

A Contextual Transformer Architecture for Proactive Recognition of Cyber Aggression in Social-Media Communication

Roopa A P, Dr. Manjunatha N, Dr. Basavaraj S Pol

Department of Computer Science and Engineering

R L Jalappa Institute of Technology, Doddaballapur, Karnataka, India

roopayadav@gmail.com, Manjunatha22@gmail.com, basavarajspol@rljit.in

Abstract—The continuous growth of social networking services has produced a large volume of user-generated language that cannot be examined manually at scale. Among the safety concerns associated with such communication, cyber aggression is particularly challenging because harmful intent may be stated directly or concealed through context, sarcasm, ridicule, or indirect wording. This paper presents a contextual transformer architecture for assigning social-media messages to three categories: non-aggressive, explicit aggression, and implicit aggression. A pretrained BERT language encoder is adapted to the classification task through supervised fine-tuning. The study uses a combined collection of 50,000 posts and comments obtained from Twitter, Instagram, and public Facebook pages and compares the transformer classifier with Bi-LSTM and SVM reference models. The reported BERT configuration achieves 95.8% accuracy and 94.9% F1-score, compared with 92.1% accuracy and 91.4% F1-score for Bi-LSTM. The results suggest that contextual representations are useful when aggressive meaning cannot be identified reliably from isolated keywords. The proposed architecture can be used as a text-analysis layer for scalable moderation systems and provides a foundation for later real-time and multimodal safety applications.

Index Terms—Cyber aggression, contextual language modelling, BERT, transformer, social-media safety, implicit aggression, text classification

I. INTRODUCTION

Social networking environments have become important channels for discussion, information sharing, entertainment, and interpersonal interaction. At the same time, these environments can facilitate hostile communication, including bullying, intimidation, insulting remarks, harassment, and other aggressive behaviour. The number of messages generated on popular platforms makes continuous human inspection expensive and difficult to sustain. Consequently, computational techniques that can screen large quantities of text are increasingly important for online safety.

A simple keyword-based filter is not sufficient for this task. Clearly offensive messages may contain recognizable abusive vocabulary, but many harmful statements do not. Users can communicate hostility through sarcasm, mockery, coded expressions, contextual references, spelling variations, slang, or combinations of words that appear harmless individually. This creates two complementary risks: harmful content can

pass through a lexical filter, while legitimate criticism can be incorrectly identified as abusive.

Traditional machine-learning approaches generally depend on manually designed linguistic features or frequency-based representations. Although such approaches remain useful as reference systems, they often have limited access to the broader context in which a sentence is written. Neural sequence models improve representation learning by modelling word order and dependencies. Bidirectional LSTM networks, for example, process information in both directions and can preserve useful contextual information. However, recurrent processing does not provide the same direct interaction among all tokens that is available in a transformer architecture.

Transformers use self-attention to determine relationships between tokens within an input sequence. BERT further strengthens this representation by learning bidirectional contextual language features during pretraining. Such representations can subsequently be adapted to classification tasks with a task-specific output layer. This property makes BERT suitable for social-media text where the interpretation of an expression may depend on surrounding words.

The present work develops a three-category cyber-aggression classifier using a pretrained BERT model. Instead of treating all harmful language as one class, the proposed formulation separates explicit aggression from implicit aggression. The distinction is useful because indirect hostility is often more difficult to detect than overt abuse. The study further evaluates the transformer approach against Bi-LSTM and SVM baselines using standard classification measures.

II. RELATED WORK

Automated analysis of hostile and abusive online communication has developed through several generations of computational methods. Early systems commonly represented text using lexical patterns, word counts, character sequences, and n-grams before applying conventional classifiers. Hybrid machine-learning and language-processing strategies have also been explored for cyberbullying detection [1]. These approaches are useful for establishing baseline performance, but feature engineering can become less effective when users

introduce new slang, spelling variations, or context-dependent expressions.

Deep recurrent architectures provided an alternative by learning representations directly from sequential text. LSTM-based approaches have been investigated for cyberbullying and harmful-language classification because recurrent memory can retain information from earlier parts of an input sequence [2]. Bidirectional processing further permits information from both directions to influence the final representation. Nevertheless, recurrent architectures process a sequence step by step, whereas transformers can establish attention-based relationships among many token positions simultaneously.

The introduction of BERT demonstrated the effectiveness of pretrained bidirectional transformer representations for downstream language understanding tasks [3]. The contextual nature of these representations is particularly relevant to offensive-language classification, where the same word may have different meanings depending on its surrounding text. Research on offensive posts has also investigated the distinction between different types and targets of harmful social-media content [4].

