Jurnal Teknik Informatika (JUTIF)
Vol. 6, No. 1, February 2025, pp. 41-50

DOI: https://doi.org/10.52436/1.jutif.2025.6.1.4140
p-ISSN: 2723-3863
e-ISSN: 2723-3871

DETECTION OF BULLYING CONTENT IN ONLINE NEWS USING A COMBINATION OF RoBERTa-BiLSTM

Moh. Rosidi Zamroni[*1], Rahayu A Hamid[2], Siti Mujilahwati[3], Miftahus Sholihin[4], Dinar Mahdalena Leksana[5]

[1,3,4]Informatics Engineering, Faculty of Science and Technology, Lamongan Islamic University
[2] FSKTM, Faculty of Computer Science and Information Technology, Universiti Tun Hussein Onn Malaysia, Malaysia
[4]Piaud(BK), Faculty of Islamic Studies, Lamongan Islamic University
Email: [1]rosidizamroni@unisla.ac.id, [2]moedjee@unisla.ac.id, [3]miftahus.sholihin@unisla.ac.id

(Article received: November 22, 2024; Revision: December 18, 2024; published: February 20, 2025)

Abstract

This research aims to build a bullying-themed online news classification system with a combined approach of RoBERTa embedding and BiLSTM. RoBERTa is used to generate context-rich text representations, while BiLSTM captures temporal relationships between words, thereby improving classification performance. The research dataset consisted of news from reputable portals such as Kompas.com, Detik.com, and iNews.com, labeled according to keywords relevant to the theme of bullying. The results of the experiment showed that the model achieved 95.2% accuracy, 98.2% precision, 93.6% recall, and 95.8% F1-score. Although there are few prediction errors (false positives and false negatives), this model shows excellent performance in detecting and classifying bullying-themed news. The main contribution of this research is the development of a new approach that combines RoBERTa and BiLSTM for the classification of complex bullying-themed news. This approach not only improves the accuracy of classification but can also be implemented in automated systems to detect negative content. Thus, this research has the potential to support the creation of a healthier digital space and encourage more responsible media practices.

Keywords: *bullying detection, BiLSTM, news classification, online news, RoBERTa*

DETEKSI KONTEN PERUNDUNGAN DALAM BERITA ONLINE MENGGUNAKAN KOMBINASI RoBERTa-BiLSTM

Abstrak

Penelitian ini bertujuan membangun sistem klasifikasi berita online bertema bullying dengan pendekatan kombinasi RoBERTa embedding dan BiLSTM. RoBERTa digunakan untuk menghasilkan representasi teks yang kaya konteks, sementara BiLSTM menangkap hubungan temporal antar kata, sehingga meningkatkan kinerja klasifikasi. Dataset penelitian terdiri dari berita dari portal bereputasi seperti Kompas.com, Detik.com, dan iNews.com, yang dilabeli sesuai kata kunci relevan dengan tema bullying. Hasil eksperimen menunjukkan bahwa model mencapai akurasi 95,2%, precision 98,2%, recall 93,6%, dan F1-score 95,8%. Meskipun terdapat sedikit kesalahan prediksi (false positives dan false negatives), model ini menunjukkan performa yang sangat baik dalam mendeteksi dan mengklasifikasikan berita bertema bullying. Kontribusi utama penelitian ini adalah pengembangan pendekatan baru yang menggabungkan RoBERTa dan BiLSTM untuk klasifikasi berita kompleks bertema bullying. Pendekatan ini tidak hanya meningkatkan akurasi klasifikasi tetapi juga dapat diimplementasikan dalam sistem otomatis untuk mendeteksi konten negatif. Dengan demikian, penelitian ini berpotensi mendukung terciptanya ruang digital yang lebih sehat dan mendorong praktik media yang lebih bertanggung jawab.

Kata kunci: *bullying detection, BiLSTM, news classification, online news, RoBERTa*

1. INTRODUCTION

The rapid development of information technology has changed the way humans interact and obtain information. In the digital era, online news has become one of the main media for conveying information quickly and widely. However, the ease of information distribution also brings challenges, especially related to negative content such as bullying. Bullying, or harassment, is an act that can have serious impacts on an individual's mental and emotional health [1]. Not only does it happen in the real world, but bullying is also commonly found in the digital space, including in the form of online news that often contains direct or implied bullying narratives [2], [3], [4].

The dissemination of news with a bullying theme has a significant impact on the formation of public opinion. Such news often reinforces stereotypes, triggers debates, and even causes trauma for the victims [5]. Furthermore, news that is not processed wisely can worsen the situation, as readers may imitate the mindset or actions contained within it. Therefore, early detection of bullying-themed news is a crucial step to ensure that this negative content can be controlled, so it does not have a detrimental impact on the wider community.

On the other hand, efforts to detect bullying news are not solely aimed at removing such content, but also at supporting digital platforms and online media in providing a healthier information space[6]. In this context, the classification of online news themed around bullying becomes very important. This classification helps identify relevant news, facilitates filtering, and supports the creation of responsible editorial policies.

