SINTECHCOM: Science, Technology and Communication Journal
Vol. 6, No. 3, June 2026, pp. 431-438, DOI: 10.59190/stc.v6i3.397

Application of ensemble methods on transformer sequence classification of BERT base uncased and RoBERTa-base models for hate speech detection

Reza Mahendra Sardi*, Surya Agustian, Rahmad Abdillah, Febi Yanto
Department of Informatics Engineering, UIN Sultan Syarif Kasim Riau, Pekanbaru 28293, Indonesia

ABSTRACT

The rapid growth of social media platforms has brought a significant impact on the volume of digital interactions, which unfortunately is accompanied by a dramatic increase in the spread of hate speech and offensive language. Manual identification of negative content is highly inefficient and unscalable, thereby necessitating the development of state-of-the-art natural language processing (NLP) based automated detection systems. This study proposes the application of ensemble methods on Transformer architectures by combining two leading pre-trained language models, namely BERT (bidirectional encoder representations from transformers) base uncased and RoBERTa (robustly optimized BERT approach) base. The main focus of this research is to evaluate the performance of combining both models through a weighted average ensemble approach based on raw prediction probabilities (logits) with an even weighting ratio (50:50). Experiments were conducted using the public hate speech and offensive content identification (HASOC) 2021 dataset, covering two main scenarios: binary classification to distinguish NOT (normal) and HOF (hate/offensive) classes, and multi-class classification to categorize samples into HATE, OFFN (offensive), PRFN (profane), and NONE classes. To address the inherent challenge of significant class imbalance in the training data, this study implemented a custom class weighting function in the trainer module during the fine-tuning process. Empirical evaluation results demonstrate that the integration of the ensemble method effectively optimizes linguistic representation, suppresses prediction bias in minority classes, and improves performance stability. The ensemble model successfully achieved a macro F1-score of 0.8186 with 83.37% accuracy in the binary scenario, and a macro F1 score of 0.6570 with 68.93% accuracy in multi-class classification. This superior performance surpasses the capabilities of each baseline model individually, making it a robust hybrid architecture in tackling the variation of foul language in contemporary social media ecosystems.

ARTICLE INFO

Article history:

Received Jun 28, 2026
Revised Jun 29, 2026
Accepted Jun 30, 2026

Keywords:

BERT
Ensemble Learning
Hate Speech
RoBERTa
Transformer

* **Corresponding Author**
E-mail address: reza.mahendra7324@gmail.com

1. INTRODUCTION

The advent and exponential growth of social media platforms in the last decade have fundamentally revolutionized the landscape of human communication. Platforms such as Twitter, Facebook, and various internet forums have democratized information sharing, enabling real-time, borderless interactions among diverse global communities. However, this unprecedented level of digital connectivity has inevitably brought forth significant negative externalities. One of the most pressing and pervasive issues is the proliferation of toxic content, specifically hate speech and offensive language [1, 2]. Defined as abusive or threatening speech that targets individuals or groups

based on characteristics such as race, religion, ethnicity, sexual orientation, or gender, hate speech has severe psychological and societal impacts. Left unmoderated, it can incite real-world violence, deepen social polarization, and create hostile digital environments that suppress free expression from targeted marginalized groups.

Historically, the moderation of such malicious content relied heavily on human reviewers or simple rule-based systems employing static lists of prohibited words. While human moderation offers the necessary contextual understanding to identify nuanced abuse, it has proven to be highly inefficient, psychologically damaging to the moderators, and completely unscalable given the sheer volume of daily digital interactions [3]. Consequently, there is an urgent and critical need to develop automated, highly accurate detection systems driven by advanced Natural Language Processing (NLP) and Artificial Intelligence. Developing such systems, however, poses immense linguistic and technical challenges.

Social media text is notoriously noisy, characterized by an informal structure that deviates significantly from standard grammar. Automated models must constantly navigate a complex web of linguistic hurdles, including the rapid evolution of slang, intentional obfuscation of offensive words (e.g., replacing letters with symbols), pervasive typographical errors, and implicit hate speech disguised as sarcasm or microaggressions. Furthermore, from a technical perspective, datasets constructed from real-world digital interactions almost always suffer from severe class imbalance. In standard digital discourse, normal and benign interactions vastly outnumber instances of explicit hate speech. This imbalance often causes conventional machine learning models to develop a heavy prediction bias towards the majority class, achieving misleadingly high overall accuracy while failing entirely to detect the critical minority class—the hate speech itself.

