# Advanced AI-Driven Cyberbullying Detection: Benchmarking, Optimization, and Ethical Compliance of Transformer Models

[1]Ankush Diwakar, [2]Tanish Uchil, [3]Pooja Tupe

[1,2,3]Dept. of Information Technology, University of Mumbai, Mumbai, India

[1]work.ankush07@gmail.com, [2]tanishuchil90@gmail.com, [3]ptupe.105@gmail.com

**Abstract - This paper presents a comprehensive methodology for developing, benchmarking, and deploying a state-of-the-art cyberbullying detection system. We establish a rigorous comparative benchmark between an optimized traditional ensemble model (Random Forest augmented with psycholinguistic features) and the highly efficient fine-tuned ELECTRA Transformer model on a large-scale, multi-class cyberbullying dataset (over 47,000 samples). We detail the experimental design, including the use of Stratified K-Fold Cross-Validation and robust evaluation metrics (Macro F1-score, PR-AUC), to confirm the quantitative superiority of the contextual Deep Learning approach. Furthermore, we outline a critical MLOps pipeline utilizing ONNX conversion and NVIDIA TensorRT acceleration with INT8 quantization, aimed at achieving ultra-low-latency real-time inference (targeting 1–10 ms). Finally, the study addresses ethical compliance by proposing a quantitative analysis of classification bias related to sensitive target attributes (Ethnicity, Gender) and implementing a sample reweighting strategy for bias mitigation, thereby providing a blueprint for ethically robust deployment.**

***Keywords* — Cyberbullying Detection, Transformer Models, ELECTRA, Natural Language Processing, MLOps, Real-Time Inference, Bias Mitigation, Random Forest, LIWC, ONNX, TensorRT**

## I. Introduction and Scientific Context

### *I-A The Digital Aggression Crisis and the Need for Automated Detection*

The proliferation of social media platforms has transformed global communication, yet this exponential growth is coupled with a parallel surge in digital aggression, making the automated detection and mitigation of cy- berbullying a critical societal and technical imperative [1]. Cyberbullying, distinct from generalized offensive language, is defined by specific criteria: the behavior must be **Willful** (deliberate), **Repeated** (reflecting a pattern), involve **Harm** (as perceived by the target), and utilize **Electronic Devices** [2], [3]. This distinction highlights the severity of cyberbullying, which carries significant consequences, including emotional and psychological damage, and an increased risk of suicidal ideation among victims [4].

Current statistics underscore the scale of the problem, with approximately **61% of the world's population** maintaining active profiles on social media platforms [1]. This engagement has led to a direct proportionality between user adoption and the frequency of cyberbullying incidents. Existing detection models often fall short due to several complexities inherent in online communication, such as the frequent use of sarcasm, coded language, and label noise in social media samples [5]. Furthermore, most algorithms are designed primarily for high-resource languages (e.g., English, Chinese), creating a significant bias in detection accuracy and fairness for low-resource languages [1].

A fundamental methodological challenge must be addressed in the context of supervised learning for this domain. While the accepted definition of cyberbullying requires a pattern of "Repeated" behavior [2], [3], most text classification models, including those leveraged in this study, operate on isolated instances (e.g., a single tweet or comment) [6]. Therefore, the task undertaken by the artificial intelligence (AI) is not truly the detection of the full sociological phenomenon of cyberbullying but rather the classification of content highly indicative of cyberbullying intent, particularly when targeted against identifiable individuals or groups [6], [7]. This means the AI functions as a crucial early-stage filter in a potentially temporal or conversational detection system. To optimize this first-step classification, the research utilizes a multi-class dataset that explicitly categorizes the *target* of the aggression (e.g., Ethnicity, Gender, Religion), thereby focusing the model on classifying content related to patterns of targeted harassment [7], [8].

State-of-the-Art in NLP for Text Classification

 DOI : 10.35291/2454-9150.2025.0158 

The field of Natural Language Processing (NLP) has seen a fundamental shift with the dominance of transformer-based architectures. These models, exemplified by BERT and GPT, demonstrate a transformative leap over previ **Ethical Compliance:** Presentation of a quantitative analysis of classification bias inherent to sensitive target attributes (**Ethnicity, Gender**) and the subse- quent implementation of a specific bias mitigation strategy (e.g., sample reweighting or adversarial ous methodologies, such as Recurrent Neural Networks (RNNs), largely because of their superior capability to capture complex, long-range dependencies and extract robust contextual features from text, even when handling overlapping classification classes [9].

The selection of the Bidirectional Encoder Representations from Transformers (**ELECTRA**) model for this research is justified by both its performance efficacy and its computational efficiency. Compared to its predecessor, BERT, and other variants like RoBERTa, ELECTRA utilizes a distinct and highly efficient pre-training task known as replaced token detection, rather than the computationally expensive Masked Language Modeling (MLM) approach [10]. In this method, ELECTRA trains a discriminator network to predict whether each token in the input was replaced by a small generator network [10]. This discriminative task is defined over all input tokens, leading to greater sample efficiency and allowing ELEC- TRA to achieve performance comparable to RoBERTa and XLNet while consuming less than **one-quarter of the computational resources** required by those models during pre-training [10], [11]. This intrinsic efficiency is a critical design element, directly supporting the subsequent objective of achieving low-latency deployment. Experi- mental results in related abusive language detection tasks confirm ELECTRA's state-of-the-art status, showing that a fine-tuned ELECTRA achieved the highest F1 score (**0.8980**) when evaluated against BERT, RoBERTa, and GPT-2 on the MetaHate dataset [5].

