuracy of the BERT, RoBERTa, and StylBERT models in classifying data. This analysis involves a one-way ANOVA test to identify differences between groups and a further test of Tukey HSD to determine model pairs that have significant differences.

The trial was carried out 5 times to obtain accuracy results on each model as follows in table 3

Table 3. *Average model accuracy*

Model	Accuracy					
	1	2	3	4	5	***Rata***
Bert	90	88	89	89	89	89,0
Roberta	93	93	94	95	95	94,0
DistilBERT	89	90	92	89	90	90,0

Based on table 3, it can be seen that the RoBERTa model has the highest average accuracy of 94%, followed by DistilBERT with 90%, and BERT with 89%.

From the accuracy table, an ANOVA test will be carried out to analyze whether there is a

significant difference in the average accuracy between the three models.
The hypotheses tested are as follows:

a. H_0 (Hypothesis zero): There is no difference in the average accuracy between models.
b. H_1 (Alternative hypothesis): There is at least one pair of models that have significant differences in average accuracy.

The results of the ANOVA test showed an F-Statistic value of 35,0000 and a P-Value of 0.0000. Since the p-value < 0.05, the null hypothesis (H_0) is rejected. This indicates that there is a significant difference in the average accuracy between at least the two models.

To find out which model pairs have significant differences, further tests were carried out using Tukey HSD. The results of the Tukey HSD test are shown in the following table 4:

Table 4. *Tukey HSD Post-Hoc Test*

Group1	Group2	M. Diff	P-Val	Low	Upp	Rej
BERT	DistilBERT	1.0	0.2908	0.6873	2.6873	F
BERT	RoBERTa	5.0	0.0000	3.3127	6.6873	T
DistilBERT	RoBERTa	4.0	0.0001	2.3127	5.6873	T

Table 4 displays the results of the Tukey HSD (Post-Hoc Test) which is used to find out which model pairs have significant differences in the average accuracy. Based on the test results, the comparison between BERT and DistilBERT shows an mean difference of 1.0 with p-value = 0.2908 (> 0.05), which means that there is no significant difference between these two models. Meanwhile, the comparison between BERT and RoBERTa has an average difference of 5.0 with p-value = 0.0000 (< 0.05), showing a significant difference between the two models. Similarly, the comparison between DistilBERT and RoBERTa showed an average difference of 4.0 with p-value = 0.0001 (< 0.05), which also showed a significant difference. Thus, it can be concluded that the RoBERTa has a significant difference in accuracy compared to the other two models.

Thus, in the tested classification scenario, RoBERTa is the model with the most statistically superior performance.

CONCLUSION

This study demonstrated that combining BiLSTM with RoBERTa achieved the highest accuracy (94%) in online bullying news classification, highlighting its robustness in detecting harmful content. While DistilBERT offers a balance between efficiency and accuracy, RoBERTa is recommended for applications prioritizing precision. Future research can explore additional datasets, integrate domain-specific embeddings, and evaluate real-time deployment feasibility in content moderation systems.

REFERENCES

[1] J. Song, K. Kim, Y. Han, and T. M. Song, "Classification of Bullying-Related Web Documents: An Ecological Systems and Machine Learning Approach," Jan. 29, 2023, *Social Science Research Network, Rochester, NY*: 4341006. doi: 10.2139/ssrn.4341006.

[2] S. Salawu, Y. He, and J. Lumsden, "Approaches to Automated Detection of Cyberbullying: A Survey," *IEEE Transactions on Affective Computing*, vol. 11, no. 1, pp. 3–24, Jan. 2020, doi: 10.1109/TAFFC.2017.2761757.

[3] F. A. Nirmala, M. Jazman, N. E. Rozanda, and F. N. Salisah, "CYBERBULLYING SENTIMENT ANALYSIS OF INSTAGRAM COMMENTS USING NAÏVE BAYES CLASSIFIER AND K-NEAREST NEIGHBOR ALGORITHM METHODS," *Jurnal Teknik Informatika (Jutif)*, vol. 5, no. 5, Art. no. 5, May 2024, doi: 10.52436/1.jutif.2024.5.5.1997.

[4] B. I. Kusuma and A. Nugroho, "CYBERBULLYING DETECTION ON TWITTER USES THE SUPPORT VECTOR MACHINE METHOD," *Jurnal Teknik Informatika (Jutif)*, vol. 5, no. 1, Art. no. 1, Jan. 2024, doi: 10.52436/1.jutif.2024.5.1.809.