Some recent directions move beyond text-only analysis. Multimodal datasets have shown that harmful meaning can arise from a combination of written language and visual information [5]. Other research has investigated large-scale personal-attack detection and automatic comment moderation [6], [11]. Surveys of hate-speech detection further demonstrate the wide range of linguistic and machine-learning techniques used for harmful-content analysis [7], [8].

Deep learning has also been applied to social-media hate-speech detection [9], while cross-platform cyberbullying detection has examined the challenge of transferring learned patterns across different social environments [10]. BERT-based approaches have additionally been studied for harmful-language detection and bias-related issues [12]. Transformer alternatives such as RoBERTa and sentence-level BERT representations have expanded the range of contextual language modelling techniques [13], [14]. The present study focuses specifically on a three-class distinction between non-aggressive, explicit, and implicit aggression.

III. OBJECTIVES

The study is designed around the following objectives:

- 1) Develop an automated text-analysis pipeline for screening cyber aggression in social-media communication.
- 2) Fine-tune a pretrained BERT encoder for the target three-class prediction task.
- 3) Differentiate ordinary communication from overt and indirect aggressive language.
- 4) Improve the identification of aggression whose meaning is distributed across the surrounding sentence rather than expressed through a single offensive word.
- 5) Establish Bi-LSTM and SVM as comparison models for evaluating the transformer-based approach.
- 6) Measure model behaviour using accuracy, precision, recall, and F1-score.

- 7) Develop an architecture that can act as a foundation for proactive and scalable moderation.
- 8) Identify extension paths involving multilingual, multimodal, conversational, and real-time analysis.

IV. PROBLEM STATEMENT

The large and continuously changing stream of messages produced by social-media users creates a significant challenge for automated cyber-aggression monitoring. Harmful communication may take the form of direct insults, threats, profanity, bullying, or less obvious interpersonal attacks. A moderation mechanism that depends entirely on human reviewers cannot efficiently inspect every message at the scale of modern platforms.

The difficulty becomes greater when aggressive intent is not stated explicitly. Sarcastic comments, indirect insults, ridicule, coded phrases, local slang, deliberate spelling changes, and context-dependent statements may convey hostility without containing a conventional offensive keyword. Conversely, an automated filter may classify an emotionally strong but legitimate statement as aggressive when it does not understand the intended context. Therefore, accurate detection requires more than matching words against a predefined list.

Social-media writing also differs substantially from formal text. Users frequently employ abbreviations, emojis, repeated characters, unconventional grammar, mixed writing styles, and newly emerging expressions. These characteristics create a continually changing linguistic environment in which static rules can lose effectiveness. A robust classifier should consequently capture relationships between words and interpret an expression using its surrounding information.

The central problem addressed by this study is the design of a scalable contextual text-classification framework capable of distinguishing non-aggressive communication, explicit aggression, and implicit aggression. The proposed framework uses a fine-tuned BERT encoder to obtain contextual representations and a classification layer to map those representations to the three target categories. The intended outcome is a stronger basis for early screening of potentially harmful communication and subsequent human moderation.

V. PROPOSED METHODOLOGY

The proposed methodology converts raw social-media messages into class predictions through a sequence of data preparation, token representation, transformer encoding, supervised adaptation, and evaluation stages. The processing strategy intentionally avoids excessive text normalization so that contextual information, informal wording, and meaningful variations are not unnecessarily removed.

A. Data Acquisition and Annotation

The experimental collection contains 50,000 posts and comments gathered from Twitter, Instagram, and public Facebook pages. The messages are organized into three categories according to an annotation guideline prepared with input from a social psychologist. The categories are defined as follows:

- **Class 0 – Non-Aggressive:** communication that does not express an identifiable aggressive intent and may be neutral or positive.
- **Class 1 – Explicit Aggression:** overtly hostile communication such as direct insults, threats, profanity, or clearly abusive statements.
- **Class 2 – Implicit Aggression:** hostile meaning expressed indirectly through ridicule, sarcasm, indirect insults, or subtle demeaning language.

This three-way organization preserves information that would be lost in a simple binary aggressive/non-aggressive formulation.

B. Text Preparation and Token Representation

Only limited normalization is applied before transformer processing. Unnecessary whitespace and unsuitable character sequences are removed while the original linguistic content is retained wherever possible. This approach is important because unusual spelling, slang, and informal construction can sometimes contribute to the interpretation of an aggressive message.