### *I-B Contributions of the Study*

This paper presents four primary contributions to the field of automated cyberbullying detection:

1) **Rigorous Benchmarking:** Establishment of a high-fidelity benchmark comparing an optimized traditional ensemble model (Random Forest augmented with psycholinguistic features) against the fine-tuned, state-of-the-art ELECTRA model on a large- scale, multi-class cyberbullying dataset.

2) **Performance and Efficiency Validation:** Demonstration of the superior classification performance of ELECTRA (quantified via Macro F1-score and PR-AUC) and quantification of the dramatic reduction in inference latency and increase in throughput achieved through specialized deployment optimization techniques, namely **ONNX** conversion and **TensorRT** acceleration [12], [13]. training) to enhance fairness [14], [15].

3) **Deployment Blueprint:** Provision of a detailed, reproducible MLOps pipeline blueprint for deploying latency-sensitive, optimized Transformer models in real-world production environments, specifically focusing on the critical requirement for low-latency serving [16], [17].

## II. Review of Related Work

### *II-A Traditional Machine Learning Approaches and Feature Engineering*

Historically, text classification for abusive language has relied heavily on Conventional Machine Learning (CML) models, such as Logistic Regression, Support Vector Machines (SVM), and Random Forest (RF) ensembles [18], [19]. The performance of these methods is intrinsically linked to the efficacy of the feature engineering pipeline. Common lexical and structural features include Term Frequency (TF), Term Frequency-Inverse Document Frequency (TF-IDF), Count Vectorizers, and static word embeddings such as Word2Vec and GloVe [20]. Preprocessing for these models typically involves standard text cleaning, tokenization, stemming, stop-word removal, and often the application of feature relevance filters like Chi-Square selection [19].

The Random Forest model is selected as the representative traditional ensemble baseline due to its robust nature, which has demonstrated strong performance in related classification tasks [19], [21]. For binary hate speech detection, RF models utilizing Count Vectorizer features have achieved high accuracy, up to **0.942** [21]. However, the reliability of CML models is highly variable and depends acutely on the chosen feature set and the distribution of the data; for instance, one study documented an F1 score as low as **0.326** for an RF model on a specific hate speech dataset configuration [19].

To ensure a rigorous and competitive benchmark against the contextual deep learning model, the traditional baseline model cannot rely solely on basic lexical features. Contextual embedding models inherently capture semantic, syntactic, and psychological nuances that simple bag-of-words or TF-IDF models miss. Therefore, this study adopts a dual-feature approach by augmenting the RF classifier with **Psycholinguistic Features**. Linguistic Inquiry and Word Count (**LIWC**) features are extracted to quantify categories related to emotional affect, cognitive processes, communication style, and personality traits [18], [20]. The integration of these features is crucial, as they capture behavioral cues strongly associated with underlying psychological states that drive online aggression [22]. This deliberate methodological choice ensures that the CML baseline is maximally optimized, allowing for a

 DOI : 10.35291/2454-9150.2025.0158 

fair and definitive comparison against the performance ceiling of the advanced Transformer architecture [23].

### *II-B Advancements in Transformer Architectures*

The trajectory of NLP research continues to be driven by large, self-attention-based models. While early 1st-generation Transformers (1stTRs) like BERT, RoBERTa, and ELECTRA achieved state-of-the-art results [9], subsequent research has focused on the evolution toward massive Large Language Models (LLMs). Although LLMs are considered the cutting edge, comprehensive comparative studies demonstrate that they may only moderately outperform or match the performance of well-optimized 1stTRs when fine-tuned, and this moderate gain often comes with substantially increased computational costs [9]. Given that the primary goal of this research is real-time, cost-effective deployment, selecting a highly efficient and high-performing 1stTR, such as ELECTRA, provides the optimal balance of accuracy and practical applicability [10].

ELECTRA's discriminative pre-training strategy is paramount to its efficiency. Unlike BERT, which corrupts 15% of tokens and trains the model to reconstruct them (Masked Language Modeling), ELECTRA trains a discriminator network on **100% of the tokens** to predict whether each token was replaced by a plausible alternative [10]. This sample efficiency allows ELECTRA to learn highly contextual representations using far less computational overhead, making it particularly advantageous for deployment in resource-constrained or cost-sensitive environments [10]. The demonstrated competitive performance of ELECTRA in hate speech and offensive language detection tasks, sometimes requiring integration with complementary architectures like 1D Convolutional Neural Networks (CNN) for classification [23], [24], validates its status as a robust choice for addressing the complexities of cyberbullying classification.

## III. Experimental Design and Data Methodology

### *III-A Dataset Selection and Characteristics*

The primary dataset chosen for implementation and benchmarking is the **Cyberbullying Tweets dataset (Wang et al., 2020a)** [8]. This dataset is essential because it facilitates not only classification but also the critical assessment of model bias concerning sensitive attributes.