[5] M. F. Hibatulloh, D. N. Suci, A. G. Puspita, and I. F. Rohmah, "A Critical Discourse Analysis on Antara English News Reports About Bullying in Education Institutions in Indonesia," *1*, vol. 5, no. 3, Art. no. 3, Oct. 2023.

[6] A. D. Gower, T. Vaillancourt, H. Brittain, K. Pletta, and M. A. Moreno,

"185. Understanding News Media Coverage On Bullying And Cyberbullying," *Journal of Adolescent Health*, vol. 64, no. 2, p. S94, Feb. 2019, doi: 10.1016/j.jadohealth.2018.10.201.

[7] K. Nemkul, "Use of Bidirectional Encoder Representations from Transformers (BERT) and Robustly Optimized Bert Pretraining Approach (RoBERTa) for Nepali News Classification," *Tribhuvan University Journal*, vol. 39, no. 1, Art. no. 1, Jun. 2024, doi: 10.3126/tuj.v39i1.66679.

[8] P. Singh and A. Jain, "A BERT-BiLSTM Approach for Socio-political News Detection," in *Proceedings of Fifth Doctoral Symposium on Computational Intelligence*, A. Swaroop, V. Kansal, G. Fortino, and A. E. Hassanien, Eds., Singapore: Springer Nature, 2024, pp. 203–212. doi: 10.1007/978-981-97-6036-7_17.

[9] W. Shi, M. Song, and Y. Wang, "Perturbation-enhanced-based RoBERTa combined with BiLSTM model for Text classification," in *ICETIS 2022; 7th International Conference on Electronic Technology and Information Science*, Jan. 2022, pp. 1–5. Accessed: Feb. 26, 2025. [Online]. Available: https://ieeexplore.ieee.org/document/9788645

[10] C. Y. Sy, L. L. Maceda, M. J. P. Canon, and N. M. Flores, "Beyond BERT: Exploring the Efficacy of RoBERTa and ALBERT in Supervised Multiclass Text Classification," *IJACSA*, vol. 15, no. 3, 2024, doi: 10.14569/IJACSA.2024.0150323.

[11] Q. An, B. Pan, Z. Liu, S. Du, and Y. Cui, "Chinese Named Entity Recognition in Football Based on ALBERT-BiLSTM Model," *Applied Sciences*, vol. 13, no. 19, Art. no. 19, Jan. 2023, doi: 10.3390/app131910814.

[12] "Python Based Machine Learning Text Classification - IOPscience." Accessed: Mar. 04, 2025. [Online]. Available: https://iopscience.iop.org/article/10.1088/1742-6596/2394/1/012015

[13] E. Hayati, M. R. Zamroni, D. H. Prayitno, and E. Rachmawati, "Sentiment analysis on Indomie vs Gaga polemic using tiktok data," *International Management Conference and Progressive Papers*, pp. 498–507, Nov. 2023.

[14] Y. Fauziah, B. Yuwono, and A. S. Aribowo, "Lexicon Based Sentiment Analysis in Indonesia Languages : A Systematic Literature Review," *RSF Conference Series: Engineering and Technology*, vol. 1, no. 1, Art. no. 1, 2021, doi: 10.31098/cset.v1i1.397.

[15] F. T. Saputra, S. H. Wijaya, Y. Nurhadryani, and Defina, "Lexicon Addition Effect on Lexicon-Based of Indonesian Sentiment Analysis on Twitter," in *2020 International Conference on Informatics, Multimedia, Cyber and Information System (ICIMCIS)*, Nov. 2020, pp. 136–141. doi: 10.1109/ICIMCIS51567.2020.9354269.

[16] A. Aribowo and S. Khomsah, "Implementation Of Text Mining For Emotion Detection Using The Lexicon Method (Case Study: Tweets About Covid-19)," *Telematika*, vol. 18, p. 49, Mar. 2021, doi: 10.31315/telematika.v18i1.4341.

[17] R. Darman, "Analisis Sentimen Respons Twitter terhadap Persyaratan Badan Penyelenggara Jaminan Sosial (BPJS) di Kantor Pertanahan," *Widya Bhumi*, vol. 3, no. 2, Art. no. 2, Oct. 2023, doi: 10.31292/wb.v3i2.61.