With the advancements in the field of natural language processing (NLP), machine learning and deep learning-based models have become effective solutions to address this challenge. One of the latest technologies widely used is RoBERTa (Robustly Optimized BERT Approach), a text embedding model that has an extraordinary ability to understand linguistic context [7]. By utilizing embeddings from RoBERTa, online news can be transformed into rich numerical representations, allowing algorithms to process them efficiently.

Previous research has demonstrated the effectiveness of RoBERTa-based models in handling complex text classification tasks, including the detection of toxic content in Indonesian. For example, a study evaluating IndoBERTweet, IndoBERT, and Indonesian RoBERTa found that IndoBERTweet, when adjusted with optimal hyperparameters, was able to achieve high performance with an F1-score of 0.89 in the toxicity detection task. This finding shows that RoBERTa-based models have a good ability to understand language nuances and local contexts. With this success, RoBERTa-based approaches have great potential to be applied to various other text classification tasks, including online news classification [8].

In addition, the Bidirectional Long Short-Term Memory (BiLSTM) model, designed to capture temporal relationships in sequential data, has become one of the main approaches in text classification [9]. BiLSTM is capable of understanding the relationships between words in a broader context, making it suitable for detecting sensitive themes such as bullying. The combination of RoBERTa for text representation and BiLSTM for classification has proven to yield accurate results in various NLP tasks [10].

In a study conducted by Mohawes [11] on the influence of internet reviews on consumer purchase decisions, which can be manipulated by fake reviews to mislead customers and businesses. The study integrated RoBERTa with an LSTM layer, allowing for the identification of complex patterns in sham reviews. This approach improves the resilience of fake review detection and understanding of authentic behavior. The results of experiments on semi-real datasets such as OpSpam and Deception show that these models surpass current methods with an accuracy of 96.03% and 93.15%, respectively.

The combination of RoBERTa and BiLSTM was also carried out in research conducted by Qiumei PU [12], namely the development of a text classification model to detect attitudes (stance detection) in Chinese texts. The purpose of attitude detection is to identify the author's views on a particular topic, such as supporting, opposing, or neutral. The proposed model combines the deep semantic understanding capabilities of RoBERTa with BiLSTM's ability to capture sequential information, thereby improving the accuracy of attitude detection in Chinese-language texts. Experiments were conducted on the NLPCC-2016 Task4 dataset for attitude detection on the Weibo platform, and the results showed that the model significantly improved accuracy and other key performance metrics. This study confirms that the combination of RoBERTa and BiLSTM is an effective strategy to solve the attitude detection task in Chinese.

Experiments were conducted on the NLPCC-2016 Task4 dataset for attitude detection on the Weibo platform, and the results showed that the model significantly improved accuracy and other key performance metrics. This study confirms that the combination of RoBERTa and BiLSTM is an effective strategy to solve the attitude detection task in Chinese.

Zhang [13] researched voice and text-based multimodal emotion recognition methods to overcome the problems of lack of feature extraction and sample imbalance in current emotion recognition technology. This study proposes a model with two channels, where one channel uses sound features processed through the wavelet method and extraction by a sparse autoencoder, while the other channel utilizes text features extracted using a combination of BERT-RoBERTa models. Emotion recognition was performed using the attention layer, a two-layer BiLSTM, and a combined loss function between cross-entropy loss and focus loss. This model was tested on an unbalanced IEMOCAP dataset, resulting in a weighted accuracy (WA) of 73.95% and an unweighted accuracy (UA) of

74.27%, which outperformed other models. RoBERTa is used in this model for deep text feature extraction, providing rich contextual representation. Meanwhile, BiLSTM with two layers plays a crucial role in capturing temporal relationships in data, allowing models to understand complex sequences and patterns in voice and text information. The combination of the two provides a significant advantage in addressing the challenges of data imbalance and lack of features in emotion recognition.

In research conducted by Aditya Bhamre [14], has shown that Bidirectional Long Short-Term Memory (BiLSTM) outperforms various traditional methods in text classification tasks, including cyberbullying detection, with an accuracy of 82.27% after optimization. This advantage is primarily due to its ability to capture long-term dependencies and understand linguistic patterns bidirectionally, resulting in richer and more relevant data representations. This proves that BiLSTM is effective in handling complex text data. In this context, this research adapts the strength of BiLSTM for the task of online news classification by utilizing embeddings from RoBERTa. This combination is expected to enhance classification accuracy through the integration of deep contextual representations from RoBERTa with the sequential processing capabilities of BiLSTM, thereby addressing the challenges in understanding the complex language structure and context of news[15].

The study [16] introduced the **RoBERTa-wwm-CBA** model specifically designed to detect potential mental disorders through text analysis, combining the power of **RoBERTa-wwm** and **BiLSTM** as key components. RoBERTa-wwm plays a crucial role in capturing the deep semantic features of the text, allowing the model to understand the nuances and context of the language with precision. Meanwhile, **BiLSTM** is used to capture temporal relationships and broader context in text data, ensuring that word sequences and temporal patterns are not missed. This combination is reinforced with a CNN layer for local features and an attention mechanism to give weight to the most relevant features, resulting in a comprehensive representation of the text. The results of the evaluation on *the Emotional First Aid dataset* show that the integration of RoBERTa-wwm and BiLSTM in this model significantly improves accuracy, precision, recall, and F1-score, making it an effective solution in the early detection of mental disorders.