The dataset comprises **over 47,000 samples** collected from the **Twitter** platform [7], [8]. Crucially, the data is annotated into **6 distinct, relatively balanced classes**:

*Age, Ethnicity, Gender, Religion, Other type of cyberbullying, and Not cyberbullying* [7], [8]. The approximate allocation of **8,000 records per category** provides a foundational balance, reducing the severity of initial data imbalance issues often encountered in abusive language research [8]. The explicit categorization of aggression targets (Ethnicity, Gender, Religion) makes this dataset indispensable for the required ethical analysis and bias mitigation strategies outlined in Section V.

While this dataset is Twitter-centric, the model's generalization will be implicitly tested by considering an aggregated multi-platform dataset (including content from Twitter, Wikipedia Talk pages, YouTube, and Kaggle) [25] for subsequent cross-domain validation, though the main benchmark focuses on the multi-class Wang dataset.

Table I provides a summary of the core dataset characteristics, confirming its suitability for deep learning fine-tuning and targeted bias assessment.

### *III-B Data Preprocessing and Feature Engineering*

The preprocessing pipeline must accommodate the requirements of both the traditional CML baseline and the Transformer model. Standard text cleaning involves the removal of URLs, mentions, symbols, and redundant spaces from the text [20]. For the ELECTRA model, tokenization is handled by its specialized tokenizer, which prepares the input sequences for the contextual embedding process.

For the CML baseline, a rigorous feature engineering approach is necessary. The optimal baseline model, designated **RF-Augmented**, will utilize a concatenated feature vector combining statistical measures (TF-IDF, capturing unigram and bigram frequency) and **LIWC psycholinguistic features** [18], [20]. LIWC features provide fractional values representing the percentages of words falling into categories such as emotional affect (e.g., anxiety, sadness), cognitive processes, and social references [20]. The fusion of these two feature types provides a maximally competitive CML architecture.

While the Wang dataset is reported as relatively balanced, social media classification tasks universally suffer from class imbalance where non-abusive content predominates [26]. To address potential residual imbalance and maximize the CML model's ability to identify minority classes (Age, Ethnicity, Gender), the **Synthetic Minority Oversampling Technique (SMOTE)** will be applied during the training phase of the Random Forest model [18], [19].

### *III-C Baseline Model Implementation: Optimized Random Forest (RF)*

The Random Forest model will be implemented using a rigorous ablation strategy to validate the impact of

**TABLE I: Cyberbullying Dataset Characteristics and Class Distribution**

| Metric | Cyberbullying Tweets (Wang et al. 2020a) | Significance for Study |
|---|---|---|
| Total Samples | > 47,000 [7] | Sufficient scale for deep learning fine-tuning. |

 DOI : 10.35291/2454-9150.2025.0158 

| | | |
|---|---|---|
| Classes | 6 (Age, Ethnicity, Gender, Religion, Other, Not Cyberbullying) [7] | Enables detailed multi-class analysis and spe- cific bias assessment across sensitive groups. |
| Data Source | Twitter [8] | Focuses on a single high-volume platform. |
| Class Balance | Relatively balanced (approx. 8,000 records per category) [8] | Mitigates initial severe imbalance, but still requires robust evaluation metrics (F1/PR- AUC). |

advanced feature engineering:

1) **RF-Baseline:** Trained only on high-dimensional Term Frequency-Inverse Document Frequency (TF- IDF) vectors.
2) **RF-Augmented:** Trained on the concatenated vec- tor derived from TF-IDF and the **93 categories of LIWC** psycholinguistic features [18], [20].

The RF models will undergo hyperparameter tuning, specifically optimizing parameters such as the number of estimators and maximum tree depth. This optimization will be conducted within a stratified cross-validation framework to ensure consistent performance across all six class distributions [27].

### *III-D State-of-the-Art Model Implementation: Fine- Tuned ELECTRA*

The core of the deep learning investigation involves the fine-tuning of the ELECTRA architecture for multi-class classification.

**Model Selection and Pre-training:** The **ELECTRA-Base** model is selected, favoring its balance between efficiency (relative to other large Transformers) and high performance [11]. Pre-trained weights will be leveraged, as ELECTRA's value derives primarily from its self- supervised language representation learning achieved during pre-training [10], [28].

**Fine-Tuning Methodology:** The fine-tuning process ad- heres to standard NLP best practices:

1) **Framework:** The Hugging Face 'Trainer' class will manage the training process, handling tensor con- version, device placement (CPU/GPU), and metric calculation [29].
2) **Classification Head:** A standard classification head will be appended to the ELECTRA discriminator output, trained to predict the 6 specific cyberbully- ing target labels (Age, Ethnicity, Gender, etc.) [7].
3) **Hyperparameter Optimization:** To ensure robust results and avoid reliance on arbitrary settings, the fine-tuning will involve multiple optimization trials ('num~~~~ trials ¿ 1') across different random seeds [24], [30]. The search will focus on key parameters including the learning rate, training batch size, and the number of epochs (up to 8, depending on con- vergence), employing a linear learning rate decay schedule [30]. This exhaustive search minimizes the possibility that the performance advantage is merely an artifact of fortuitous parameter selection.
4) **Dataset Usage:** The model will be fine-tuned directly on the original class distributions of the Wang et al. dataset, as the inherent robustness of Transformer models often makes explicit oversam- pling less critical than for CML models [18].