The BERT tokenizer subsequently divides each message into WordPiece subword units. Special sequence markers are inserted according to the BERT input format, and an attention mask is created to distinguish valid tokens from padding positions. The token identifiers and corresponding masks are then passed to the transformer encoder.

C. Transformer Architecture

The proposed classifier uses the pretrained `bert-base-uncased` model. The processing path begins with a social-media message and proceeds through subword tokenization, contextual transformer encoding, extraction of the classification representation, a fully connected layer, and a three-class probability output.

1) *Self-Attention:* The principal operation used by the transformer is self-attention. For one attention head, the transformation is represented by

$$\text{Attention}(Q, K, V) = \text{softmax}\left(\frac{QK^T}{\sqrt{d_k}}\right)V, \quad (1)$$

where Q , K , and V denote the query, key, and value matrices, respectively, and d_k is the dimension associated with the key representation.

Several attention heads are combined to form the multi-head representation:

$$\text{MultiHead}(Q, K, V) = \text{Concat}(\text{head}_1, \dots, \text{head}_h)W^O. \quad (2)$$

For the individual heads,

$$\text{head}_i = \text{Attention}(QW_i^Q, KW_i^K, VW_i^V). \quad (3)$$

The attention mechanism permits the encoder to assign different importance to tokens according to their relationships within the input. This property is useful for cyber-aggression

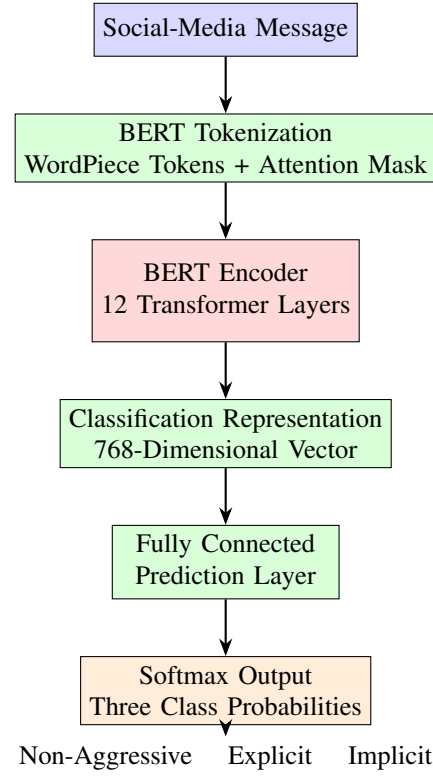


Fig. 1: Processing pipeline of the proposed contextual BERT classifier.

analysis because the interpretation of one phrase may depend strongly on another phrase appearing elsewhere in the same message.

2) *Supervised Adaptation:* The pretrained encoder is adapted to the target classes by minimizing the categorical cross-entropy objective:

$$\mathcal{L} = -\frac{1}{N} \sum_{i=1}^N \sum_{c=1}^C y_{i,c} \log(\hat{y}_{i,c}), \quad (4)$$

where N denotes the number of examples considered in the batch, $C = 3$ represents the number of output categories, $y_{i,c}$ is the target indicator, and $\hat{y}_{i,c}$ is the probability predicted for class c .

The reported configuration uses four training epochs, a batch size of 16, AdamW optimization, a learning rate of 2×10^{-5} , and a weight-decay coefficient of 0.01.

D. Bi-LSTM Comparison Model

A bidirectional LSTM is employed as a neural comparison model. Input text is represented using 100-dimensional GloVe embeddings. The embedding sequence is processed by forward and backward recurrent units containing 128 units in each direction. Their representations are combined, regularized using dropout with a rate of 0.5, and supplied to a dense layer containing three output units with softmax activation.

The recurrent computations are represented by

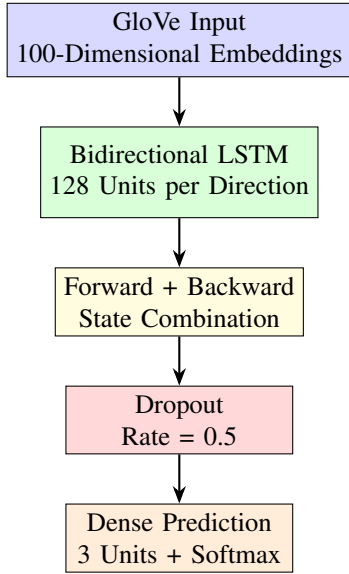


Fig. 2: Structure of the Bi-LSTM comparison model.