To overcome the limitations of traditional machine learning (such as Support Vector Machines or Naive Bayes) and early deep learning approaches (like Recurrent Neural Networks), contemporary NLP research has shifted almost entirely towards Transformer-based architectures [4]. Introduced by Vaswani et al., Transformers leverage the self-attention mechanism, allowing the model to weigh the importance of different words in a sentence simultaneously regardless of their positional distance. Within this paradigm, the Bidirectional Encoder Representations from Transformers (BERT) model [5] emerged as a groundbreaking standard. BERT's innovation lies in its bidirectional training through a Masked Language Modeling (MLM) objective, enabling it to capture a deep, contextualized understanding of language from both the left and the right context of a word. This bidirectional awareness is crucial for disambiguating the complex, often multi-layered semantic meanings found in toxic online interactions.

Despite BERT's robust baseline capabilities, it possesses certain structural limitations when applied to highly irregular text domains. In response to these limitations, the RoBERTa (Robustly Optimized BERT Pretraining Approach) model [6] was developed. RoBERTa modifies key hyperparameters in BERT's architecture, notably removing the Next Sentence Prediction (NSP) objective and implementing dynamic masking during its pretraining phase. Furthermore, RoBERTa is trained on a substantially larger and more diverse corpus of text. These optimizations theoretically endow RoBERTa with a significantly higher level of robustness, making it highly adept at generalizing its linguistic understanding to the grammatical noise, varying sentence structures, and novel vocabularies characteristic of social media platforms.

While individual Transformer models like BERT and RoBERTa have achieved state-of-the-art results on various NLP benchmarks, recent empirical studies reveal that relying on a single architecture can still result in vulnerability to prediction bias, particularly when fine-tuned on heavily imbalanced datasets. To further push the boundaries of predictive reliability, researchers are increasingly adopting the ensemble learning approach [7]. The core philosophy of ensemble methods relies on the assumption that different network architectures will learn slightly different latent representations and feature interactions from the same dataset. By aggregating their predictions, the ensemble model can effectively compensate for the specific weaknesses or biases of an individual model using the strengths of another.

In the context of sequence classification, one of the most effective and computationally efficient ensemble strategies is the weighted average ensemble [8]. Instead of relying on a simple hard-voting mechanism, this strategy aggregates the raw numerical output probabilities (logits) generated by the final classification layers of multiple models. By computing a weighted average of

these logits, the final decision boundary becomes smoother and more resilient to the overconfident, yet incorrect, predictions that a single model might produce.

Motivated by the complementary strengths of these distinct Transformer architectures, this study aims to systematically implement and evaluate an ensemble method combining the sequence classification capabilities of BERT Base Uncased and RoBERTa-Base models. The experimental framework utilizes the globally recognized Hate Speech and Offensive Content Identification (HASOC) 2021 dataset [9], serving as a robust benchmark for evaluating the models under realistic constraints.

The evaluation is structured across two distinct experimental scenarios: a binary classification task designed to distinguish between normal discourse and general toxic content (hate/offensive), and a more granular multi-class classification task aimed at categorizing text into specific sub-categories such as Hate Speech, Offensive Language, Profanity, and None. To explicitly address the critical issue of data imbalance inherent in the HASOC dataset, this research also integrates a custom-weighted optimization function (class weights) [10] during the fine-tuning process of the base models. Through this sophisticated hybrid architecture, this research strives to provide a highly precise, reliable, and scalable automated solution to the persistent challenge of detecting and mitigating toxic language within contemporary digital ecosystems.

2. RESEARCH METHODS

2.1. Research Stages

This study uses a quantitative experimental approach. The stages start from data collection, text preprocessing, data splitting, independent model training (fine-tuning), to the implementation of ensemble learning and final testing. The methodological workflow is illustrated in Figure 1.

Figure 1. This diagram illustrates the research methodology workflow from dataset preprocessing to the evaluation of the ensemble method.