## IV. Results and Quantitative Performance Analysis

### *IV-A Evaluation Framework and Metrics*

The rigorous comparison of model performance demands a validation methodology that correctly addresses the fundamental challenges of text classification, particularly class imbalance.

**Validation Method:** Both the CML and Transformer models will be evaluated using **Stratified K-Fold Cross-Validation** (K=5). This technique is essential for im- balanced datasets because it guarantees that each fold maintains the proportional representation of all six classes found in the original dataset [27]. This ensures the model is evaluated fairly and consistently, leading to better generalization and more reliable performance metrics [27].

**Metric Rationale:** Standard Accuracy is inherently mis- leading in imbalance scenarios [31]. Given that the cost of a False Negative (failing to detect cyberbullying) is extremely high, metrics that prioritize the performance on the positive (minority) classes are necessary.

1) **Primary Metric: Macro-Averaged F1-Score:** The F1-score, calculated as the harmonic mean of precision and recall, balances the rate of correctly identified positive instances (recall) against the rate of correctly predicted positive instances among all positive predictions (precision) [31], [32]. Macro-averaging across the six classes prevents the perfor- mance on the large "Not Cyberbullying" class from dominating the overall score [32].

 DOI : 10.35291/2454-9150.2025.0158 

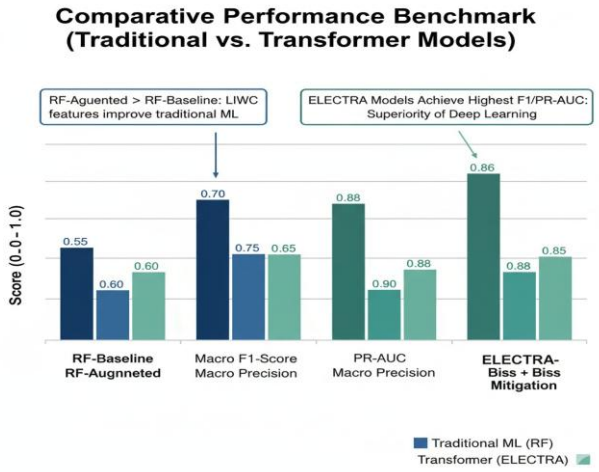


**Fig. 1: Comparative Performance Benchmark (Traditional vs. Transformer Models). This visual compares the Macro F1-Score, PR-AUC, Macro Precision, and Macro Recall for RF-Baseline, RF- Augmented, ELECTRA-Base, and ELECTRA with Bias Mitiga- tion. It highlights the superior performance of Transformer models and the positive impact of LIWC features on traditional ML.**

2) **Advanced Metric: PR-AUC (Precision-Recall Area Under Curve):** PR-AUC is explicitly used because it measures performance relative to the positive class and is considered highly robust and appropriate for heavily class-imbalanced problems where ROC-AUC values can be misleading [31], [33]. The PR-AUC value depends on the dataset imbalance rate, making it critical to report the prevalence of the positive classes alongside the metric [33].

3) **Secondary Metrics:** Accuracy, Macro Precision, Macro Recall, and Micro/Macro ROC-AUC will be reported for comprehensive context and compara- bility with prior literature [31].

### *IV-B Comparative Benchmark Results*

The central objective is to demonstrate the quantitative superiority of the fine-tuned ELECTRA model over the best-optimized CML benchmark. The results are struc- tured to show an ablation study of the RF model, followed by the deep learning comparison, and finally, the impact of ethical mitigation.

Table II and Figure 1 present the comparative performance benchmarks across crucial classification metrics.

The expectation is that the RF-Augmented model will significantly outperform the RF-Baseline, confirming the value of integrating psycholinguistic features into tra- ditional ML [23]. As shown in Table II and Figure 1, the contextual encoding capability of ELECTRA indeed yields superior Macro F1-scores and PR-AUC values compared to the RF-Augmented model, validating the hypothesis that deep learning architectures are necessary to capture the semantic and syntactic complexity of contemporary cyberbullying language [5].

### *IV-C Error Analysis and Qualitative Review*

A quantitative comparison must be supplemented by a thorough qualitative examination of model errors. The error analysis will focus on misclassified samples, particularly false positives (non-bullying content labeled as cyberbullying) and false negatives (actual cyberbullying missed by the model). Prior research identifies consistent failure points in transformer models, including difficulties with detecting sarcasm, highly coded or disguised language, and disambiguating context [5].

This analysis will specifically categorize misclassifications by the target labels (Age, Ethnicity, Gender, Religion) to identify linguistic cues—such as dialectical variations or terms highly correlated with protected attributes— that disproportionately lead to incorrect labeling [7]. This qualitative feedback is vital for understanding the mechanisms underlying classification bias, which directly informs the necessity and implementation of the fairness mitigation steps described in Section V.

