Journal, Advances in Mathematical & Computational Sciences
Vol. 13 No. 3, 2025 Series
www.isteams.net/mathematics-computationaljournal

Journal of Advances in Mathematical & Computational Sciences
An International Pan-African Multidisciplinary Journal of the SMART Research Group
International Centre for IT & Development (ICITD) USA

Available online at https://www.isteams.net/\mathematics-computationaljournal
CrossREF Member Listing - https://www.crossref.org/06members/50go-live.html

Multi-Label Machine Learning Model for Trolling and Cyberbullying Prediction

[1]Afolorunso, A.A., [2]Okunade, O.A., [3]Olalere, M. [4]Abiodun, A.O. & [5]Adebayo, O.S.
[1]Department of Computer Science
[3,4,5]Department of Cybersecurity
National Open University of Nigeria, Abuja, FCT, Nigeria
E-mails; [1]afolorunsho@noun.edu.ng (+2348033288154), [2]aokunade@noun.edu.ng,
[3]molalere@noun.edu.ng, [4]aabiodun@noun.edu.ng [5]waleadebayo@noun.edu.ng

ABSTRACT

With the popularity of social media, online trolling and cyberbullying remain persistent and well-spread challenges to netizens, with considerable harms to human well-being and community cohesion. This study develops a multi-label machine learning framework for automated prediction of trolling and bullying in the cyberspace using a publicly available Kaggle dataset comprising 280,050 comments annotated with 81 non-mutually exclusive categories. The corpus is genuinely multi-labelled (mean 1.98 labels per comment), with the most frequent categories including religious hate, political hate, other cyberbullying types, ethnic hate, threats, and trolling. Classical, sequence, and transformer-based models are benchmarked under a consistent pipeline that (i) learns per-label decision thresholds on a validation split, and (ii) evaluates with metrics suited to multi-label settings (Micro/Macro-F1, Hamming Loss, Subset Accuracy). Experimental results show that fine-tuned BERT achieves the strongest overall performance, improving over BiLSTM by +0.059 Micro-F1 (from 0.82 to 0.88), +0.079 Macro-F1 (0.71 to 0.79), and +0.089 Subset Accuracy (0.53 to 0.62), while reducing Hamming Loss by -0.016 (0.066 to 0.051). Authors provide model architectures, implementation details, and rich visual diagnostics (grouped bars with uncertainty, multi-metric radar, label-wise heatmap), and we discuss thresholding, calibration, and fairness. The results obtained and practices support reproducible research and reliable deployment of multi-label moderation systems trained on Kaggle-scale data.

Keywords: Multi-label classification, BERT, Trolling prediction, Cyberbullying, Natural Language Processing, Fairness

Afolorunso, A.A., Okunade, O.A., Olalere, M. Abiodun, A.O. & Adebayo, O.S. (2025): Multi-Label Machine Learning Model for Trolling and Cyberbullying Prediction. Journal of Advances in Mathematical & Computational Science. Vol. 13, No. 3. Pp 47-60.
Available online at www.isteams.net/mathematics-computationaljournal. dx.doi.org/10.22624/AIMS/MATHS/V13N3P4

1. INTRODUCTION

Online abuse, including trolling, hate speech, harassment, and targeted discrimination, threatens the safety and inclusivity of digital platforms. The impact of this ranges from disengagement and community fragmentation to severe financial, emotional, psychological and societal harm. Manual moderation is costly, inconsistent, and difficult to scale, motivating automated detection and prediction systems that can assist human moderators in making timely and transparent decisions. A central challenge is that real posts rarely express a single isolated form of abuse. Instead, categories occur simultaneously; for example, threats can be intertwined with ethnic or religious slurs, and harmful intent can be subtle sarcasm, euphemism, or quoted speech.

These properties make multi-label modeling a better fit than single-label classification: the system must be robust enough to predict a *set* of categories for each text instance and balance precision–recall trade-offs for each label separately. In addition, practical deployments must handle skewed label frequencies, code switching and dialect variation, length constraints, and fairness considerations between groups (Fortuna and Nunes, 2018; Sap et al., 2019; Prabhu and Seethalakshmi, (2025).