$$f_t = \sigma(W_f[h_{t-1}, x_t] + b_f), \quad (5)$$

$$i_t = \sigma(W_i[h_{t-1}, x_t] + b_i), \quad (6)$$

$$\tilde{C}_t = \tanh(W_C[h_{t-1}, x_t] + b_C), \quad (7)$$

$$C_t = f_t * C_{t-1} + i_t * \tilde{C}_t, \quad (8)$$

$$o_t = \sigma(W_o[h_{t-1}, x_t] + b_o), \quad (9)$$

$$h_t = o_t * \tanh(C_t). \quad (10)$$

These operations allow the recurrent baseline to retain and update information while processing the sequence in both directions.

VI. EXPERIMENTAL RESULTS AND ANALYSIS

A. Experimental Configuration

The reported experiments divide the dataset into training and testing portions using an 80:20 stratified split. Stratification is applied to preserve the relative representation of the three target categories. The experiments are reported as being executed on an NVIDIA RTX 3080 GPU with 10 GB memory using PyTorch 1.12 together with the Hugging Face Transformers library.

B. Performance Measures

Model performance is examined using accuracy, precision, recall, and F1-score. Accuracy measures the proportion of correctly classified examples:

$$\text{Accuracy} = \frac{TP + TN}{TP + TN + FP + FN}. \quad (11)$$

Precision and recall are calculated as

$$\text{Precision} = \frac{TP}{TP + FP}, \quad \text{Recall} = \frac{TP}{TP + FN}. \quad (12)$$

The F1-score combines precision and recall through their harmonic mean:

$$F1 = 2 \frac{\text{Precision} \times \text{Recall}}{\text{Precision} + \text{Recall}}. \quad (13)$$

For the three-class experiment, these measures can be aggregated according to the selected multiclass evaluation strategy.

C. Comparative Results

The reported values for the three evaluated approaches are listed in Table I.

TABLE I: Reported performance of the evaluated classifiers

Model	Accuracy	Precision	Recall	F1
BERT	95.8%	95.5%	94.9%	94.9%
Bi-LSTM	92.1%	91.8%	91.0%	91.4%
SVM	85.3%	84.1%	83.7%	83.9%

Based on the reported measurements, the BERT configuration produces the highest overall scores among the three models. Its accuracy exceeds that of Bi-LSTM by 3.7 percentage points, and the corresponding F1-score is 3.5 percentage points higher. The SVM baseline records the lowest values in the comparison. These results are consistent with the motivation for using contextual representations: the interpretation of an aggressive statement may depend on relationships extending beyond neighbouring words.

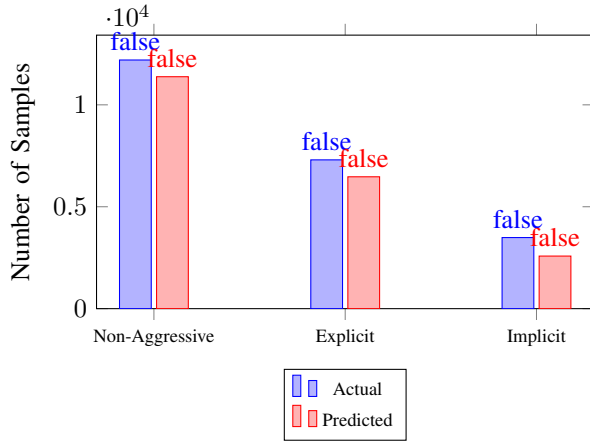
The reported analysis further highlights the implicit-aggression category. Compared with BERT, the Bi-LSTM system is described as producing greater confusion between implicit and non-aggressive examples. The stronger separation obtained with BERT suggests that attention-based contextual representations can provide useful information when hostile intent is distributed across a sentence rather than exposed through an individual offensive token.