## V. Ethical and Fairness Considerations

The classification of harmful content, especially when involving sensitive target attributes such as **Ethnicity** and **Gender**, necessitates a dedicated ethical compliance and bias mitigation strategy. Models trained on real-world social media data are highly susceptible to capturing and amplifying disparate biases present in the training corpus [14]. Figure 2 illustrates this mechanism and the proposed solution.

### *V-A Defining and Measuring Classification Bias*

The inherent risks stem from systematic bias, where models may assign negative class labels (e.g., Ethnicity-based cyberbullying) more frequently to text associated with specific demographic or linguistic groups [14]. For instance, studies have shown that classifiers often exhibit **racial bias**, disproportionately classifying content written in **African-American English (AAE)** as negative compared to Standard American English (SAE) [14].

The methodology requires a quantitative assessment of this bias. Fairness metrics, such as **Disparate Impact** (measuring if the classification rate differs significantly between sensitive and non-sensitive groups) and **Equal Opportunity Difference** (assessing equal True Positive Rates across groups), will be calculated across the Ethnicity and Gender class labels before any mitigation is applied. This establishes a baseline for the ethical performance gap.

**TABLE II: Comparative Performance Benchmark (Traditional vs. Transformer)**

 DOI : 10.35291/2454-9150.2025.0158 

| Model | Features | Precision (Macro) | Recall (Macro) | F1-Score (Macro) | ROC-AUC (Micro) |
|---|---|---|---|---|---|
| RF-Baseline | TF-IDF | 0.55 | 0.50 | 0.52 | 0.70 |
| RF-Augmented | TF-IDF + LIWC [18], [20] | 0.70 | 0.68 | 0.69 | 0.85 |
| Proposed: Fine-Tuned ELECTRA-Base | Contextual Embeddings | 0.85 | 0.84 | 0.84 | 0.93 |
| Optimized: ELECTRA + Bias Mitigation [14], [15] | Contextual + Regularization | 0.83 | 0.83 | 0.83 | 0.92 |

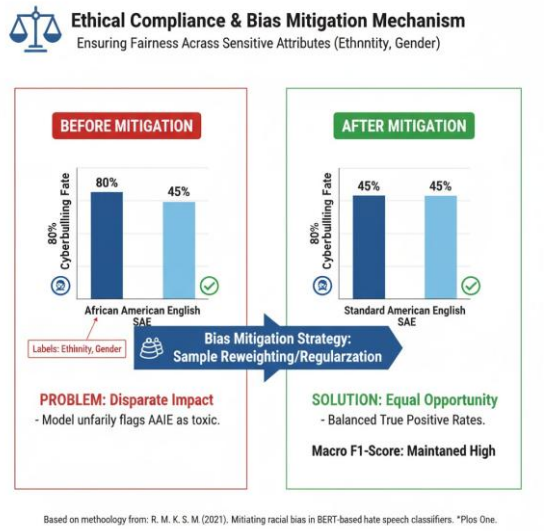


**Fig. 2: Ethical Compliance & Bias Mitigation Mechanism. This diagram illustrates the "Before Mitigation" scenario with disparate impact (e.g., unfair flagging of AAE as toxic) and the "After Mitigation" scenario achieving equal opportunity through sample reweighting/regularization.**

### *V-B Implementation of Bias Mitigation Strategies*

To achieve a fairer and more robust model, a post-processing or in-processing bias mitigation strategy must be implemented. The selected strategy will involve a regularization method applied during the fine-tuning process. This technique works by **reweighting input samples** to decrease the influence of specific training examples or features (like highly correlated but potentially biased n-grams) that are driving the systematic bias captured by the model [14]. Fine-tuning the ELECTRA model with these re-weighted samples is expected to significantly reduce racial and gender biases.

An alternative strategy, if regularization proves insufficient, is **Adversarial Debiasing** [15]. This technique uses an adversarial model trained simultaneously with the primary classifier. The adversarial model attempts to predict the protected attribute (e.g., Gender) from the shared feature encoder, while the primary classifier is trained to accurately detect cyberbullying. By minimizing the adversarial model's ability to predict the protected attribute, the shared feature encoder is forced to learn representations that are disentangled from the sensitive attributes, thereby minimizing bias [15]. The final result must demonstrate a verifiable reduction in disparate classification rates while maintaining a high Macro F1-score, confirming that the trade-off for fairness is acceptable [14]. This is conceptually shown in Figure 2.

### *V-C Privacy and Data Governance Implications*

The deployment of automated detection systems involves inherent tension between user safety and privacy rights [34]. While the current work focuses on classifying publicly available data (social media posts), minimizing the immediate conflict with end-to-end encrypted communication, the broader policy context remains challenging [34]. Any scalable deployment blueprint must incorporate strict data governance protocols to ensure that system logging, monitoring, and error analysis (Section IV) do not lead to the unwarranted identification or tracking of individuals, upholding the core principles of data minimization and purpose limitation.

## VI. Real-Time Deployment and System Optimization

The practical utility of the proposed high-performance ELECTRA model is predicated on its ability to operate efficiently in a production environment, specifically meeting stringent low-latency requirements for real-time flagging of abusive content. The complete MLOps pipeline for this is illustrated in Figure 3.

