Tri-RoBiAtt: A Hybrid Deep Learning Model for Effective Classification of Offensive Speech in Multilingual Corpora

1st Mei-Ling Huang
Department of Industrial Engineering & Management
National Chin-Yi University of Technology
Taiping, Taichung, Taiwan.
huangml@ncut.edu.tw

2nd Chih-Chen Yang
Department of Industrial Engineering & Management
National Chin-Yi University of Technology
Taiping, Taichung, Taiwan
3b015146@gm.student.ncut.edu.tw

***Abstract*—With the rapid development of social media, abusive and offensive speech online has become increasingly serious. To enhance the safety of cyberspace, this study proposes a novel hybrid model, Tri-RoBiAtt. Tri-RoBiAtt enriches language representations by combining the outputs of RoBERTa's last three layers, employs a dual-channel BiLSTM for language modeling, and uses a learnable attention mechanism to dynamically adjust weights according to context and focus on key tokens. In addition, we incorporate the LIME interpretability method to reveal the model's decision logic and thereby improve transparency and credibility. Experimental results show that Tri-RoBiAtt achieves 87% accuracy and an 83% macro-average F1 on the English OLID corpus, and 83% accuracy and an 83% macro-average F1 on the Chinese COLD corpus. These findings indicate that Tri-RoBiAtt can effectively capture subtle semantic shifts and attains strong classification performance on both English and Chinese datasets, providing an innovative and reliable technical solution for offensive speech classification.**

Keywords—Offensive text; Deep learning model; Social media; Transformer; Bidirectional Long Short-Term Memory; Local Interpretable Model-agnostic Explanations

I. Introduction

With the rapid popularization of social media, platforms such as Twitter, Facebook, Instagram, and emerging platforms such as Threads and TikTok have not only become important channels for obtaining information and exchanging opinions, but have also triggered the spread of hate and offensive speech [1].

Offensive speech poses serious harms at both individual and societal levels. It can trigger short-term anxiety, depression, anger, and helplessness and, if left unaddressed, may lead to long-term psychological trauma, particularly among psychologically immature adolescents who face elevated risks of self-harm or suicide [2]. At the societal level, hate speech directed at specific groups can inflame online tensions and, in extreme cases, precipitate real-world violence. To mitigate these risks, an automated offensive-speech classification system that leverages text-analysis techniques can detect and intercept harmful content in real time, limit its spread.

Recent research has focused on using deep learning models to detect hateful and offensive speech on social media. Khan et al. (2022) proposed BiCHAT, which combines BERT, CNN, and BiLSTM with a hierarchical attention mechanism to effectively identify hateful tweets [3]. Subsequent studies have introduced various hybrid models, such as Aljohani et al. (2024), who combined CNN, attention layers, and a random forest for Arabic text classification [4]; Mahajan et al. (2024) with EnsMulHateCyb, which leverages multiple RNN base models and heterogeneous fusion methods [5]; Kibriya et al. (2024), who used a multi-layer deep architecture combined with explainable AI techniques to analyze model performance [6]; Kothuru et al. (2025), who applied OLS feature selection with the SBiLSTM-MAM model to improve accuracy [7]; and Gashe et al. (2024), who created an Amharic hate speech dataset and used SBi-LSTM for classification [8].

This study proposes a hybrid model, Tri-RoBiAtt, which integrates multiple technologies for offensive text. The model combines the hidden states of the last three layers of RoBERTa and uses BiLSTM to capture the dependencies between texts. It also focuses on key information through the attention mechanism, improving the classification performance. To verify the effectiveness and stability of the method, we conducted training and testing on two public corpora, English and Chinese. In addition, in order to explore the internal operating mechanism of the model more deeply, This study adopts the Local Interpretable Model-agnostic Explanations (LIME)[9] to make the decision process more transparent and provide rich interpretation insights.

The main contributions of this study are as follows:

(1) A new hybrid model called Tri-RoBiAtt is proposed, which successfully integrates RoBERTa, BiLSTM, and attention mechanism.

(2) Comparison of English text models: we conducted an in-depth comparison of the RoBERTa, DistilBERT and ToxicBERT models for multi-class English offensive speech text to explore their respective advantages and disadvantages.

(3) Chinese text model analysis: A detailed comparison of Chinese texts was conducted to analyze the performance differences between the Chinese-RoBERTa-WWM-Ext, MacBERT, and ERNIE 3.0 models in offensive speech classification.