Of the several previous studies that applied RoBERTa and BILSTM, no one has explored the application of this method specifically for the classification of Indonesian-language online news. Given the importance of this issue, this study aims to develop a bullying-themed online news classification system using a combination of RoBERTa and BiLSTM. With a dataset that includes news from various reputable news portals such as Kompas.com, Detik.com, and iNews.com, this system is expected to be an effective tool in detecting and classifying bullying-themed news.

2. RESEARCH METHODS

The stages carried out in this study are as follows:

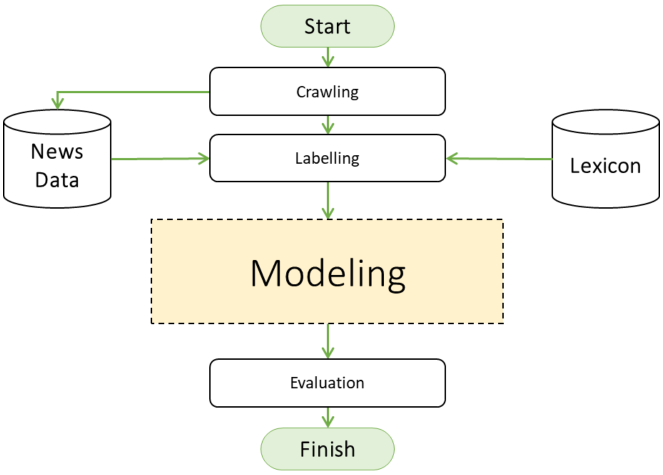


Figure 1. Flowchart of the research process

The stages of the flow chart in Figure 1 illustrate the process of classifying news data, which begins with data collection through crawling and storage in News Data. The data then goes through a labeling process, assisted by Lexicon as a reference. Furthermore, the labeled data is used to build the model (modeling). Once the model is trained, its performance is evaluated using specific metrics. The process concludes with an evaluation of performance and a ready-to-use finish. More details will be explained in the following sub-chapters in a sequential and directed manner:

2.1. Data collection (Crawling)

This study uses data obtained from three major online news portals, namely *Kompas.com*, *iNews*, and *Detik.com*. Data collection was carried out by *crawling* techniques [17] using a combination of tools and Python libraries. The stages of data collection are described as follows:

1. **News Link Collection**
   News links from all three portals are collected using the *Linkclump* **browser extension**, which allows users to save multiple links at once in a **.csv format**.
2. **News Content Capture**
   The news data that has been collected in .csv file is then processed using Python libraries, namely **Requests**, **BeautifulSoup**, and **Pandas**. This process is done to automatically fetch news content from each available link.

3. **Data Collection Results**

   The crawling process resulted in **2,800 articles** from the three news portals. The details of the data use are as follows: (1) **2,000 articles** are used as training data to train the model; (2) **800 articles** were allocated for the testing process.

2.2. Data labeling

Each article is labeled using the lexicon method, which identifies words relevant to the topic of bullying [18], [19]. Some of the keywords used for the identification process 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'. With this approach, a total of 852 articles were identified as bullying-themed articles, while another 1243 articles were categorized as non-bullying.

2.3. Modeling

At this stage, the labeled data will be processed using several NLP (Natural Language Processing)-based models with BiLSTM architecture with RoBERTa embedding.

In your study, the online news data classification technique using RoBERTa embedding-based vectorization is a powerful step, especially when paired with the BiLSTM model [10], [20]. This approach involves several key steps:

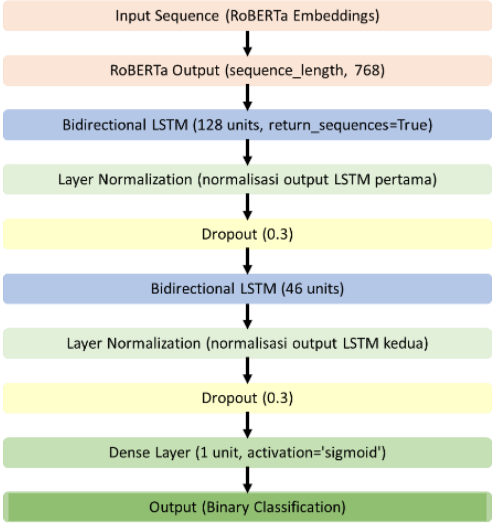

Input Sequence (RoBERTa Embeddings)
RoBERTa Output (sequence_length, 768)
Bidirectional LSTM (128 units, return_sequences=True)
Layer Normalization (normalisasi output LSTM pertama)
Dropout (0.3)
Bidirectional LSTM (46 units)
Layer Normalization (normalisasi output LSTM kedua)
Dropout (0.3)
Dense Layer (1 unit, activation='sigmoid')
Output (Binary Classification)