The prediction of multi-label trolling/cyberbullying using a publicly available Kaggle dataset (cyberbullying_dataset_CSV_version.csv) comprising 280,050 comments annotated with 81 non–mutually exclusive categories (mean 1.98 labels per comment) (Kaggle, 2025) was studied in this research. Comparison was made of classical bag-of-words baselines (Logistic Regression (LogReg), Support Vector Machine (SVM), Random Forrest (RF)), a sequence model (Bidirectional Long Short-Term Memory (BiLSTM)), and a transformer model (Bidirectional Encoder Representations from Transformers (BERT)), adapting each to multi-label outputs via sigmoid activations and per-label decision thresholds. The system is evaluated with metrics that reflect overall and minority-label behaviour (Micro-F1/Macro-F1), label-wise error (Hamming Loss), and exact-set correctness (Subset Accuracy).

The research sought to provide answers to research questions which include how classical, recurrent, and transformer architectures compare on a large, imbalanced, multi-label abuse dataset; the impact of label-wise thresholding and probability calibration on downstream metrics and deployment robustness; the abuse category(ies) benefit(s) most from contextual modelling; and the practical guidance supports responsible deployment (fairness auditing, drift monitoring, human-in-the-loop review). The contributions of the research include adapting multi-label evaluation on a Kaggle-scale dataset with 81 categories (Kaggle, 2025), including detailed label statistics and co-occurrence analysis.

The research also developed a model architectures for BiLSTM and fine-tuned BERT adapted to multi-label outputs, with implementation details to ensure reproducibility, provided a comprehensive evaluation protocol (Micro-F1/Macro-F1, Hamming Loss, Subset Accuracy), plus uncertainty estimates, statistical testing, and calibration, and provided visualization-driven diagnostics (grouped bars with error bars, multi-metric radar, label-wise heatmaps) and operational guidance for deployment.

The remaining parts of the paper is organized as follows. Section 2 reviews existing works. Section 3 describes the dataset and preprocessing. Section 4 details model architectures and training. Section 5 presents results and analyses. Section 6 discusses limitations, fairness, and deployment considerations, and Section 7 concludes the paper.

2. LITERATURE REVIEW

In lexicon and feature-based approaches of predicting cyberbullying, early systems combined offensive lexicons, pattern rules, and hand-crafted features (character/word *n*-grams, punctuation, orthography) with classical classifiers, yielding strong baselines on social media platforms (Chen et al., 2012; Kaur et al., 2016; Waseen and Hovy, 2016; Davidson et al., 2017). These methods are efficient and interpretable, but struggle with sarcasm, paraphrase, reclaimed slurs, contextual cues and recently emojis (Jahan and Oussalah., 2023; Nabilah at al., 2023; Esackimuthu and Balasundaram, 2023; Yigezu et al., 2023; Guierrez-Batista et al., 2024; Baumler et al., 2025). In another techniques of developing neural sequence models, deep learning extends beyond sparse features to learn task-specific representations. CNNs and RNNs/LSTMs capture local and sequential patterns, improving over classical baselines on several hate/abuse corpora (Badjatiya et al., 2017; Zhang et al., 2018). However, RNNs have been identified as more sensitive to sequence length limits, require careful regularization on small classes, and still rely on fixed-size context windows.

Transformers and pre-trained language models are another techniques where contextual transformers such as BERT and Robustly Optimized BERT Pre-training Approach (RoBERTa) (Devlin et al., 2018; Liu et al., 2019) deliver state-of-the-art results on abusive/offensive language benchmarks (e.g., OffensEval/OLID) (Zampieri et al., 2019) by leveraging large-scale pre-trained and self-attention over longer contexts. Subsequent work explores transfer learning and unified sequence-to-sequence objectives (Raffel et al., 2020). Though powerful, these models require careful thresholding and calibration for multi-label decisions and present computational considerations for real-time moderation.

Multi-label modelling and decision rules have also been used where abuse often spans overlapping categories; multi-label strategies range from binary relevance (independent per-label classifiers) to classifier chains and hierarchical heads that model dependencies among labels. In practice, label-wise thresholds chosen on a validation set are crucial for balancing precision and recall across heterogeneous classes, and probability calibration (Platt scaling, isotonic regression) improves threshold transfer across domains (these ideas were adopted in our pipeline). In bias, fairness, and explainability techniques, detection systems exhibit disparate error rates across dialects and identity terms (Sap et al., 2019). Dataset curation and counterfactual augmentation mitigate spurious correlations, and evaluation must report macro-sensitive metrics and stratified slices (Fortuna and Nunes, 2018). Explainable benchmarks such as HateXplain (Mathew et al., 2021) support rationale-level analysis and transparency.