(4) Explore explainable AI methods such as LIME to reveal the decision logic within the model, thereby improving the transparency and credibility of the results.

II. PROPOSED TRI-ROBIATT MODEL

2.1 Architecture of the proposed model

This study conducted training and testing on two public corpora, English OLID and Chinese COLD. First, the original text was pre-processed for standardization. The English part included converting all letters to lowercase, expanding abbreviations (such as "can't" to "cannot"), removing URLs, punctuation and extra spaces, while the Chinese part reduced corpus noise by converting simplified to traditional Chinese, normalizing spaces, and appropriately segmenting Chinese words, removing URLs and punctuation to ensure the integrity of language features. Next, the processed English and Chinese texts are input into the pre-trained language models RoBERTa-base and HF/CHINESE-RoBERTa-WWM-EXT respectively to obtain deep semantic representations. In order to make full use of language features at different levels, the model extracts features from the last three hidden states of RoBERTa and concatenates them in the feature dimension, enabling it to integrate semantic information at different depths and improve text representation capabilities. The concatenated vector is then input into the BiLSTM layer, which learns the forward and backward dependencies, enabling the model to better capture the semantic coherence and structural information of the text. Furthermore, the importance weight of each time step is calculated through global soft attention, and the output of BiLSTM is weighted summed to obtain a context vector that can focus on key information. Finally, the vector is regularized by the Dropout layer to reduce the risk of overfitting, and finally classified by the fully connected layer. **Figure 1** shows the model architecture proposed in this study.

- Multi-layer feature fusion: Extract features from the last three hidden states of RoBERTa and concatenate them in dimensions to achieve effective fusion of semantic information at different depths.

- Sequence dependency modeling: BiLSTM is used to simultaneously capture the forward and backward dependencies of text, enhancing semantic coherence and structural information learning.

- Focus on key information: Through global soft attention, weighted summation is performed according to the importance of the time step to highlight key contextual information.

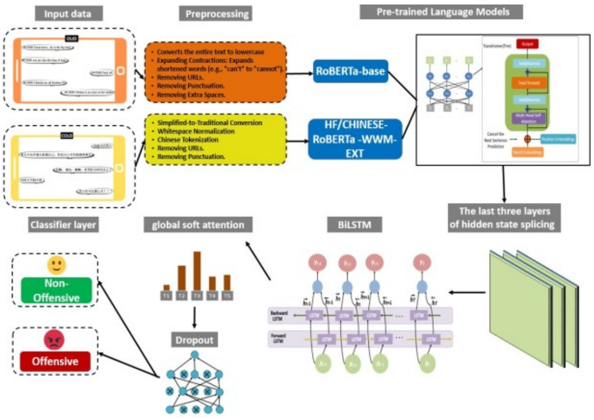

Figure 1 Architecture of study

2.2 Corpus

OLID (Offensive Language Identification Dataset) was created by SemEval-2019 Task 6 [10]. The main purpose of this dataset is to identify offensive language on Twitter. The training set contains 13,240 tweets and the test set contains 860 tweets, for a total of 14,100 tweets.

Model	Acc	Macro-Average		
		P	R	F1
RoBERTa	86%	84%	79%	81%
DistilBERT	85%	83%	78%	80%
ToxicBERT	83%	79%	80%	80%
Tri-RoBiAtt	87%	84%	83%	83%

COLD (Chinese Offensive Language Dataset) was collected and released by Deng *et al*. in 2022 [11] to study and analyze offensive language in Chinese social media. This dataset comes from texts on Chinese social platforms such as Weibo, covering a wide range of offensive speech, including insults, discrimination, hatred, etc. The training set has 25726 tweets and the test set has 5323 tweets.

2.3 Preprocessing

For English preprocessing, the text is first converted to lowercase to eliminate case sensitivity, and contractions.fix is applied to expand abbreviations (for example, “don’t” is transformed into “do not”) in order to preserve semantic integrity. Subsequently, URLs, punctuation marks, and redundant spaces are removed to produce cleaner input. For Chinese preprocessing, OpenCC is employed to convert simplified Chinese into traditional Chinese to ensure language consistency, followed by the use of regular expressions to normalize whitespace and remove URLs and punctuation. Finally, the jieba toolkit is applied for word segmentation, dividing the text into structured tokens suitable for subsequent analysis.