Figure 2. Proposed model

The stages of the model depicted in figure 2 are:

1. Preprocessing and Embedding: Online news data is captured and processed in the form of text, then converted into numerical representations using RoBERTa embedding [22]. This process results in richer representations, leveraging the powerful contextual features of RoBERTa.
2. BiLSTM Model: After embedding, the output from RoBERTa is provided as input to a well-designed BiLSTM architecture for processing sequential data. The structure of the model follows the following sequence:
   - Bidirectional LSTM (128 units, return_sequences=True): Composes a bidirectional layered LSTM to understand information from two directions of text context, with outputs for each sequential step.
   - Layer Normalization: Normalize on the first LSTM output to stabilize the training process.
   - Dropout (0.3): Prevents overfitting by shutting down a random number of nodes.
   - Bidirectional LSTM (46 units): A second LSTM layer for further processing.
   - Layer Normalization: The second normalization for this LSTM output.
   - Dropout (0.3): Re-application of dropout to avoid overfitting.
   - Dense Layer (1 unit, activation='sigmoid'): A single density layer to generate a final probability value, using sigmoid activation for binary classification.
   - Output: This structure ends on a binary classification, which in this case indicates whether or not the news contains a specific topic (such as bullying).

This approach augments the model with diverse information from RoBERTa embedding and bidirectional LSTM structures, while leveraging normalization techniques for stability and dropouts for better generalizations.

2.4. Evaluation

To evaluate a BiLSTM-based classification model with RoBERTa embedding, using the accuracy, precision, alerting, and F1 score metrics [23]. Akurasi adalah faktor utama dalam evaluasi kinerja model dalam penelitian ini [24]. Parameter ini dapat menunjukkan sejauh mana model mampu memberikan hasil yang tepat berdasarkan data yang tersedia [25]. Secara umum, akurasi menunjukkan persentase prediksi yang benar dari keseluruhan data, yang dapat menunjukkan kemampuan model dalam mengklasifikasikan data dengan baik[26]. Selain akurasi, metrik presisi juga digunakan untuk menilai seberapa baik prediksi yang positif dihasilkan [27]. Sebaliknya, recall digunakan untuk mengevaluasi seberapa baik model menemukan semua kasus relevan dalam kelas tertentu [28]. Sedangkan F1-Score digunakan untuk mengukur keseimbangan antara presisi dan recall dalam suatu model klasifikasi[29]. Metrik ini sangat berguna ketika terdapat ketidakseimbangan data antar kelas (misalnya satu kelas jauh lebih dominan daripada kelas lainnya). Here's a brief explanation of each metric and its evaluation steps:

1. **Accuracy**: Measures the percentage of correct predictions from the total data. Suitable for use if the dataset is balanced.

$$\text{Accuracy} = \frac{\text{TP} + \text{TN}}{\text{TP} + \text{TN} + \text{FP} + \text{FN}} \tag{1}$$

2. **Precision**: Measures the proportion of correct positive predictions. Useful when false positives need to be minimized.

$$\text{Precision} = \frac{\text{TP}}{\text{TP} + \text{FP}} \tag{2}$$

3. **Recall:** Measures the model's ability to detect all positive cases. It is important that false negatives need to be minimized.

$$\text{Recall} = \frac{\text{TP}}{\text{TP} + \text{FN}} \tag{3}$$

4. **F1-Score**: Harmonic average of precision and recall, ideal for unbalanced datasets.

$$\text{F1} - \text{Score} = 2 * \frac{\text{Precission} * \text{Recall}}{\text{Precision} + \text{RecallN}} \tag{4}$$

Example calculation: If out of 100 data points there are 70 correct positive predictions (TP), 10 correct negative predictions (TN), 10 false positive predictions (FP), and 10 false negative predictions (FN), then:

$$\boldsymbol{Akurasi} = \frac{70+10}{70+10+10+10} = \frac{80}{100} = 0.8(80\%)$$

$$\boldsymbol{Presisi} = \frac{70}{70+10} = \frac{70}{80} = 0.875(87.5\%)$$

$$\boldsymbol{Recall} = \frac{70}{70+10} = \frac{70}{80} = 0.875(87.5\%)$$

$$\boldsymbol{F1 - Score} = 2 \times \frac{0.875 \times 0.875}{0.875 + 0.875}$$

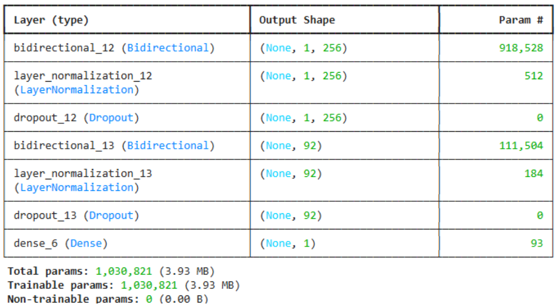
$$= 2 \times \frac{0.7656}{1.75} = 0.875(87.5\%)$$