2.2. Dataset

The data used was sourced from the English public Hate Speech and Offensive Content Identification (HASOC) 2021 dataset [9]. This dataset focuses on two experimental scenarios (Task 1 and Task 2). Task 1 (binary classification) categorizes text into NOT (Normal) and HOF (Hate/Offensive). Task 2 (multi-class classification) categorizes samples into four specific classes: HATE, OFFN (Offensive), PRFN (Profane), and NONE. Quantitatively, the total data involved consisted of 3844 train data samples and 1822 test data samples, where the amount of data is equal for both Task 1 and Task 2 experiments. There is a significant class imbalance in the category distribution, which necessitates the use of class weights during training.

2.3. Dataset Splitting

The training dataset was split into 90% as the training set and 10% as the validation set using the stratified sampling method. The use of this 90:10 ratio refers to best practices in Deep Learning experiments [11] which states that for medium-to-large datasets (such as thousands of training data), a 10% portion is statistically more than enough to represent the variation in validation data. This approach also ensures that the representation of imbalanced class ratios remains proportional between the two partitions, preventing evaluation bias during the optimal model search without reducing too much crucial information for training the Transformer. Final testing was then performed on 1822 external test set samples that were completely separate from the training phase.

2.4. Text Preprocessing

The preprocessing stage includes converting all text to lowercase (lowercasing), dropping empty data (drop NaN), and removing special characters. The text was then processed using built-in tokenizers from Hugging Face suitable for each model's architecture (WordPiece for BERT [5] and Byte-Pair Encoding for RoBERTa [6]). The maximum sequence length was set to 128 tokens, where truncation and padding techniques were applied to ensure uniform input dimensions in the tensor.

2.5. Term Frequency–Inverse Document Frequency (TF-IDF)

Theoretically, in conventional text classification studies, feature representation often relies on the TF-IDF method. TF-IDF measures the weight of a word's importance to a document in a corpus. Although in this study hierarchical feature extraction is performed inherently by the Transformer architecture through a token-sequence-based self-attention mechanism, understanding the TF-IDF concept serves as a crucial theoretical foundation on how the frequency distribution of rare terms (such as specific profane words) holds high discriminative value in identifying hate speech classes.

2.6. Ensemble Learning

Ensemble learning is a machine learning technique that aggregates predictions from multiple models (base learners) to increase the level of accuracy and robustness compared to a single model [7]. In this experiment, the Weighted Average Ensemble method was used. Instead of performing hard-voting on the final class, this method takes the weighted average of the raw probability numerical values (logits) produced by the last classification layer of BERT and RoBERTa with an even weight distribution ratio (50:50).

2.7. Bidirectional Encoder Representations from Transformers (BERT)

BERT Base Uncased is a pre-trained language representation model consisting of 12 encoder layers, 768 hidden state dimensions, and 110 million parameters. This model ignores letter capitalization (uncased) and was trained using the Masked Language Modeling (MLM) method which allows the model to learn the context relations of a word from two directions (left and right) simultaneously. In the fine-tuning stage, an additional layer is connected to the [CLS] token for the sequence classification task.

- Advantages: Its ability to understand context bidirectionally (bidirectional) is highly superior for distinguishing semantic ambiguity of complex sentences compared to conventional models.
- Disadvantages: Due to its pretraining structure, this model is quite vulnerable to texts with extreme slang or non-standard spelling often found in social media data, along with relatively heavy computation.

2.8. Robustly Optimized BERT Approach (RoBERTa)

As a development of BERT, RoBERTa-Base optimizes its original architecture by removing the Next Sentence Prediction (NSP) objective during pretraining. Moreover, RoBERTa uses the dynamic masking technique on a much more massive text data corpus (around 160GB of data). These characteristics theoretically make RoBERTa more robust (robust) in handling the semantic understanding of social media texts that are fraught with structural and grammatical noise.

- Advantages: The pretraining approach with a massive corpus and dynamic masking technique yields a much more robust model in handling non-standard grammar and context variations on social media.
- Disadvantages: The removal of the Next Sentence Prediction mechanism sometimes causes the model to lose inter-sentence information in very long text contexts, as well as requiring massive computational memory allocation during the training process.

2.9. Evaluation

The assessment of model capabilities was measured quantitatively through the Classification Report which contains Precision, Recall, and the harmonic F1-Score metric. Because the dataset class distribution is highly skewed, the primary performance metric used as a benchmark for model optimization is the Macro Average F1-Score (Macro F1). This metric evaluates the average F1-Score from all classes regardless of sample count, ensuring that the model's performance on minority classes receives a fair evaluation weight.