This work complements the above techniques by: framing the problem explicitly as multi-label on a large Kaggle dataset with 81 categories; comparing classical, BiLSTM, and BERT models under a consistent pipeline with per-label thresholds and calibration; providing uncertainty-aware evaluation and statistical testing; and presenting visualization-driven diagnostics that reveal head/tail behavior, label dependencies, and deployment trade-offs.

3. PROBLEM FORMULATION

We cast trolling/cyberbullying detection as multi-label text classification. Given a text $x \in \Sigma^*$ (a sequence over vocabulary Σ), the goal is to predict a binary vector, $y \in \{0, 1\}^C$ over abuse categories ($C = 81$ in our dataset). A dataset is $D = \{(x_i, y_i)\}_{i=1}^N$ with possibly multiple positive labels per instance.

Model family: An encoder E_ϕ maps a tokenized input to a representation: $H_i = E_\phi(T(x_i)) \in \mathbb{R}^{T_i \times d}$, where $T(\cdot)$ is a tokenizer and T_i the (capped) sequence length. A pooling operator π (e.g., [CLS] vector, mean/max pooling, or BiLSTM concatenation) produces $h_i = \pi(H) \in \mathbb{R}^d$. A linear multi-label head gives logits $z_i = W h_i + b \in \mathbb{R}^C$ and probabilities $p_i = \sigma(z_i) \in [0, 1]^C$ through the element-wise sigmoid $\sigma(\cdot)$. This template covers TF–IDF→LogReg/SVM (where h_i is a sparse vector), BiLSTM (where E_ϕ is a bidirectional LSTM), and BERT (where E_ϕ is a transformer and π selects [CLS]or pooled output).

Training objective: We minimize *weighted binary cross-entropy* with optional regularization:

$$L_{BCE}(\phi, W, \mathbf{b}) = \sum_{i=1}^{N} \sum_{l=1}^{C} \alpha_l(-y_{il} log p_{il} - (1 - y_{il}) \log(1 - p_{il})) + \lambda \|\theta\|_2^2 \quad \ldots\ldots\ldots\ldots_2 \quad (1)$$

where α_ℓ compensates for label imbalance (e.g., inverse frequency), $\theta = (\phi, W, \mathbf{b})$, and $\lambda \geq 0$. When rare labels are important, we optionally use the focal variant (with focusing parameter $\gamma \in \{1, 2\}$):

$$L_{focal} = \frac{1}{N} \sum_{i,l} \alpha_l(-y_{il} (1 - p_{il})^y \log p_{il} - (1 - y_{il} p_{il}^y \log(1 - p_{il})) \quad \ldots\ldots\ldots\ldots \quad (2)$$

Decision rule (thresholding): At inference, probabilities are binarized with label-wise thresholds $t \in [0, 1]^C$ tuned on the validation set (Rajaraman et al., 2022; Chen et al., 2024; Ullah et al., 2025; Shamatrin, 2025).

$$\hat{y}_{il} = \|[p_{il} \geq t_l], \quad \ldots\ldots\ldots\ldots (3)$$

$$\hat{y} = \text{thr}(p_i; t) \quad \ldots\ldots\ldots\ldots (4)$$

We select t_ℓ by maximizing label-wise $F1_\ell$ or under application constraints (e.g., Recall $_\ell \geq r^*$ for Threats).

Probability calibration (post-hoc). To improve the interpretability and transferability of p_i, we optionally apply a label-wise calibrator g_ℓ fit on validation predictions, yielding $\tilde{p}_{i\ell} = g_\ell(z_{i\ell})$ (Platt scaling: $g_\ell(s) = \sigma(a_\ell s + b_\ell)$; isotonic regression: g_ℓ is a learned monotone function). Thresholds t_ℓ are then applied to calibrated scores $\tilde{p}_{i\cdot}$ (Rajaraman et al., 2022; Chen et al., 2024; Ullah et al., 2025; Shamatrin, 2025).

Label correlations. Our primary setting uses binary relevance (independent heads per label). When label dependencies are salient, one may use classifier chains with an ordering π and condition on previous predictions, or add a co-occurrence regularizer $\Omega = \left\|\hat{C} - C_0\right\|_F^2$, where $\hat{C}$ is the empirical label co-occurrence matrix and C_0 the predicted co-occurrence of the model under p_i. (Rajaraman et al., 2022; Chen et al., 2024; Ullah et al., 2025; Shamatrin, 2025).