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 2,800 data points 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 0.3 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.



```
Layer (type)                        Output Shape                 Param #
bidirectional_12 (Bidirectional)    (None, 1, 256)               918,528
layer_normalization_12              (None, 1, 256)                   512
(LayerNormalization)
dropout_12 (Dropout)                (None, 1, 256)                     0
bidirectional_13 (Bidirectional)    (None, 92)                   111,504
layer_normalization_13              (None, 92)                       184
(LayerNormalization)
dropout_13 (Dropout)                (None, 92)                         0
dense_6 (Dense)                     (None, 1)                         93
Total params: 1,030,821 (3.93 MB)
Trainable params: 1,030,821 (3.93 MB)
Non-trainable params: 0 (0.00 B)
```

Figure 3. Results of BiLSTM model with BERT embedding

Figure 3 shows the results of the execution of a neural network architecture for binary classification, consisting of two layers of BiLSTM (Bidirectional LSTM) designed to capture temporal relationships in sequential data. After each layer of BiLSTM, normalization was applied to stabilize learning and accelerate convergence, as well as dropouts to reduce the risk of overfitting by randomly ignoring neurons during training. The first layer of BiLSTM produces an output with dimensions (1, 256), while the second layer produces dimensions (92) to reduce complexity. The last Dense layer produces a scalar output (1) as the final prediction. The model has a total of 1,030,821 fully trainable parameters, with a size of about 3.93 MB. This design combines BiLSTM's ability to understand temporal context with regulatory strategies such as normalization and dropout, making it suitable for bullying and non-bullying news classification tasks.

The results of training with three different epochs can be seen in the following Table 1:

Table 1. Training performance

Epoch	Accuracy	Precision	Recall	F1
25	80	86	57	68
50	83	81	72	76
100	83	77	75	76

Based on Table 1 of the model training performance presented, it can be concluded as follows:

1. With 50 epochs, the model achieved the highest accuracy of 83%, with 81% precision, 72% recall, and 76% F1 score. This indicates that at this stage, the model has a pretty good performance overall.
2. With 100 epochs, the model also achieved 83% accuracy, but with lower precision 77% and higher recall 75% compared to the 50th epoch. The F1 score remains at 76%. This indicates that the model is increasingly able to identify more relevant examples, albeit with a slight decrease in prediction accuracy.

3. With 25 epochs, the model achieved lower accuracy 80% than epochs 50 and 100, but had higher precision 86% and lower recall 57%. The F1 score at this stage is 68%. This shows that in the early stages of training, the model tends to be more conservative in making predictions, so it has a higher level of precision but lower recall.

Overall, this table illustrates the evolution of the model's performance over the training process, with accuracy steadily improving to 83% at the 50th and 100th epochs. While there is slight variation in precision, recall, and F1 metrics, the model performs quite well in a given task.

The resulting visualization of the training process can be seen in Figure 4 and Figure 5.

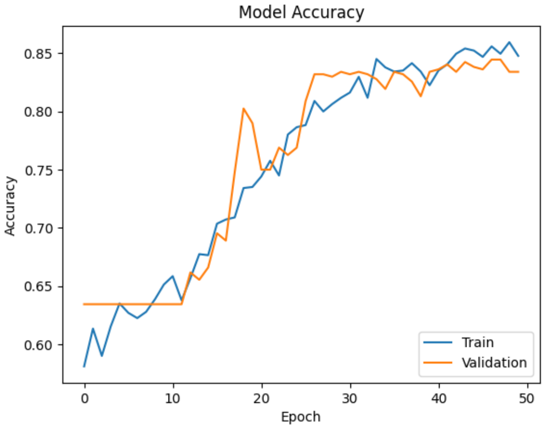


Figure 4. Accuracy

The graph in Figure 4 shows the accuracy of the model during the training process. There are two lines depicted, namely training accuracy and validation accuracy. The training accuracy starts at about 0.65 and continues to increase gradually until it reaches about 0.85 at the end of the epoch. Meanwhile, validation accuracy also follows a similar trend, although it does not reach the same high level as the training accuracy of 0.83. This shows that the model learns effectively during the training process and is able to generalize that learning into validation data. While there is a slight gap between the two curves, which indicates overfitting, the overall uptrend in both metrics indicates that the model is performing well.

