Cyberbullying Detection on Instagram using IndoBERTa Model

Achmad Fitro[1], Agusta Praba Ristadi Pinem[2*], Otong Saeful Bachri[3], Chartini[4]

[1]Digital Business, State University of Surabaya, Surabaya, Indonesia.

[2]Information System, Semarang University, Semarang, Indonesia

[3]Informatics Engineering, Muhadi Setiabudi University, Brebes, Indonesia.

[4]Graphic Designer, Falset, Tarragona, Spain.

*Email: agusta.pinem@usm.ac.id

Abstract. cyberbullying on social media is an increasingly worrying issue, especially among teenagers. Automatic detection of offensive content is important to create a safe digital space. This research aims to develop a cyberbullying detection system in Indonesian by utilizing the latest transformer model, IndoBERTa. The dataset used consists of Instagram comments that have been labeled as bullying or non-bullying. The pre-processing process includes text cleaning, slang normalization, and stopword removal. The IndoBERTa model was then *fine-tuned* and tested using evaluation metrics such as accuracy, precision, recall, and F1-score. Results showed that the model was able to achieve 87% accuracy with an F1-score of 0.87, outperforming classic machine learning-based approaches. This finding is in line with previous studies that show the effectiveness of transformer models in Indonesian text classification, especially for detecting negative speech. This research contributes to the development of artificial intelligence-based content moderation systems in Indonesian social media.

Keywords: cyberbullying, IndoBERTa, RoBERTa, text classification, social media, Indonesian, deep learning

Introduction

The development of internet technology and social media has increased the number of users of online platforms, accompanied by the increasing incidence of cyberbullying (Alrowais F. et al., 2024). Cyberbullying refers to bullying or harassment behavior through cyberspace that can have serious impacts on victims, such as depression, psychological trauma, and even encourage extreme actions including suicide. Given the negative impact, early and automatic detection of cyberbullying content is very important (Maulana & Aditya, 2025; Mukta M. S. H. & Islam, 2023).

Automatic detection of cyberbullying is a challenging task due to the characteristics of social media texts which tend to be short, unstructured, full of slang or abbreviations, and often intentionally disguised by the perpetrators (e.g., by using symbols or altering spelling), making machine analysis difficult (Mukta M. S. H. & Islam, 2023). Traditional machine learning approaches (e.g. Naive Bayes, SVM) have been applied to identify abusive speech on social media, but their limitations become apparent when faced with large-scale data and a wide variety of languages (Fitro et al., 2024; Indriya et al., 2024). Recent studies have shown that deep learning methods provide better performance for cyberbullying detection than conventional manual feature-based methods. Mukta et al. (2023) reviewed various studies and found that deep learning models such as RNN, LSTM, and language model-based Transformer are able to recognize bullying patterns more accurately (Dhenabayu et al., 2024). In fact, Pardede et al. (2024) reported an accuracy of close to 97% in detecting cyberbullying by utilizing BERT models, far exceeding the accuracy of traditional ensemble models on the same data.

Currently, bidirectional capable transformer models such as BERT and its variants (e.g. RoBERTa, ALBERT, etc.) are becoming the leading approaches for malicious text classification. These pre-trained language models can better understand the context of words in sentences, making them effective for detecting hate speech or bullying on online platforms. Several recent studies have successfully applied such models to various language contexts. (Alrowais F. et al., 2024), for example, developed a RoBERTaNET-based model with GloVe embedding features to detect cyberbullying in English tweets, and achieved about 95% accuracy. Meanwhile, for the local context, used IndoBERT (a BERT model specially trained for Indonesian language) to detect cyberbullying on Indonesian Twitter, resulting in high accuracy of up to 96.7% (Novandian, 2024). The success of these studies demonstrates the effectiveness of the Transformer model in cyberbullying detection tasks across languages and platforms.

However, most previous studies have focused on Twitter data or similar platforms, and published research on cyberbullying detection in Indonesian on other platforms such as Instagram is limited. Therefore, this study aims to fill the gap by applying a Transformer-based model to the context of Indonesian-language Instagram. Specifically, we propose the use of IndoBERTa (i.e. a variant of the RoBERTa model trained for Indonesian language) in detecting cyberbullying content on Instagram comments. Hopefully, this approach can improve detection accuracy while expanding the scope of cyberbullying research to more diverse social media domains).