VII. SCOPE OF THE WORK

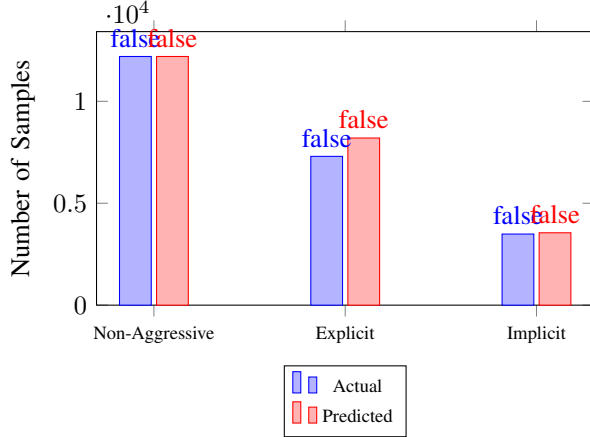
The current framework concentrates on the classification of textual communication collected from social-media environments. Its primary purpose is to provide a contextual language-analysis component that can identify whether a message is non-aggressive, explicitly aggressive, or implicitly aggressive.

The current scope includes:

- processing of textual material obtained from multiple social-media services;
- three-category recognition of non-aggressive, explicit, and implicit aggression;
- contextual feature extraction using a pretrained BERT encoder;
- comparative evaluation against Bi-LSTM and SVM models;
- assessment using accuracy, precision, recall, and F1-score; and
- preparation of a model architecture that can later be incorporated into moderation pipelines.



(a) BERT



(b) Bi-LSTM

Fig. 3: Actual and predicted category distributions used for the reported comparison. Numerical values are intentionally not printed above the bars to avoid label overlap.

The present implementation does not directly process photographs, audio, video, social-network graphs, or detailed behavioural histories. These sources can be introduced in future versions when additional data and processing modules are available. Other possible extensions include multilingual and code-mixed text, conversation-level reasoning, explanation generation, and real-time deployment.

VIII. NOVELTY AND KEY CONTRIBUTIONS

The distinctive aspect of the proposed work is the formulation of cyber-aggression recognition as a three-category contextual language problem. Instead of assigning every harmful message to a single broad class, the framework explicitly separates overt aggression from less visible forms of hostile communication.

The principal contributions are:

- **Context-sensitive text modelling:** A pretrained BERT encoder is fine-tuned so that the classifier can use relationships among tokens when assigning an aggression category.

- **Three-level aggression representation:** The framework distinguishes ordinary communication, explicit aggression, and implicit aggression.
- **Indirect-aggression focus:** Sarcasm, ridicule, subtle insults, and demeaning expressions are treated as a separate analytical category.
- **Transfer-learning strategy:** A pretrained language representation is adapted to the target application rather than learning all language features from an initially untrained network.
- **Cross-model comparison:** The proposed architecture is evaluated against both a recurrent neural model and a conventional machine-learning classifier.
- **Multi-source evaluation:** The experimental collection combines social-media material from Twitter, Instagram, and public Facebook pages.
- **Moderation-oriented architecture:** The resulting classifier can operate as a text-screening component before messages are forwarded to a larger moderation or human-review system.

IX. LIMITATIONS

Several limitations should be considered when interpreting the reported findings.

- 1) **Dataset dependence:** Predictions depend on the characteristics of the training collection, including its balance, diversity, annotation quality, and time of collection. Language patterns that appear later may therefore be underrepresented.
- 2) **Indirect meaning remains difficult:** Sarcasm, humour, coded expressions, and culturally specific references can remain ambiguous even when contextual transformer representations are used.
- 3) **Platform-specific language:** Different platforms encourage different communication styles. Performance obtained from one combination of platforms may not transfer unchanged to another environment.
- 4) **Restricted language coverage:** The reported implementation is centred on English. Regional languages, multilingual messages, and code-mixed communication require additional training data and appropriate language models.
- 5) **Computational demand:** Transformer fine-tuning and inference generally require more memory and computation than simpler classifiers, which can affect deployment on low-resource devices.
- 6) **Message-level analysis:** The primary unit of analysis is an individual text sample. Conversation history, relationships among users, posting frequency, and network-level evidence are not directly included.
- 7) **False-positive and false-negative cases:** Strong criticism can be mistaken for aggression, while carefully disguised hostile messages may escape detection.
- 8) **Production deployment:** The presented work evaluates a classification architecture rather than a complete in-

dustrial moderation service with monitoring, alerting, human escalation, and operational controls.

- 9) **Privacy and fairness concerns:** Practical deployment must account for data protection, annotation bias, model bias, responsible data handling, and the potential impact on legitimate expression.
- 10) **Limited interpretability:** High predictive performance does not automatically provide a simple explanation of why the model selected a particular aggression category.

X. CONCLUSION AND FUTURE WORK

This paper presented a contextual transformer architecture for identifying cyber aggression in social-media text. The proposed formulation divides messages into non-aggressive, explicit-aggression, and implicit-aggression categories and employs a fine-tuned BERT encoder to capture contextual relationships. According to the reported experimental values, the BERT model obtains the strongest performance among the evaluated BERT, Bi-LSTM, and SVM approaches.