2.4 Models

This study compares its proposed method with baseline models in both English and Chinese. For English, RoBERTa improves upon BERT by focusing on masked language modeling, DistilBERT reduces complexity through knowledge distillation while retaining much of BERT’s performance, and ToxicBERT is fine-tuned to detect toxic and hateful content. For Chinese, Chinese-RoBERTa-WWM-Ext enhances learning with whole word masking, MacBERT introduces dynamic masking and improved word representations, and ERNIE 3.0 integrates external knowledge and cross-modal learning to strengthen contextual understanding.

2.4.1 The last three layers of hidden state concatenation

Use RoBERTa to return the last three layers $H_{\text{final}} = [H^{(12)}, H^{(11)}, H^{(10)}]$ to all hidden states and concatenate the vectors of these three layers along the last dimension $H_{\text{concat}} = [H^{(12)}; H^{(11)}; H^{(10)}] \in \mathbb{R}^{n\times(768\times3)}$. The dimension becomes: $H_{\text{concat}} \in \mathbb{R}^{n\times2304}$. This can fuse semantic features at different levels together to capture deeper language information, while also expanding the feature dimension and improving the performance of downstream tasks.

2.4.2 BiLSTM

Forward LSTM computes a new hidden state at each time step based on the current input and the hidden state of the previous time step. First, determine which new information should be stored in the cell state through input gate calculation; then determine which past information needs to be discarded through forget gate calculation; thirdly, determine what information to output from the cell state through output gate calculation; at the same time, use the tanh function to generate new candidate cell state information; combine with the forget gate and input gate results to update and retain necessary long-term dependency information; finally, calculate the final hidden state based on the updated cell state and output gate.

Backward LSTM processes the input in reverse order, and uses information from future time steps to make predictions. Its calculation steps are similar to forward LSTM, but in the opposite direction, allowing the model to capture the impact of future scenarios on the current state.

2.4.3 Global soft attention and Classifier layer

Dynamically assign different importance weights to the LSTM hidden state at each time step in the entire input sequence. First perform a linear transformation on each hidden state to calculate the attention score, where is the weight matrix in the attention mechanism and is the bias vector of the attention mechanism. Then use the softmax function to normalize all into probability distributions to ensure that each falls between 0 and 1 and the sum is 1; finally, a context vector that integrates the entire sequence information is obtained through weighted summation, dynamically adjust the model's attention allocation based on the needs of the current decoding or prediction step. Finally, the context vector after dropout is sent to the fully connected layer to calculate the probability distribution of the two categories.

III. Results

3.1 Results for OLID

This study uses RoBERTa, DistilBERT, ToxicBERT, and the proposed Tri-RoBiAtt model on the English dataset OLID to evaluate the performance of the modelS in Non-Offensive and Offensive text classification. **Table 1 s**hows the performance of the models in terms of overall accuracy, precision, recall and F1 score.

Among them, RoBERTa shows a high precision rate when identifying non-offensive text, but the recall rate for offensive text is relatively insufficient; DistilBERT uses knowledge distillation technology to compress the model size, but the overall performance is not better than other models; and ToxicBERT, which is fine-tuned for harmful content, does have advantages in identifying harmful content, but the balance between positive and negative samples still needs to be optimized. Finally, the Accuracy, Macro-Average-Precision, Macro-Average-Recall and Macro-Average-F1-score from the proposed Tri-RoBiAtt model were 87%, 84%, 83% and 83%, respectively, which further improved the ability to capture potentially offensive features in text.

Ref.	Acc.	Macro-Average		
		P	R	F1
[12]	85.3%	82.1%	80.8%	81.4%
[13]	74.5%	-	-	73.5%
[14]	84%	-	-	79%
[15]	77.89 %			72%
[16]	82%	81%	71%	73%
[7]	81.25%	73.91%	72.83%	72.77%
Tri-RoBiAtt	**87%**	**84%**	**83%**	**83%**