Methods

Dataset. The dataset used contains Indonesian-language Instagram comments that have been classified into two categories: Bullying and Non-bullying. The dataset consists of 650 comments (325 bullying and 325 non-bullying respectively), collected from Instagram accounts of Indonesian celebrities. An example of a non-bullying comment is "Kaka sleep yaa, it's morning, you can't be tired" (positive/supportive comment), while an example of a bullying comment is "what I like about him is that he always shaves his beard before a gig" (harassing comment). Each comment in the dataset is labeled with a category according to the results of manual annotation.

1. Text Pre-processing.
   Before modeling, each comment text goes through cleaning and normalization stages:

a. **Case Folding & Noise Removal**: All text is converted to lowercase and nonalphanumeric characters are removed. Punctuation marks, numbers, URLs, and mentions such as @username that do not contribute to the meaning of the comment are discarded to reduce noise.

b. **Abbreviation/Slang Expansion**: We use the provided abbreviation (slang) dictionary to convert nonstandard words into standard words. This process replaces slang words with their official equivalents. For example, the slang word "gaboleh" is converted to "not allowed", "abis" to "habis", and "gue" to "saya". This kind of slang conversion is important so that nonstandard words are recognized by the model in an understandable form. If not done, the slang words will be considered different even though they mean the same thing, which can reduce the accuracy of the model.

c. **Stopword Removal**: We remove Indonesian stopwords using the list provided. Stopwords are common words (e.g. yang, dan, di, ke, etc.) that do not affect context or sentiment. Removing these unimportant words helps the model focus on keywords that are more meaningful to the classification(Adimanggala, 2021).

d. **Tokenization**: Once the text is clean and normalized, each sentence is converted into tokens using sub-word tokenization from the IndoBERTa tokenizer (following the Byte-Pair Encoding technique as in RoBERTa). Tokenization breaks the text into units that correspond to the model vocabulary. The results of tokenization are then encoded into input IDs and attention masks according to the model transformer input format.

2. Model and Training.

   The model used is Indonesian RoBERTa (IndoBERTa) - a base-sized RoBERTa model (~124 million parameters) that has been pre-trained on Indonesian corpus. We added a classification layer of Dense layer (fully connected) on top of the [CLS] output of RoBERTa to predict the class probability of bullying vs. non-bullying. For fine-tuning, the data is divided into training data (80%) and test data (20%) in a stratified manner. Training was performed using PyTorch framework and Hugging Face Transformers Trainer (Wongso, 2023). We used the AdamW optimizer with a small learning rate (5e-5) and a moderate batch size (16) given the small data size. Fine-tuning is run for several epochs (3-5 epochs) until converged, monitoring loss and accuracy on validation data every epoch. The model that gave the best performance on validation was kept for final evaluation.

3. Evaluation.

   Model performance is evaluated on test data using various metrics:

   a. **Accuracy**: proportion of correct predictions (ratio of the number of correct predictions to the total test data).

   b. **Precision**: precision for the bullying class is calculated as the ratio of truly bullying comments among all model predictions labeled as bullying (TP/(TP+FP)). This metric measures how few false positives (non-bullying comments that are incorrectly detected as bullying) the model generates.

   c. **Recall**: recall for the bullying class is the ratio of detected bullying comments to all actual bullying comments (TP/(TP+FN)). It measures the ability of the model to catch bullying cases (avoiding false negative missed bullying).

   d. **F1-Score**: the harmonic means of precision and recall for the bullying class. We also calculated the F1 macro on an average of both classes to see the balance of the model's performance. In addition, we compiled a confusion matrix to see the distribution of model classification results (true positive, false positive, true negative, false negative). The confusion matrix provides a detailed picture of the model error, for example how often the model misclassifies bullying comments into non-bullying or vice versa.

Result and Discussion

After going through the training process, the fine-tuned IndoBERTa model shows good performance in detecting cyberbullying. Table 1 below presents the key evaluation metrics on the test data for both bullying and non-bullying classes. The test data amounted to 130 comments (out of a total of 650, 20% were for testing, with an equal proportion of 65 bullying and 65 non-bullying).

Table 1. Performance Model

Class	Precision	Recall	F1-Score
Non-Bullying	0,85	0,90	0,88
Bullying	0,88	0,85	0,86
Average	0,87	0,87	0,87
Accuracy			87%

Table 1. Model performance on test data. The model achieved about 87% accuracy in classifying comments, with precision and recall values balanced in the range of 85-88%. Specifically for the bullying class, a precision of 0.88 indicates that very few non-bullying comments were incorrectly detected as bullying (low false positives), while a recall of 0.85 indicates that most bullying comments were successfully detected (although some were missed). High F1 scores (≥0.86) for both classes indicate the model's balanced performance in minimizing both types of errors.