[18] Z. Li and Z. Zou, "Punctuation and lexicon aid representation: A hybrid model for short text sentiment analysis on social media platform," *Journal of King Saud University - Computer and Information Sciences*, vol. 36, no. 3, p. 102010, Mar. 2024, doi: 10.1016/j.jksuci.2024.102010.

[19] "Sentiment score-based classification for fake news using machine learning and LSTM-BiLSTM | Soft Computing." Accessed: Nov. 10, 2024. [Online]. Available: https://link.springer.com/article/10.1007/s00500-024-09884-9

[20] "Combining Bi-LSTM And Word2vec Embedding For Sentiment Analysis Models Of Application User Reviews | The Indonesian Journal of Computer Science." Accessed: Nov. 07, 2024. [Online]. Available: http://ijcs.net/ijcs/index.php/ijcs/article/view/3647

[21] G. Meera and Dr. R. Murugesan, "Improving sentiment analysis of financial news headlines using hybrid Word2Vec-TFIDF feature extraction technique," *Procedia Computer Science*, vol. 244, pp. 1–8, Jan. 2024, doi: 10.1016/j.procs.2024.10.172.

[22] J. Devlin, M.-W. Chang, K. Lee, and K. Toutanova, "BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding," May 24, 2019, *arXiv*: arXiv:1810.04805. doi: 10.48550/arXiv.1810.04805.

[23] A. Rogers, O. Kovaleva, and A. Rumshisky, "A Primer in BERTology: What We Know About How BERT Works," *Transactions of the Association for Computational Linguistics*, vol. 8, pp. 842–866, Jan. 2021, doi: 10.1162/tacl_a_00349.

[24] M. Zaib, Q. Z. Sheng, and W. Emma Zhang, "A Short Survey of Pre-trained Language Models for Conversational AI-A New Age in NLP," in *Proceedings of the Australasian Computer Science Week Multiconference*, in ACSW '20. New York, NY, USA: Association for Computing Machinery, Feb. 2020, pp. 1–4. doi: 10.1145/3373017.3373028.

[25] R. Alshalan and H. Al-Khalifa, "A Deep Learning Approach for Automatic Hate Speech Detection in the Saudi Twittersphere," *Applied Sciences*, vol. 10, no. 23, Art. no. 23, Jan. 2020, doi: 10.3390/app10238614.

[26] Z. Mu, S. Zheng, and Q. Wang, "ACL-RoBERTa-CNN Text Classification Model Combined with Contrastive Learning," in *2021 International Conference on Big Data Engineering and Education (BDEE)*, Aug. 2021, pp. 193–197. doi: 10.1109/BDEE52938.2021.00041.

[27] R. Mengi, H. Ghorpade, and A. Kakade, "Fine-tuning T5 and RoBERTa Models for Enhanced Text Summarization and Sentiment Analysis".

[28] V. Sanh, L. Debut, J. Chaumond, and T. Wolf, "DistilBERT, a distilled version of BERT: smaller, faster, cheaper and lighter," Mar. 01, 2020, *arXiv*: arXiv:1910.01108. doi: 10.48550/arXiv.1910.01108.

[29] R. Silva Barbon and A. T. Akabane, "Towards Transfer Learning Techniques—BERT, DistilBERT, BERTimbau, and DistilBERTimbau for Automatic Text Classification from Different Languages: A Case Study," *Sensors*, vol. 22, no. 21, Art. no. 21, Jan. 2022, doi: 10.3390/s22218184.

[30] S. Akpatsa *et al.*, "Online News Sentiment Classification Using DistilBERT," *JQC*, vol. 4, no. 1, pp. 1–11, 2022, doi: 10.32604/jqc.2022.026658.

[31] O. Karakaya and Z. H. Kilimci, "An efficient consolidation of word embedding and deep learning techniques for classifying anticancer peptides: FastText+BiLSTM," *PeerJ Comput. Sci.*, vol. 10, p. e1831, Feb. 2024, doi: 10.7717/peerj-cs.1831.

[32] J. Juarros-Basterretxea, G. Aonso-Diego, Á. Postigo, P. Montes-Álvarez, Á. Menéndez-Aller, and E. García-Cueto, "Post-hoc tests in one-way ANOVA: The case for normal distribution," *Methodology*, vol. 20, no. 2, pp. 84–99, Jun. 2024, doi: 10.5964/meth.11721.

[33] D. C. Montgomery, *Design and Analysis of Experiments*, 8th ed. Arizona State University, 2022.