Table 1 The performance of the models on the English dataset OLID

3.2 Results for COLD

This study uses Chinese-RoBERTa-WWM-Ext, MacBERT, ERNIE 3.0, and the proposed Tri-RoBiAtt model on the Chinese dataset COLD. These models are pre-trained and optimized for Chinese language characteristics, aiming to capture the implicit semantics and subtle grammatical structures in Chinese texts. Experimental results **(Table 2)** show that each model achieves good recognition results when processing Chinese text, but in the classification of Non-Offensive and Offensive samples, their respective precision rate, recall rate, F1 score and other indicators are slightly different. Chinese-RoBERTa-WWM-Ext shows higher accuracy in identifying non-offensive text, while MacBERT has a better recall rate in capturing offensive remarks; ERNIE 3.0 uses knowledge integration technology to maintain stable classification performance in processing more complex contexts. The Tri-RoBiAtt model proposed in this study has Accuracy, Macro-Average-Precision, Macro-Average-Recall and Macro-Average-F1-score of 83%, 82%, 83% and 83% respectively, which further improves the ability to capture potentially offensive features in text.

Table 2 The performance of the models on the Chinese dataset COLD

Model	Acc	Macro-Average		
		P	R	F1
Chinese-RoBERTa-WWM-Ext	82%	82%	83%	82%
MacBERT	82%	82%	83%	82%
ERNIE 3.0	83%	82%	83%	82%
Tri-RoBiAtt	83%	82%	83%	83%

3.3 Comparison with SOTA

Table 3 shows a comparison of OLID among related literature. The Tri-RoBiAtt proposed in this study has reached 87%, 84%, 83% and 83% in accuracy, Macro-Average-precision, Macro-Average-recall, and Macro-Average-F1-score, respectively, in the English classification task. The results were better than methods such as CNN, RoBERTa fine-tuning, HateBERT, FastText embedded LSTM, BiGRU combined with FastText, and Random Forest used in other studies. Tri-RoBiAtt integrates multiple feature extraction and attention mechanisms in the model architecture, which can more comprehensively capture the semantic and emotional information in the text to improve the performance of classification.

Table 3 The comparison of OLID

Table 4 shows a comparison of COLD among related literature. The Tri-RoBiAtt proposed in this study achieved 83%, 82%, 83% and 83% in accuracy, Macro-Average-precision, Macro-Average-recall and Macro-Average-F1-score, respectively in the Chinses classification task, which is better than the models and methods used in other studies. The RoBERTa+Bi-GRU+ Multi-Head Attention proposed by Xu and Liu [17] also shows its advantages by leveraging RoBERTa's powerful pre-trained language understanding capabilities, BiGRU's extraction of pre- and post-sequence information, and the multi-head attention mechanism to capture features from multiple angles.

Table 4 The comparison of COLD

Ref.	Acc.	Macro-Average		
		P	R	F1
[11]	81%	80%	82%	81%
[17]	82.931%	82.26%	83.44%	82.84%
[18]	81.87%	-	-	79.09%
[19]	82.46%	-	-	-
[20]	80.65%	80.12%	81.36%	80.29%
Tri-RoBiAtt	**83%**	**82%**	**83%**	**83%**

3.4 Interpreting classification results using LIME

Figure 2 shows the visualization of LIME's explanation of the decision process of the English text classification model. The bar chart on the left side of the figure reflects the impact of each word on the final classification result: the red bar indicates that the word prompts the model to classify the text as Offensive, while the blue bar indicates that it tends to promote the judgment of Non-Offensive. As can be seen from the figure, the word fuck has the longest red bar, which clearly shows that it is the main factor that drives the sentence to be judged as offensive language; although other words such as type, me, saying and about also show positive effects, their influence is relatively light; in contrast, up, the, speak, of and cool have less influence, and even tend to reduce the judgment of aggressiveness. The probability graph on the right shows that the model considers the sentence to be offensive with a probability of 77%, which further verifies the distribution of the influence of each word in the bar graph on the left. This visual explanation not only helps us intuitively understand the model's sensitivity to keywords in the decision-making process, but also provides valuable basis for further optimizing the text classification model.