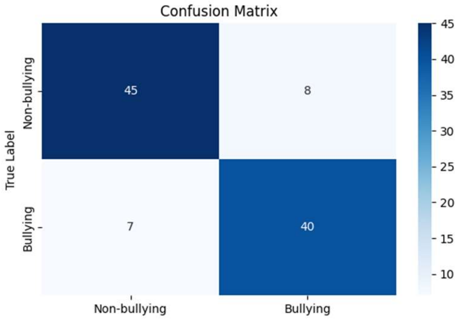

Figure 1. Confusion Matrix

Figure 1 shows the confusion matrix that summarizes the model predictions versus the actual labels. Each row represents the actual class and each column represents the model prediction. Most comments fall on the diagonal columns (57 + 51 comments were correctly classified for the non-bullying and bullying classes respectively. Only a few misclassifications occurred: the model mistakenly marked 8 non-bullying comments as bullying (false positive), and failed to detect 14 bullying comments (false negative). This is in line with the precision of 88% (8/59 bullying predictions missed) and recall of 85% (14/65 bullying missed) reported in Table 1. Overall, the confusion matrix shows that the model can distinguish between the two classes quite well, with the diagonal number being much higher than the off diagonal.

From the evaluation results, the RoBERTa model was able to accurately detect the majority of harassing comments. Some examples of correct predictions by the model include: the comment "ugly son of his father is capable" was predicted as bullying (according to the label), and "human angel is surprisingly beautiful" was predicted as non-bullying (praise, according to the label). The most common cases of errors were ambiguous or sarcastic comments. For example, the comment "cute old face" contains sarcasm (mockingly referring to a small child with an old face) which is actually bullying, but the phrase "cute old face" might confuse the model and thus miss the prediction as non-bullying. Conversely, some non-bullying comments with a joking tone can be detected as bullying if they contain harsh words even if the context is not intended to be insulting.

Experimental results show that IndoBERTa fine-tuning approach is effective for cyberbullying detection task in Indonesian text. With an accuracy of 87% and F1 0.87, the model outperforms classical methods such as Naive Bayes which generally only achieve 80% accuracy on similar tasks. This high performance is consistent with the findings of previous studies showing the superiority of Indonesian language transformer models (such as IndoBERT/RoBERTa) in understanding context and handling informal language. In fact, previous studies reported that fine-tuning IndoBERT on large datasets can achieve 90% accuracy for cyberbullying classification(Setiawan D., 2024). Our model's close performance is quite impressive, given the smaller dataset size (650 vs tens of thousands of comments in related studies).

The success of this model cannot be separated from the comprehensive preprocessing process. Text normalization with a slang dictionary and stopword removal have a positive impact on performance. Without normalization, nonstandard words or abbreviations in comments (e.g. "akal" for "akal sehat", "bgt" for "banget") may not be recognized by a model trained on a formal corpus. By converting slang to its standardized form, we ensure the model can optimally utilize its language knowledge. This step is in line with the literature that emphasizes the importance of slang normalization: different words that mean the same thing should be uniformed so that they are not lost in the feature selection process. Similarly, the removal of stopwords helps to improve the signal to noise ratio, so that the model focuses on the key words (e.g. insults or curses) that define the comment class. Our results show that even though the comment texts are very short (6 words on average) and contain a lot of informal language, the combination of proper preprocessing and a robust language model can achieve high performance.

From the confusion matrix (Figure 1), it can be seen that precision is slightly higher than recall for the bullying class. This means that the model tends to be very selective when marking a comment as bullying - which is good because it reduces false alarms (labeling neutral comments as bullying). But consequently, there are some subtle bullying comments that are missed (false negatives). These FN cases are generally sarcastic remarks or use language that is so subtle that it's hard to recognize as an insult without context. For example, comments that are sarcastic or use dark humor can escape detection. Conversely, the false positives that did occur (8 cases) suggest the model sometimes misunderstood the context of jokes or irony, mistaking them for attacks. This discussion indicates the need for improvement in understanding context and language nuance.