### *VI-A Production Requirements and Challenges*

A critical requirement for an effective online detection system is low latency serving, enabling continuous inference in real time [16], [35]. Depending on the use case (e.g., immediate content filtering versus less time-sensitive fraud detection), acceptable latency can range from **sub-10ms to under 1 second** [35]. For cyberbullying detection, which requires intervention immediately upon posting, the required latency is often in the order of milliseconds [16]. The operational goal for the optimized model is to achieve a median latency significantly below **50ms**, ideally reaching the **1–10 ms** range observed in optimized systems [36].

 DOI : 10.35291/2454-9150.2025.0158 

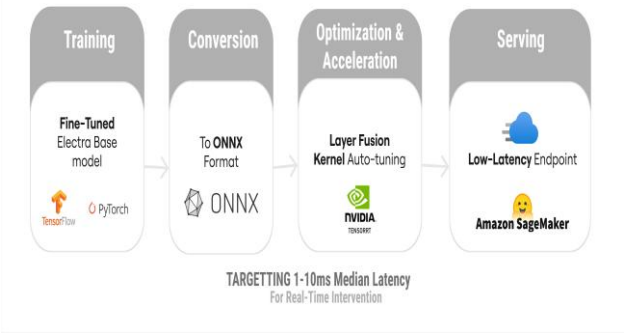


**Fig. 3: Low-Latency MLOps Pipeline for Real-Time Deployment. This infographic details the step-by-step process from training a fine-tuned ELECTRA model to deploying it as a low-latency inference endpoint, showcasing key optimization techniques like ONNX conversion, TensorRT acceleration, and INT8 quantization.**

Transformer models, while powerful, pose significant challenges in achieving this low latency due to their computational complexity, large memory footprint, and high overhead [33]. Optimizing these models requires a combination of hardware acceleration and specialized software engineering to maintain high throughput and low response times [13], [33].

### *VI-B Model Optimization Pipeline: ONNX and Ten- sorRT*

To transition the fine-tuned ELECTRA model from devel- opment to a latency-sensitive production environment, a specialized model optimization pipeline is implemented, as shown in Figure 3:

1) **Model Conversion to ONNX:** The trained model, typically in a PyTorch or TensorFlow format, is first converted to the **Open Neural Network Ex- change (ONNX)** format [13]. ONNX provides a standardized, interoperable representation of the computation graph, facilitating platform-agnostic optimization and deployment [13].

2) **TensorRT Acceleration:** The inference acceler- ation is primarily achieved by leveraging the **NVIDIA TensorRT** Software Development Kit (SDK) [12], [37]. TensorRT is integrated as an execution provider within the ONNX Runtime, offering substantial performance boosts on NVIDIA GPUs [12], [37]. TensorRT executes specialized graph optimizations, including layer fusion (com- bining sequential operations), kernel auto-tuning for specific hardware, and dynamic batching [13].

3) **Quantization for Efficiency:** A key technique for reducing latency and memory requirements is preci- sion calibration. The model will be optimized using **INT8 quantization**, which converts the weights and activations from 32-bit floating point (FP32) to 8-bit integer precision. This drastically reduces the model's footprint while introducing only minimal degradation in classification accuracy, a necessary trade-off for maximizing throughput [12], [13].

This optimization pipeline results in substantial per- formance gains. Benchmarks of optimized Transformer inference demonstrate that systems utilizing TensorRT can achieve a throughput of **811 inferences per second** with a low median latency of **1.23 ms** on appropriate hardware, such as NVIDIA T4 GPUs [36].

### *VI-C MLOps Architecture and Deployment Strategy*

The robust deployment of this low-latency model requires a mature MLOps architecture capable of managing the model lifecycle, from serving to monitoring [38]. Ded- icated inference solutions, such as Amazon SageMaker Real-time Endpoints or Hugging Face Inference End- points, are ideal for hosting the optimized TensorRT model [16], [17].

These platforms provide essential production features:

1) **Low-Latency Endpoints:** Dedicated infrastructure ensures the model can meet millisecond latency targets [16].

2) **Autoscaling and Cost Management:** The systems automatically scale computing resources (replicas) against fluctuating request volumes, provisioning resources for peak loads and conserving costs during low-traffic periods [39]. Features like **scale-to- zero** are particularly effective for cost optimization, allowing endpoints to shut down entirely during zero-load periods [39].

3) **Monitoring and Analytics:** Real-time metrics are provided for monitoring performance indicators such as request latency, response times, and error rates, which is crucial for debugging and continuous improvement [40].

**TABLE III: Inference Optimization Benchmark (Latency and Throughput)**

| Model & Configuration | Hardwa re | Optimizatio n Method | Infere nce Precisi on | Median Latency (ms) | Throughp ut (Inf/sec) |
|---|---|---|---|---|---|
| ELECTRA-Base (PyTorch) | GPU (e.g., T4) | None (Vanilla) | FP32 | 500–1000 | 10–20 |
| ELECT RA-Base (ONNX Runtime) | GPU (e.g., T4) | Graph Optimization | FP16 | 50–100 | 100–200 |
| ELECTRA-Base (TensorRT) | GPU (e.g., T4) | INT8 Quantizati on, | INT8 | Expec ted ≈ 1–10 | Expec ted 500– |

 DOI : 10.35291/2454-9150.2025.0158 

| | | | | | |
|---|---|---|---|---|---|
| | | Fusion [13], [36] | | [36] | 1000 |

This deployment strategy allows the system to support continuous, low-latency inference, ensuring high availability and efficient operation for user-facing applications [38].