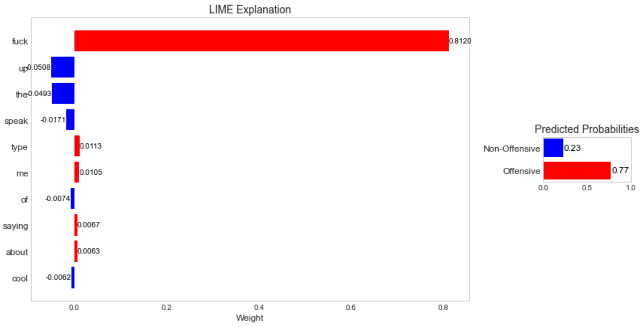

Figure 2 The visualization of LIME's explanation of the English text

Figure 3 shows the process of judging offensive Chinese text through the LIME explanation model. The word "disgusting" has the longest red bar, indicating that it has the greatest impact on the model's judgment of the text as offensive; "prodigal son" and "woman" also show positive effects, although the impact is less. The chart on the right shows the final prediction result of the sentence. The model judged the text as offensive with a probability of 99%, and non-offensive with only 1%. This is consistent with the influence distribution of sensitive words in the bar chart on the left, fully reflecting the degree of reliance of the model on keywords in the judgment process.

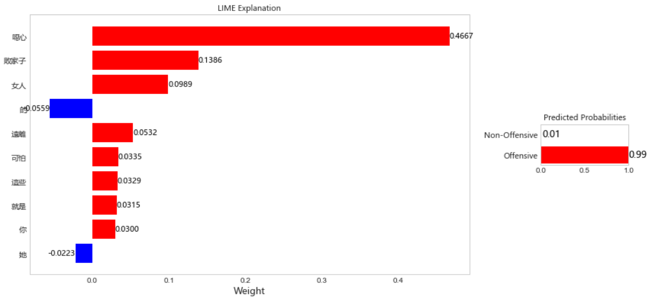

Figure 3 The visualization of LIME's explanation of the Chinese text

IV. Conclusion and Future Work

This study proposes an innovative hybrid model-Tri-RoBiAtt for the problem of aggressive language classification. This model combines the hidden states of the last three layers of the RoBERTa model, the bidirectional long short-term memory network (BiLSTM), and the attention mechanism to more accurately capture subtle changes in semantics. This combination takes advantage of deep context embedding and enhances the model's ability to model dependencies in language sequences, thereby demonstrating higher accuracy and robustness than other models when processing complex and varied expressions in language.

Detailed tests were conducted on two data sets, OLID and COLD. On the OLID dataset, the Tri-RoBiAtt model achieved an accuracy of 87% and a macro-average F1 score of 83%; while in the COLD dataset, the model also achieved an accuracy of 83% and a macro-average F1 score of 83%, both significantly better than the existing baseline model. These results fully demonstrate that by combining RoBERTa's deep context understanding capabilities, BiLSTM's sequence dependency modeling, and the role of the attention mechanism in key information extraction, the performance of offensive language detection can be significantly improved. This study also introduces the LIME (Local Interpretable Model-agnostic Explanations) explanatory method. In-depth analysis of the decision-making logic made by the model when processing input text reveals which words or sentences play a decisive role in the final classification result, improving the transparency of the model. This study not only proposes a hybrid model that performs well in offensive language classification tasks, but also improves the transparency and credibility of the model by introducing interpretive technology, providing an innovative and effective technical solution for online hate speech detection.

REFERENCES

[1] B. Mathew, R. Dutt, P. Goyal, and A. Mukherjee, “Spread of hate speech in online social media,” in Proc. 10th ACM Conf. Web Sci., pp. 173–182, 2019. doi:10.1145/3292522.3326034.

[2] R. J. Boeckmann and J. Liew, “Hate speech: Asian American students' justice judgments and psychological responses,” J. Soc. Issues, vol. 58, no. 2, pp. 363–381, 2002. doi:10.1111/1540-4560.00265.

[3] S. Khan, M. Fazil, V. K. Sejwal, M. A. Alshara, R. M. Alotaibi, A. Kamal, and A. R. Baig, “BiCHAT: BiLSTM with deep CNN and hierarchical attention for hate speech detection,” J. King Saud Univ. Comput. Inf. Sci., vol. 34, no. 7, pp. 4335–4344, 2022. doi:10.1016/j.jksuci.2022.05.006.

[4] A. Aljohani, N. Alharbe, R. E. Al Mamlook, and M. M. Khayyat, “A hybrid combination of CNN attention with optimized random forest with grey wolf optimizer to discriminate between Arabic hateful, abusive tweets,” J. King Saud Univ. Comput. Inf. Sci., vol. 36, no. 2, art. no. 101961, 2024. doi:10.1016/j.jksuci.2024.101961.