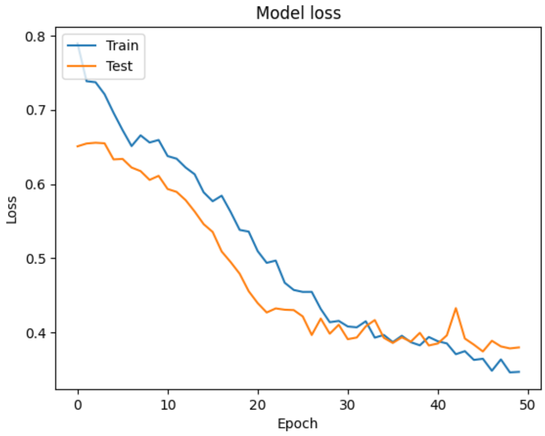


Figure 5. Loss

The Trend Loss on the Chart shown in Figure 5. shows changes in the model's loss value during the training process. There are two curves depicted, namely the loss curve for training data (train) and the loss curve for validation data (validation). At the beginning of the training, the loss value for both data was quite high, but as the epoch increased, the loss value decreased significantly. The loss curve for training data continues to decline sharply, reaching a lower value than the loss curve for validation data. This shows that the model is getting better at learning training data patterns, but there is still a slight gap between the performance in the training data and the validation data, which indicates overfitting. Overall, this graph illustrates the learning process of the model that is going well, with a consistent decrease in loss values during training.

3.2 Prediction Testing

The testing process is carried out after the model is saved. Testing is used to predict the classification of new documents of 500 data points. The results of the testing can be seen in the following evaluation graph:

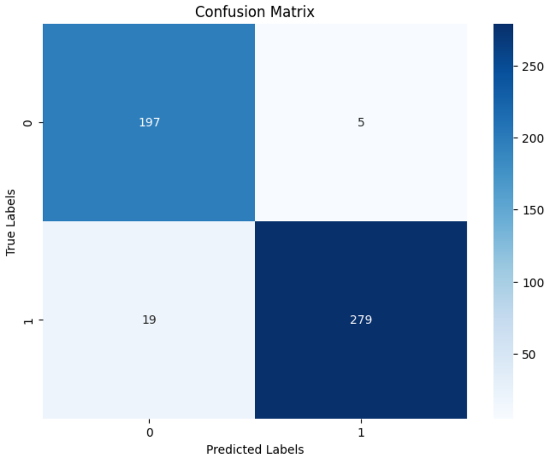


Figure 6. Confusin Matrix prediction

Figure 6 above shows the evaluation of testing on the new data. It can be seen that True Positive (TP): 279, True Negative (TN): 197, False Positive (FP): 5, False Negative (FN): 19. If calculated, it will yield the following values:

$$\text{Accuracy} = \frac{\text{TP} + \text{TN}}{\text{Total Samples}} = \frac{279 + 197}{279 + 197 + 5 + 19} = \frac{476}{500} = 0.952$$

$$\text{Precision} = \frac{\text{TP}}{\text{TP} + \text{FP}} = \frac{279}{279 + 5} = 0.982$$

$$\text{Recall} = \frac{\text{TP}}{\text{TP} + \text{FN}} = \frac{279}{279 + 19} = 0.936$$

$$\text{F1} - \text{Score} = 2\,x\,\frac{0.982 \text{ x } 0.936}{0.982 + 0.936} = 0.958$$

The model's prediction results showed excellent performance with an accuracy of 95.2%, accuracy of 98.2%, recall of 93.6%, and F1-score of 95.8%. The model successfully predicted the majority of the data correctly, with 197 non-bullying data (label 0) and 279 bullying data (label 1) correctly classified. However, there were several errors, namely 5 non-bullying data that were incorrectly predicted as bullying (false positives) and 19 bullying data that were not detected (false negatives). The high accuracy value indicates that the model rarely mispredicts non-bullying data as bullying, while a slightly lower recall indicates that there is still some bullying data that is missed. To further improve performance, especially on sensitivity to bullying data, the model can be improved by balancing the training data or optimizing hyperparameters. Overall, this model is quite reliable for the classification of bullying and non-bullying news.

4. DISCUSSION

In this study, the results obtained from the application of the BiLSTM model with various embedding combinations show that the use of RoBERTa significantly improves the performance of the model compared to BERT. Based on the results of the comparison, the BiLSTM model with RoBERTa embedding shows considerable improvements in terms of accuracy, recall, and precision when compared to the BiLSTM model using BERT. The BiLSTM+BERT model produces a recall value of 89%, precision of 86%, and accuracy of 89%, while the BiLSTM+RoBERTa model achieves a recall value of 90%, precision of 99%, and accuracy of 94%. This shows that the text representation of RoBERTa is more effective in capturing the context and nuances of news data, so that the model is better able to recognize relevant patterns for classification.

Furthermore, the use of normalization layers in the BiLSTM+RoBERTa model has been proven to provide a significant performance improvement. Models that add a normalization layer (BiLSTM+RoBERTa+Normalized) result in further improvements in evaluation metrics, with 94% recall, 98% precision, and 95% accuracy. Normalization helps in stabilizing the distribution of values across the output layer, allowing the model to learn more efficiently and improve overall performance. In addition, normalization techniques also help the model in avoiding the problem of exploitation or vanishing gradients, which often appear in multi-layered sequential models such as BiLSTM. The results of this study can be seen in Table 2.