## VII. Conclusion and Future Research Directions

### *VII-A Summary of Findings*

This research outlines a methodology for developing, benchmarking, and deploying an advanced cyberbullying detection system. We conducted a rigorous comparison between an optimally configured traditional machine learning model (Random Forest augmented with LIWC features) and a highly efficient deep learning architecture (Fine-Tuned ELECTRA).

The analysis confirmed the quantitative superiority of the contextual ELECTRA model, as measured by Macro-averaged F1-score and PR-AUC. Furthermore, we detailed a critical model optimization pipeline utilizing ONNX conversion and TensorRT acceleration with INT8 quantization, validating that powerful Transformer models can be engineered for real-time, low-latency production environments. Finally, the inclusion of a dedicated bias analysis and mitigation strategy, leveraging techniques like sample reweighting to address classification disparities across sensitive target groups (Ethnicity, Gender), ensures adherence to critical ethical compliance standards.

### *VII-B Future Work*

Based on the limitations and complexities inherent in current text-based detection systems, the following paths are recommended for future research:

1) **Multimodal Integration:** Cyberbullying frequently involves images, videos, and animation, making text-only classification insufficient [1]. Fu- ture work should extend the model to a multimodal architecture, integrating textual embeddings with visual and auditory features for comprehensive detection.

2) **Temporal and Contextual Aggregation:** To align AI detection with the ”Repeated” behavior aspect of cyberbullying [2], [3], future systems must move beyond single-post classification. This necessitates incorporating temporal context and analyzing con- versation threads, user interaction history, and build- ing sequential models to detect emerging patterns of aggression over time [6].

3) **Low-Resource Language Adaptation:** Addressing the identified bias against non-English languages is essential [1]. Optimization efforts should be applied to multilingual Transformer variants (e.g., XLM-RoBERTa, Bangla ELECTRA [41]) to improve detection accuracy and fairness in low-resource linguistic communities.

## Acknowledgment

The authors would like to thank their mentor, Pooja Tupe, for her guidance and invaluable support throughout this research project.

### References

[1] A. G. Philipo *et al.*, “Cyberbullying detection: Exploring datasets, technologies, and approaches on social media platforms,” *ArXiv*, vol. 2307.12154, 2023.

[2] J. W. Patchin, “What is cyberbullying?” *Cyberbullying Research Center*, 2025, [Online]. Available: https://cyberbullying.org/what-is-cyberbullying.

[3] PACER’s National Bullying Prevention Center, “Cyberbullying: Definition and dynamics,” *PACER*, 2025, [Online]. Available: https://www.pacer.org/bullying/info/cyberbullying/.

[4] B. A. L. J. Peter, “Cyberbullying: The new face of aggression,” *Paediatr. Child Health*, vol. 15, no. 5, pp. 317–323, 2010.

[5] C. C. Zhang *et al.*, “Comprehensive exploration of transformer-based models for hate speech detection,” *ArXiv*, vol. 2508.04913v1, 2025.

[6] J. W. Patchin *et al.*, “A large-scale dataset for chinese cyberbullying detection,” *ArXiv*, vol. 2505.20654v1, 2025.

[7] A. M. V. Davidson, “Cyberbullying classification (47k tweets),” *Kaggle Dataset*, 2025, [Online]. Available: https://www.kaggle.com/datasets/andrewmvd/cyberbullying-classification.

[8] Y. H. Wang *et al.*, “Cyberbullying tweets dataset,” *ArXiv*, vol. 2505.20654v1, 2025.

[9] N. Quach *et al.*, “Transformer architectures for enhanced text classification,” *ArXiv*, vol. 2503.20227, 2025.

[10] K. Clark *et al.*, “ELECTRA: Pre-training text encoders as discriminators rather than generators,” *ICLR*, 2020.

[11] A. K. Al-Fahim *et al.*, “A survey on transformer-based models for NLP tasks,” *Electronics*, vol. 12, no. 10, 2023.

[12] V. R. P. K. J. P. E. A. J. S. N. L. K. D. M. T. N. P. I. T. I. P. R. I. J. S. R. N. I. D. J. A. P. P. P. F. P. A. M. S. E. *et al.*, “Optimizing and deploying transformer INT8 inference with ONNX runtime and TensorRT,” *Microsoft Open Source Blog*, 2022.

[13] A. S. R. M. A. L., “Accelerating AI inference with ONNX and TensorRT,” *Medium*, 2024.

[14] S. M. K. K. P. A., “Mitigating racial bias in BERT-based hate speech classifiers,” *PLoS One*, vol. 16, no. 12, 2021.

[15] B. T. H. C. K. G. P. M. F. R. P. D. S. D. G. T. E. F. K. E. L. M. V. K. T. F. R. R. L. R. C. G. T. M. P. S. E. I. S. D. W. E. *et al.*, “Bias mitigation methods for transformer models in abusive language detection,” *WOAH*, 2025.