[5] E. Mahajan, H. Mahajan, and S. Kumar, “EnsMulHateCyb: Multilingual hate speech and cyberbully detection in online social media,” Expert Syst. Appl., vol. 236, art. no. 121228, 2024. doi:10.1016/j.eswa.2023.121228.

[6] H. Kibriya, A. Siddiqa, W. Z. Khan, and M. K. Khan, “Towards safer online communities: Deep learning and explainable AI for hate speech detection and classification,” Comput. Electr. Eng., vol. 116, art. no. 109153, 2024. doi:10.1016/j.compeleceng.2024.109153.

[7] S. Kothuru and A. Santhanavijayan, “Orthogonal least square based feature selection for an automatic hate speech detection and classification,” Comput. Electr. Eng., vol. 123, art. no. 110131, 2025. doi:10.1016/j.compeleceng.2025.110131.

[8] S. M. Gashe, S. M. Yimam, and Y. Assabie, “Hate speech detection and classification in Amharic text with deep learning,” arXiv preprint arXiv:2408.03849, 2024. (CorpusID: 271744934)

[9] M. T. Ribeiro, S. Singh, and C. Guestrin, “Why should I trust you? Explaining the predictions of any classifier,” in Proc. 22nd ACM SIGKDD Int. Conf. Knowl. Discov. Data Min., pp. 1135–1144, 2016. doi:10.1145/2939672.2939778.

[10] M. Zampieri, S. Malmasi, P. Nakov, S. Rosenthal, N. Farra, and R. Kumar, “SemEval-2019 task 6: Identifying and categorizing offensive language in social media (OffensEval),” arXiv preprint arXiv:1903.08983, 2019. doi:10.18653/v1/S19-2010.

[11] J. Deng, J. Zhou, H. Sun, C. Zheng, F. Mi, H. Meng, and M. Huang, “COLD: A benchmark for Chinese offensive language detection,” arXiv preprint arXiv:2201.06025, 2022. doi:10.48550/arXiv.2201.06025.

[12] M. Sharma, I. Kandasamy, and V. Kandasamy, “Deep learning for predicting neutralities in offensive language identification dataset,” Expert Syst. Appl., vol. 185, art. no. 115458, 2021. doi:10.1016/j.eswa.2021.115458.

[13] D. R. Beddiar, M. S. Jahan, and M. Oussalah, “Data expansion using back translation and paraphrasing for hate speech detection,” Online Soc. Netw. Media, vol. 24, art. no. 100153, 2021. doi:10.1016/j.osnem.2021.100153.

[14] N. Badri, F. Kboubi, and A. Habacha Chaibi, “Combining FastText and GloVe word embedding for offensive and hate speech text detection,” Procedia Comput. Sci., vol. 207, pp. 769–778, 2022. doi:10.1016/j.procs.2022.09.132.

[15] N. Oswal, “Identifying and categorizing offensive language in social media [Preprint],” arXiv, 2021. [Online]. Available: https://arxiv.org/abs/2104.04871.

[16] M. U. Ali and R. Lefticaru, “Detection of cyberbullying on social media platforms using machine learning,” in Advances in Computational Intelligence Systems, pp. 220–233, 2024. doi:10.1007/978-3-031-47508-5_18.

[17] M. Xu and S. Liu, “RB_BG_MHA: A RoBERTa-based model with Bi-GRU and multi-head attention for Chinese offensive language detection in social media,” Appl. Sci., vol. 13, no. 19, art. no. 11000, 2023. doi:10.3390/app131911000.

[18] L. Zhou, L. Cabello, Y. Cao, and D. Hershcovich, “Cross-cultural transfer learning for Chinese offensive language detection,” in Proc. First Workshop Cross-Cultural Considerations NLP (C3NLP), pp. 8–15, 2023. doi:10.18653/v1/2023.c3nlp-1.2.

[19] B. Peng, K. Han, L. Zhong, S. Wu, and T. Zhang, “A head-to-head attention with prompt text augmentation for text classification,” Neurocomputing, vol. 595, art. no. 127815, 2024. doi:10.1016/j.neucom.2024.127815.

[20] B. Hou, X. Xie, D. Zhang, L. Zheng, and G. Yan, “Chinese offensive language detection algorithm based on pre-trained language model and pointer network augmentation,” in 2024 5th Int. Seminar Artif. Intell., Netw. Inf. Technol. (AINIT), pp. 800–805, 2024. doi:10.1109/AINIT61980.2024.10581762.