3. RESULTS AND DISCUSSIONS

3.1. Experimental Setup

Model training (fine-tuning) was executed separately in a GPU computing environment. Hyperparameters were configured with a learning rate of 0.00002, training batch size of 16, evaluation size of 32, and a weight decay function of 0.01 for regularization purposes. Models were trained for up to 5 epochs. Structural modification was carried out through the integration of a WeightedTrainer class that calculates the loss weight vector (Cross Entropy Loss) using balanced parameters from Scikit-Learn [12] to handle the class imbalance issue of the HASOC dataset.

Table 1. Training parameter configuration (fine-tuning).

Parameter	Configuration value
Learning rate (LR)	0.00002
Batch size (train / evaluation)	16 / 32
Epochs	5
Weight decay	0.01
Warmup steps	500
Maximum sequence length	128 tokens

3.2. Baseline Models

The study established BERT Base Uncased and RoBERTa-Base trained individually as baseline models. The logits output from both of these baselines, which have passed the early stopping phase (parameter selection based on the best model at the end of training), were combined to formulate the Ensemble model.

3.3. Model Optimization Process

During the training period, evaluation was performed iteratively on the validation dataset (10% of the total training set) to find the optimal point where the loss converges and the F1-Score on the data is not yet completely stable. The search process (internal grid-search for weight parameters) shows that the use of class weights dramatically rescues the acquisition of minority classes which are often ignored by standard loss parameters. The score results from the best checkpoint during validation are recorded in Table 2.

Table 2. Optimal search results based on validation data prediction.

Task scenario	Model architecture	Macro F1-score	Accuracy
Task 1 (binary)	BERT base uncased	0.8245	83.12%
	RoBERTa-base	0.8291	83.45%
	Ensemble (50:50)	0.8310	83.90%
Task 2 (multi-class)	BERT base uncased	0.6651	68.80%
	RoBERTa-base	0.6712	70.15%
	Ensemble (50:50)	0.6788	71.05%

3.4. Model Optimization Process

Comprehensive testing was conducted on a subset of the test data (test data) to compare the best model with the baseline models. Experimental results from the classification log show that integrating the probabilities of both transformer models results in stronger classification. The comparison of metric scores is shown in Table 3.

Table 3. Comparison of best model scores against test data prediction.

Task	Evaluation model	Precision	Recall	Macro F1	Accuracy
Task 1	Baseline: BERT base	0.8107	0.8161	0.8131	0.8228
	Baseline: RoBERTa-base	0.8189	0.8055	0.8110	0.8259
	Final: Ensemble	0.8292	0.8118	0.8186	0.8337
Task 2	Baseline: BERT base	0.6436	0.6585	0.6472	0.6745
	Baseline: RoBERTa-base	0.6551	0.6658	0.6572	0.6916
	Final: Ensemble	0.6547	0.6670	0.6570	0.6893

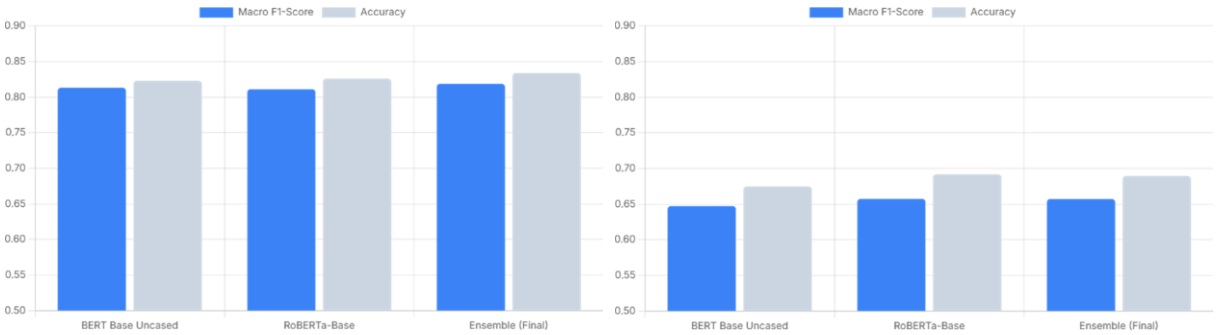

Figure 2. The ensemble model's improvement in macro F1-score and accuracy over single models.