An important aspect of the framework is the separate treatment of implicit aggression. Hostile intent is not always represented by an obvious offensive word; it can emerge from the relationship between otherwise ordinary terms, sentence structure, or contextual cues. The transformer-based representation is therefore appropriate for investigating such cases. Nevertheless, practical deployment must address limitations related to language diversity, evolving slang, platform differences, computation, privacy, fairness, and model interpretation.

Future development can extend the framework toward multilingual and code-mixed communication, conversational context, and multimodal content. Images, audio, and video could be incorporated when textual evidence alone is insufficient. User-level and conversation-level signals could provide additional context while requiring careful privacy controls. A future implementation could also provide confidence scores and human-readable explanations so that automated predictions support rather than replace moderator decisions. Finally, the trained model could be integrated into a scalable real-time service for early screening and prioritization of potentially harmful content.

ACKNOWLEDGMENT

The authors acknowledge the annotators involved in the preparation of the labelled data and the institutional computational facilities used for the reported experiments.

REFERENCES

- [1] C. Raj, A. Agarwal, G. Bharathy, B. Narayan, and M. Prasad, "Cyberbullying detection: Hybrid models based on machine learning and natural language processing techniques," *Electronics*, vol. 10, no. 22, p. 2810, 2021.
- [2] M. H. U. Rahman, M. Divya, B. R. Reddy, K. S. Kumar, and P. R. Vani, "Cyberbullying detection using LSTM networks on social media text," *International Journal of Research in Applied Science and Engineering Technology*, vol. 10, 2022.
- [3] J. Devlin, M.-W. Chang, K. Lee, and K. Toutanova, "BERT: Pre-training of deep bidirectional transformers for language understanding," in *Proc. NAACL-HLT*, 2019, pp. 4171–4186.
- [4] M. Zampieri, S. Malmasi, P. Nakov, S. Rosenthal, N. Farra, and R. Kumar, "Predicting the type and target of offensive posts in social media," in *Proc. NAACL-HLT*, 2019, pp. 1415–1420.
- [5] D. Kiela, H. Firooz, A. Mohan, V. Goswami, A. Singh, P. Ringshia, and D. Testuggine, "The hateful memes challenge: Detecting hate speech in multimodal memes," in *Advances in Neural Information Processing Systems*, vol. 33, 2020.
- [6] E. Wulczyn, N. Thain, and L. Dixon, "Ex Machina: Personal attacks seen at scale," in *Proc. 26th International Conference on World Wide Web*, 2017, pp. 1391–1400.
- [7] A. Schmidt and M. Wiegand, "A survey on hate speech detection using natural language processing," in *Proc. 5th International Workshop on Natural Language Processing for Social Media*, 2017, pp. 1–10.
- [8] P. Fortuna and S. Nunes, "A survey on automatic detection of hate speech in text," *ACM Computing Surveys*, vol. 51, no. 4, pp. 1–30, 2018.
- [9] P. Badjatiya, M. Gupta, M. Gupta, and V. Varma, "Deep learning for hate speech detection in tweets," in *Proc. 26th International Conference on World Wide Web Companion*, 2017, pp. 759–760.
- [10] S. Agrawal and A. Awekar, "Deep learning for detecting cyberbullying across multiple social media platforms," in *Advances in Information Retrieval*, Springer, 2018, pp. 141–153.
- [11] J. Pavlopoulos, P. Malakasiotis, and I. Androutsopoulos, "Deep learning for user comment moderation," in *Proc. 1st Workshop on Abusive Language Online*, 2017, pp. 25–35.
- [12] N. Mozafari, R. Farahbakhsh, and N. Crespi, "Hate speech detection and racial bias mitigation in social media based on BERT," *PLoS ONE*, vol. 15, no. 8, 2020.
- [13] Y. Liu *et al.*, "RoBERTa: A robustly optimized BERT pretraining approach," arXiv preprint arXiv:1907.11692, 2019.
- [14] N. Reimers and I. Gurevych, "Sentence-BERT: Sentence embeddings using Siamese BERT-networks," in *Proc. EMNLP-IJCNLP*, 2019, pp. 3982–3992.
- [15] A. Vaswani *et al.*, "Attention is all you need," in *Advances in Neural Information Processing Systems*, vol. 30, 2017.