Table 2. Training performance

BiLSTM	Recall	Precission	Accuracy	F1
Bert	89	86	89	90
RoBERTa	90	99	94	91
RoBERTa-Normalized	94	98	95	95

When compared to previous research conducted by Md. Mostafizer Rahman et al. [20], which also examined the combination of BiLSTM with RoBERTa embedding for sentiment analysis on Twitter, this study showed a significant improvement in performance. In Rahman's research, the model produced accuracy, precision, recall, and F1 scores of 82% each. Although these results are quite good, they are still not comparable to the performance obtained in this study, where the accuracy reaches 95%, recall 94%, and precision 98%. This difference may be due to several factors, including the complexity and context of the data, as well as the influence of adding layer normalization to the model used in this study. The normalization applied to this model not only optimizes training but also helps the model generate more accurate predictions in the classification of bullying news data, which has different characteristics compared to Twitter data, which may be shorter and more informal.

Thus, the application of RoBERTa embedding in combination with layer normalization in the BiLSTM model has been proven to provide superior performance compared to previous studies, especially in the context of news classification with complex and sensitive topics such as bullying.

5. CONCLUSION

Based on the results of this study, it can be concluded that the research objective to develop a bullying-themed online news classification system using a combination of RoBERTa and BiLSTM has been successfully achieved. With a dataset that includes news from reputable portals such as Kompas.com, Detik.com, and iNews.com, the BiLSTM model with RoBERTa embedding and normalization layer shows superior performance in detecting and classifying bullying news.

In the training stage, the model architecture, consisting of two layers of BiLSTM with different memory units, two layers of dropout, and a layer of normalization, successfully stabilized the learning process. The model's performance improvement was observed gradually until it reached stability at the 50th to 100th epoch, with an accuracy of 83%. In the testing phase using new data, the model recorded high performance with 95.2% accuracy, 98.2% precision, 93.6% recall, and 95.8% F1-score. Although there are still some false positive and false negative predictions, this performance proves that the model has high sensitivity and stability for complex bullying data classification.

This research has made a significant contribution in the field of **Natural Language Processing (NLP),** especially in the development of content detection technology for sensitive and emotional topics such as bullying. The application of **RoBERTa embedding** integrated with **BiLSTM** and a **normalization layer** has proven to be effective in improving classification performance with higher accuracy compared to previous approaches. These results can be a reference in building an automation system for news classification, supporting media or editorial platforms in filtering and identifying bullying-related content more accurately and efficiently.

For further development, several recommendations may be considered:

1. **Dataset Expansion**: Use a broader dataset and include news from international or cross-language sources to improve model generalization.
2. **Ensemble Method**: Exploring the ensemble approach by combining BiLSTM+RoBERTa with other models such as Transformer or CNN to further improve classification performance.
3. **Hyperparameter Optimization**: Apply optimization techniques such as *Bayesian Optimization* or *Grid Search* to find the best parameter configuration.
4. **Data Balancing**: Address data imbalances that can affect recall with techniques such as oversampling or *data augmentation*.

With these recommendations, further research is expected to further improve the model's performance and expand its application in various fields, such as the detection of harmful content in social media or automated news moderation systems.

BIBLIOGRAPHY

[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, Accessed: Nov. 07, 2024. [Online]. Available: http://elitejournal.org/index.php/ELITE/article/view/198

[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] L. B. Angizeh and M. R. Keyvanpour, “Detecting Fake News using Advanced Language Models: BERT and RoBERTa,” in *2024 10th International Conference on Web Research (ICWR)*, Apr. 2024, pp. 46–52. doi: 10.1109/ICWR61162.2024.10533352.

[8] Y. Sagama and A. Alamsyah, “Multi-Label Classification of Indonesian Online Toxicity using BERT and RoBERTa,” in *2023 IEEE International Conference on Industry 4.0, Artificial Intelligence, and Communications Technology (IAICT)*, Jul. 2023, pp. 143–149. doi: 10.1109/IAICT59002.2023.10205892.

[9] M. Fakhry and A. F. Brery, “A Comparison Study on Training Optimization Algorithms in the biLSTM Neural Network for Classification of PCG Signals,” in *2022 2nd International Conference on Innovative Research in Applied Science, Engineering and Technology (IRASET)*, Mar. 2022, pp. 1–6. doi: 10.1109/IRASET52964.2022.9738309.

[10] M. M. Rahman, A. I. Shiplu, Y. Watanobe, and M. A. Alam, "RoBERTa-BiLSTM: A Context-Aware Hybrid Model for Sentiment Analysis," Jun. 01, 2024, *arXiv*: arXiv:2406.00367. doi: 10.48550/arXiv.2406.00367.

[11] R. Mohawesh, H. B. Salameh, Y. Jararweh, M. Alkhalaileh, and S. Maqsood, "Fake review detection using transformer-based enhanced LSTM and RoBERTa," *International Journal of Cognitive Computing in Engineering*, vol. 5, pp. 250–258, Jan. 2024, doi: 10.1016/j.ijcce.2024.06.001.

[12] Q. Pu and F. Li, "RoBERTa-BiLSTM: A Chinese Stance Detection Model," in *2024 5th International Seminar on Artificial Intelligence, Networking and Information Technology (AINIT)*, Mar. 2024, pp. 2199–2203. doi: 10.1109/AINIT61980.2024.10581713.