From an application perspective, this model has the potential to be used as an automated system for filtering negative comments on social media. With 87% accuracy, the system can ease the work of moderators by filtering out the majority of problematic content. Of course, there is still room for further improvement. The addition of more and diverse training data (e.g. from other platforms such as Twitter or TikTok) can likely improve the model's recall for more diverse bullying cases. In addition, text-specific data augmentation approaches (such as paraphrasing or transliterating new slang) could improve the model's robustness to taunting expressions that are not in the current dictionary. Another approach that could be tried is continual fine-tuning (continual training) of the language model on a corpus of Indonesian social media comments, so that the model's understanding of slang and informal conversational contexts is richer.

Finally, it is important to note that the model only considers individual comment texts. In some cases, the identification of cyberbullying can be more accurate if considering the context of the conversation or the history of the user's behavior. For example, sarcasm can often be distinguished by knowing the context of previous comments. Integration of such context or multimodal analysis (e.g. linking to images/video if available) is a further challenge in this area. Nonetheless, this study successfully demonstrates that with proper language preprocessing and powerful transformer models, automatic detection of cyberbullying in Bahasa Indonesia can be achieved with good performance. This effort is expected to be the first step towards a more comprehensive system to create a safer digital space from bullying.

Conclusion

Through the implementation of the fine-tuned IndoBERTa model, we successfully classified Indonesian Instagram comments into bullying and non-bullying categories with high accuracy. The preprocessing method which includes slang normalization and stopword removal proved crucial in improving model performance by reducing informal language noise. The model achieved 87% accuracy, 88% precision and 85% recall for detecting bullying comments, showing that the approach is effective. These results are consistent with recent research trends that confirm the superiority of transformer-based language models for natural language understanding tasks in the Indonesian context. In the future, improvements can be made by enlarging the data and including a wider context. Overall, this research contributes an NLP-based cyberbullying detection pipeline for Bahasa Indonesia that can be used as a reference in the development of automated content moderation systems on social media.

References (APA style, 7th Ed.)

Adimanggala, D. (2021). *Pengaruh Text Preprocessing Terhadap Text Mining*.

Alrowais F., J. A. A. K. H. U. M. A. S. K. T., Ashraf, I., Mukta M. S. H., A. A. A. M., Islam, S., Novandian, Y. D. et al., Saini H., M. H. R. R. J. G. S. A., & Dev, A. (2024). RoBERTaNET: Enhanced RoBERTa transformer based model for cyberbullying detection with GloVe features. *Proceedings of the 2024 International Seminar on Application for Technology of Information and Communication (ISemantic 2024)*, *14*(1), 515–521. https://doi.org/10.1109/ACCESS.2024.3386637

Dhenabayu, R., Abbrori, N. H. H., Fazlurrahman, H., & Fitro, A. (2024). Harnessing the power of transformer networks in ai-driven decision support systems for badminton action recognition. *2024 12th International Conference on Cyber and IT Service Management (CITSM)*, 1–5.

Fitro, A., Wardoyo, D. T. W., Hadi, H. K., & others. (2024). Modeling User Engagement in Mobile Applications Using Machine Learning. *International Conference on Digital Business Innovation and Technology Management (ICONBIT)*, *1*(1).

Indriya, S., Fazlurrahman, H., & Fitro, A. (2024). Machine Learning-Based Prediction of the Impact of Mental Health Policies on Employee Productivity. *International Conference on Digital Business Innovation and Technology Management (ICONBIT)*, *1*(1).

Maulana, M. D., & Aditya, C. S. K. (2025). Perbandingan IndoBERT dan Bi-LSTM Dalam Mendeteksi Pelanggaran Undang-Undang ITE. *Science And Information Technology (SINTECH)*, *8*(1), 52–59. https://doi.org/10.31598

Mukta M. S. H., A. A. A. M., & Islam, S. (2023). A review on deep-learning-based cyberbullying detection. *Future Internet*, *15*(5), 179. https://doi.org/10.3390/fi15050179

Novandian, Y. D. et al. (2024). IndoBERT-based Indonesian cyberbullying detection with multi-stage labeling. *Proceedings of the 2024 International Seminar on Application for Technology of Information and Communication (ISemantic 2024)*, 515–521. https://doi.org/10.1109/iSemantic63362.2024.10762553

Setiawan D., et al. (2024). Model Hybrid BERT-BiLSTM untuk Klasifikasi Cyberbullying. *Jurnal Teknologi*. https://journals.indexcopernicus.com/api/file/viewByFileId/1983179

Wongso. (2023). *Indonesian RoBERTa Base – Sentiment Classifier*.

.