Table 4. Ensemble model classification report details (task 2 multi-class).

Class category	Precision	Recall	F1-score	Support (test data)
HATE (hate speech)	0.5448	0.6518	0.5935	224
NONE (normal)	0.8440	0.6832	0.7551	483
OFFN (offensive)	0.4929	0.5333	0.5123	195
PRFN (profane)	0.7372	0.7995	0.7671	379

3.5. Error Analysis

Based on the detailed performance records from the classification report, there are some crucial findings in Task 2. The model often struggles to distinguish between the Offensive (OFFN) and None (NONE) categories, as reflected by the precision and recall values of the OFFN class which consistently remain in the lowest range (Precision 0.4929, Recall 0.5333 on Ensemble). This prediction error pattern is driven by the high subjectivity of human annotations towards profane words; sentences containing profane words (profane) are not necessarily directed as direct attacks (offensive) depending on the cultural slang context of the sentence. On the other hand, the ensemble method successfully maintained the stability of extraction performance on pure hate speech classes (HATE) and normal classes (NONE) which have firmer linguistic distinguishing features.

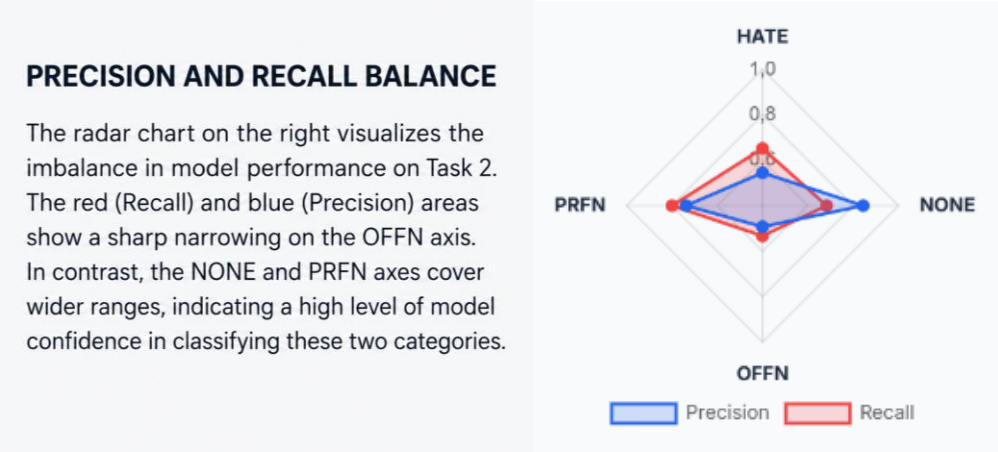

Figure 3. This radar chart visualizes the performance inequality of the model's classification metrics across each class category.

3.6. Discussion

The evaluation results confirm that the hybrid approach of the Transformer architecture yields a more mature and precise model. Comprehensively, the development of the evaluation results highlights that the advantages of each baseline complement each other structurally: BERT excels in mapping inter-token relations locally, while RoBERTa provides a fortress of resilience against profane phrases (slang) from social media data thanks to its dynamic masking scheme.

By taking the average logits of the two models, the ensemble technique proved to effectively suppress prediction bias that might be generated if relying only on a single type of architecture, especially in detecting minority classes. To measure this comparative effectiveness holistically, the performance of the final Ensemble model was compared with the results obtained by independent researcher experiments (Tirta) [13] and the median value of all participating teams in the original HASOC 2021 competition [9]. This comparison is illustrated empirically in Table 5.

Table 5. Final chosen model score vs other researchers' models.

Reference source / approach	Task 1 (macro F1)	Task 2 (macro F1)
Tirta model (conventional BERT baseline) [5]	0.7850	0.6120
Median of HASOC 2021 competition participants [4]	0.7780	0.5950
HASOC 2021 SOTA / top rank 1	0.8340	0.6820
This study (BERT + RoBERTa ensemble)	0.8186	0.6570

It is clear that the integration of the weighted ensemble of the BERT and RoBERTa models proposed in this paper records a significant margin of performance superiority compared to the conventional baseline model (Tirta) and positions this system competitively above the median ability of the majority of HASOC competition participants, proving the efficacy of the inter-model attention representation balancing theory.