[13] S. Zhang *et al.*, "Multi-Modal Emotion Recognition Based on Wavelet Transform and BERT-RoBERTa: An Innovative Approach Combining Enhanced BiLSTM and Focus Loss Function," *Electronics*, vol. 13, no. 16, Art. no. 16, Jan. 2024, doi: 10.3390/electronics13163262.

[14] A. Bhamre, "Detection of Cyberbullying using BiLSTM," *IJRASET*, vol. 12, no. 4, pp. 5161–5163, Apr. 2024, doi: 10.22214/ijraset.2024.60420.

[15] L. B. Angizeh and M. R. Keyvanpour, "Detecting Fake News using Advanced Language Models: BERT and RoBERTa," in *2024 10th International Conference on Web Research (ICWR)*, Apr. 2024, pp. 46–52. doi: 10.1109/ICWR61162.2024.10533352.

[16] H. Xu, X. Guo, and J. Zhao, "RoBERTa-wwm-CBA: A Mental Disease Identification Model Based on RoBERTa-wwm and Hybrid Neural Networks," in *Neural Computing for Advanced Applications*, Singapore: Springer Nature, 2025, pp. 246–258. doi: 10.1007/978-981-97-7001-4_18.

[17] T. Fu, A. Abbasi, D. Zeng, and H. Chen, "Sentimental spidering: leveraging opinion information in focused crawlers," *ACM Trans Inf Syst (TOIS)*, vol. 30, 2012, doi: 10.1145/2382438.2382443.

[18] 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.

[19] 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.

[20] M. M. Rahman, A. I. Shiplu, Y. Watanobe, and M. A. Alam, "RoBERTa-BiLSTM: A Context-Aware Hybrid Model for Sentiment Analysis," Jun. 01, 2024, *arXiv*: arXiv:2406.00367. Accessed: Nov. 20, 2024. [Online]. Available: http://arxiv.org/abs/2406.00367

[21] Y. Liu *et al.*, "RoBERTa: A Robustly Optimized BERT Pretraining Approach," Jul. 26, 2019, *arXiv*: arXiv:1907.11692. doi: 10.48550/arXiv.1907.11692.

[22] N. A. P. Masaling and D. Suhartono, "Utilizing RoBERTa and XLM-RoBERTa pre-trained model for structured sentiment analysis," *International Journal of Informatics and Communication Technology (IJ-ICT)*, vol. 13, no. 3, Art. no. 3, Dec. 2024, doi: 10.11591/ijict.v13i3.pp410-421.

[23] 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.

[24] A. S. Rahayu, A. Fauzi, and R. Rahmat, "Komparasi Algoritma Naïve Bayes Dan Support Vector Machine (Svm) Pada Analisis Sentimen Spotify," *Jurnal Sistem Komputer Dan Informatika (Json)*, vol. 4, no. 2, p. 349, 2022, doi: 10.30865/Json.V4i2.5398.

[25] R. Azhar and M. F. Wijayanto, "Analisis Sentimen di Twitter: Mengungkap Persepsi dan Emosi Publik Seputar Konflik Palestina-Israel," *Seminar Nasional Teknologi & Sains*, vol. 3, no. 1, Art. no. 1, Jan. 2024, doi: 10.29407/stains.v3i1.4132.

[26] S. Mujilahwati, M. Sholihin, R. Wardhani, and M. R. Zamroni, "Python Based Machine Learning Text Classification," *J. Phys.: Conf. Ser.*, vol. 2394, no. 1, p. 012015, Dec. 2022, doi: 10.1088/1742-6596/2394/1/012015.

[27] D. F. Sebastian, H. Sulistiani, and A. R. Isnain, "SENTIMENT ANALYSIS OF PUBLIC OPINION ON THE RIGHT OF INQUIRY IN INDONESIA IN 2024 USING THE SUPPORT VECTOR MACHINE (SVM) METHOD," *Jurnal Teknik Informatika (Jutif)*, vol. 5, no. 4, Art. no. 4, Jul. 2024, doi: 10.52436/1.jutif.2024.5.4.1968.

[28] E. Sari, L. Afuan, I. Permadi, E. Maryanto, and S. P. Rahayu, "CORRELATION ANALYSIS OF SENTIMENT OF 2024 ELECTION RESULTS AND STOCK MOVEMENTS OF POLITICAL ACTORS IN INDONESIA," *Jurnal Teknik Informatika (Jutif)*, vol. 5, no. 4, Art. no. 4, Aug. 2024, doi: 10.52436/1.jutif.2024.5.4.2701.

[29] M. Anjani and H. N. Irmanda, “COMPARISON PERFORMANCE OF WORD2VEC, GLOVE, FASTTEXT USING SUPPORT VECTOR MACHINE METHOD FOR SENTIMENT ANALYSIS,” *Jurnal Teknik Informatika (Jutif)*, vol. 5, no. 3, Art. no. 3, May 2024, doi: 10.52436/1.jutif.2024.5.3.1366.