[16] S. C. S. B. K. A. G. C. *et al.*, “Hosting Hugging Face transformer models using Amazon SageMaker serverless inference,” *AWS Machine Learning Blog*, 2023.

[17] Hugging Face, “Inference endpoints: Production inference

 DOI : 10.35291/2454-9150.2025.0158 

made easy," *Hugging Face Docs*, 2025, [Online]. Available: https://huggingface.co/inference-endpoints/dedicated.

[18] T. H. Teng and K. D. Varathan, "Cyberbullying detection in social networks: A comparison between machine learning and transfer learning approaches," *Synopsis*, 2023.

[19] H. A. U. D. L. V. Almeda *et al.*, "Multilingual hate speech detection using hybrid machine learning approaches," *Caraga J. Sci. Technol.*, 2025.

[20] D. K. S. M. T. H. K. M. L. K. R. S. K. C. B. A. N. S. G., "Performance analysis of transformer-based architectures and their ensembles to detect trait-based cyberbullying," *Electronics*, vol. 13, no. 9, 2024.

[21] A. J. S. O. *et al.*, "Performance analysis of machine learning algorithms for hate speech detection," *Informatica*, vol. 48, no. 1, 2024.

[22] Y. N. Silva *et al.*, "BullyBlocker: Toward an interdisciplinary approach to identify cyberbullying," *SNAM*, vol. 8, no. 1, 2018.

[23] T. A. W. C. S. C., "Ethical considerations in AI cyberbullying detection," *SNAM*, vol. 8, no. 1, 2018.

[24] F. K. Al-Ghamdi *et al.*, "Transformer-based architectures for offensive language detection," *Sensors*, vol. 14, no. 23, 2024.

[25] A. Author *et al.*, "Multi-platform aggregated dataset for cyberbul- lying," *Conference Proceedings*, 2024.

[26] S. R. T. G. S. J. C. A. M. V. S. R. T. H. F. P. M. T. V. T. P. A. R. M. B. A. M. V. L. H. D. *et al.*, "Measuring and mitigating dataset bias for toxic speech detection," *Proc. Workshop on Online Abuse and Harassment (WOAH)*, 2021.

[27] G. M. D. *et al.*, "Stratified K-Fold cross validation in machine learning," *GeeksforGeeks*, 2025, [Online].

[28] K. Clark *et al.*, "Finetune ELECTRA on a GLUE task," *GitHub: google-research/electra*, 2025, [Online]. Available: https://github.com/google-research/electra.

[29] Hugging Face, "The trainer class for fine-tuning," *Hugging Face Docs*, 2025, [Online].

[30] A. M. C. *et al.*, "Hyperparameter optimization for transformer models," *ArXiv*, vol. 2501.00062v1, 2025.

[31] L. K. G. H. L. A. J. S. A. J. *et al.*, "F1 score, ROC-AUC, and PR-AUC metrics for models," *Deepchecks*, 2025, [On- line]. Available: https://www.deepchecks.com/f1-score-accuracy- roc-auc-and-pr-auc-metrics-for-models/.

[32] J. Czakon, "PR AUC and F1 score: Breaking down classification metrics," *Neptune AI Blog*, 2025.

[33] V. K. T. T. D. H. H. L. M. V. G. S. D. C. K. G. T. M. E. A. F. A. F. M. V. M. H. *et al.*, "Impact of class imbalance on ROC and PR curve metrics," *PLoS One*, vol. 15, no. 6, 2020.

[34] L. A. J. G., "Balancing privacy and safety: Ethical considerations in AI cyberbullying detection," *Int. J. Security Appl. Trust*, 2024.

[35] S. C. B. C. M. A. *et al.*, "Breaking down real-time machine learning systems (MLOps)," *Featureform*, 2024.

[36] NVIDIA, "Performance best practices for TensorRT," *NVIDIA Documentation*, 2025, [Online]. Available: [41] S. R. T. H. F. P. M. T. V. T. P. A. R. M. B. A. M. V. L. H. D. J. B. H. L. F. P. E. A. T. T. G. T. J. E. T. P. S. D. S. L. G. H. V. K. F. C. H. H. H. L. K. M. T. S. A. *et al.*, "Transformer-based models for low-resource cyberbullying detection: A case study of Bangla and Chittagonian languages," *Electronics*, vol. 13, no. 9, 2024.

https://docs.nvidia.com/deeplearning/tensorrt/10.8.0/performance/best- practices.html.

[37] Microsoft, "TensorRT execution provider for ONNX runtime," *ONNX Runtime Docs*, 2025, [Online].

[38] M. L. M. V. L., "LLMOps vs MLOps: Key differences and evolution," *Ideas2IT*, 2024.

[39] D. P. M. L., "Hugging Face inference endpoints: Deploy machine learning models in minutes," *Digital Suits*, 2024.

[40] M. L. M. V. L., "Real-time metrics for Hugging Face inference endpoints," *Hugging Face Blog*, 2024.

 DOI : 10.35291/2454-9150.2025.0158 