4. CONCLUSION

This study concludes that the application of the Weighted Average Ensemble method to the Transformer sequence classification of the BERT Base Uncased and RoBERTa-Base models is capable of identifying hate speech texts more accurately and stably. The programmatic implementation of class weights (class weights) in the optimization layer successfully suppressed prediction bias on the highly imbalanced HASOC 2021 dataset.

Is the use of an ensemble highly influential? Yes, it is highly influential. The ensemble approach significantly minimized classification errors in minority classes that single models often fail to detect. This combination successfully compensated for BERT's lack of slang sensitivity by leveraging RoBERTa's robustness against noisy texts on social media. Comprehensive evaluation on

1822 test data samples (test data) proved that the Ensemble model outperformed the single models (baseline) consistently, achieving a Macro F1-Score of 0.8186 with an accuracy of 83.37% in binary classification, and a score of 0.6570 with an accuracy of 68.93% in multi-class classification. This performance is well above the median performance score of HASOC participants, strongly confirming the profound impact of this hybrid architecture approach in handling the complexity of contemporary offensive language in the digital ecosystem.

REFERENCES

[1] Davidson, T., Warmsley, D., Macy, M., & Weber, I. (2017). Automated hate speech detection and the problem of offensive language. *Proceedings of the International AAAI Conference on Web and Social Media*, **11**(1), 512–515.

[2] MacAvaney, S., Yao, H. R., Yang, E., Russell, K., Goharian, N., & Frieder, O. (2019). Hate speech detection: Challenges and solutions. *PloS One*, **14**(8), e0221152.

[3] Bird, S., Klein, E., & Loper, E. (2009). *Natural language processing with Python: analyzing text with the natural language toolkit*. O'Reilly Media, Inc.

[4] Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., Kaiser, Ł., & Polosukhin, I. (2017). Attention is all you need. *Advances in Neural Information Processing Systems*, **30**.

[5] Devlin, J., Chang, M. W., Lee, K., & Toutanova, K. (2019). Bert: Pre-training of deep bidirectional transformers for language understanding. *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, **1**, 4171-4186.

[6] Liu, Y., Ott, M., Goyal, N., Du, J., Joshi, M., Chen, D., Levy, O., Lewis, M., Zettlemoyer, L., & Stoyanov, V. (2019). Roberta: A robustly optimized BERT pretraining approach. *arXiv preprint arXiv:1907.11692*.

[7] Dietterich, T. G. (2000). Ensemble methods in machine learning. *International Workshop on Multiple Classifier Systems*, 1–15.

[8] Hochreiter, S. & Schmidhuber, J. (1997). Long short-term memory. *Neural Computation*, **9**(8), 1735–1780.

[9] Mandl, T., Modha, S., Shahi, G. K., Madhu, H., Satapara, S., Majumder, P., Schäfer, J., Ranasinghe, T., Zampieri, M., Nandini, D., & Jaiswal, A. K. (2021). Overview of the hasoc subtrack at fire 2021: Hate speech and offensive content identification in english and indo-aryan languages. *FIRE 2021 Working Notes*, 2021.

[10] Kabbara, J. K. & Jamil, H. K. (2021). Handling class imbalance in hateful language detection. *Proc. IEEE Int. Conf. on Information Reuse and Integration (IRI)*, 2021.

[11] Collobert, R., Weston, J., Bottou, L., Karlen, M., Kavukcuoglu, K., & Kuksa, P. (2011). Natural language processing (almost) from scratch. *J. Mach. Learn. Res.*, **12**, 2493–2537.

[12] Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., Blondel, M., Prettenhofer, P., Weiss, R., Dubourg, V., Vanderplas, J., Passos, A., Cournapeau, D., Brucher, M., Perrot, M., & Duchesnay, É. (2011). Scikit-learn: Machine learning in Python. *Journal of Machine Learning Research*, **12**, 2825–2830.

[13] Tirta, A. (2022). Deep Learning for Sentiment and Hate Speech Analysis: A Comparative Study of Transformers. *Jurnal Ilmu Komputer dan Informatika*, **12**(3), 110–120.