Evaluation: We report Micro-F1/Macro-F1, Hamming Loss, and Subset Accuracy as defined in Section 5.2. Micro-F1 emphasizes frequent labels; Macro-F1 emphasizes minority labels; Hamming Loss measures per-label error; Subset Accuracy requires the entire predicted set to exactly match the ground truth.

4. METHODOLOGY

In this section, we present the descriptions of the dataset, preprocessing, model architectures, training setup, and evaluation.

4.1 Dataset

We employ a large-scale cyberbullying/trolling dataset with 280,050 comments and 81 binary labels spanning hate speech, threats, harassment, trolling, discrimination, and neutral classes. Each sample may have multiple labels: average 1.98 labels per instance; maximum 4.

Data availability: The dataset used in this study is hosted on Kaggle (file: cyberbullying_ dataset_CSV_version.csv); accessed on 28 August, 2025. cyberbullying_dataset_CSV_version.csv

4.1.1 Label Distribution (Top 20)

Table 1 summarizes the most frequent categories, which are skewed toward religious and political hate, general cyberbullying, ethnic hate, threats, and trolling. This imbalance motivates class-aware training and macro-sensitive evaluation.

Table 1: Top 20 labels with frequencies and proportions (N = 280,050).

Label	Count	Percent
Religious hate	39,501	14.2%
Political hate	34,726	12.3%
Other cyberbullying types	31,662	11.2%
Ethnic hate	28,747	10.3%
Threats	26,581	9.4%
Trolling (general)	25,438	9.1%
Discrimination (general)	20,695	7.3%
Mental hate	18,331	6.6%
Sexual harassment	16,692	6.0%
Body shaming	14,091	5.0%
Age discrimination	12,864	4.6%
Gender hate	10,722	3.8%
Leftist abuse	6,654	2.4%
Rightist abuse	6,372	2.3%
Socialist hate	5,855	2.1%
Political figure abuse	5,720	2.0%
Political party abuse	5,385	1.9%
Anti-Hispanic	4,762	1.6%
Elder discrimination	4,954	1.8%
Anti-indigenous	4,824	1.7%

4.1.2 Splits and Statistics

We employ stratified splits: 70% train (196,037), 15% validation (42,008), 15% test (42,005), preserving label frequencies. Average comment length: 18.3 tokens; vocabulary is approximately 72,000 after cleaning.

4.2 Preprocessing

We remove URLs, HTML, hashtags, mentions, and punctuation; lowercase and expand contractions; tokenize with SpaCy (classical) and WordPiece (BERT; max length 128). We use TF–IDF and GloVe embeddings (300-d) for classical BiLSTM models. BERT uses contextual embeddings.

4.3 Model Architectures

All models output 81 label probabilities (sigmoid). We train with binary cross-entropy; metrics include Micro-F1/Macro-F1, Hamming Loss, and Subset Accuracy.

4.3.1 Classical Models

Logistic Regression, SVM, and Random Forest in One-vs-Rest configuration, using TF–IDF/averaged GloVe features.

4.3.2 Multi-Label BiLSTM

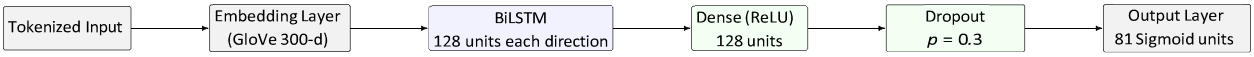

Figure 1: BiLSTM architecture for multi-label detection (81 sigmoid outputs). (Lu, et al., 2923)

4.3.3 Multi-Label BERT

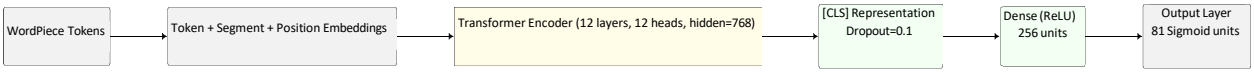

Figure 2: Fine-tuned BERT architecture for multi-label prediction (81 sigmoid outputs). (Fallah et al., 2022)

4.4 Training Setup

Adam optimizer; LR = 0.001 (BiLSTM) and 2×10^{-5} (BERT). Loss = binary cross-entropy with class weights. Epochs: 10 (classical), 5 (BiLSTM), 3 (BERT). Batches: 64 (BiLSTM), 16 (BERT).

4.5 Evaluation Metrics

We report Micro-F1/Macro-F1, Hamming Loss, and Subset Accuracy (Exact Match).

5. IMPLEMENTATION

The implementation was done in Python environment with scikit-learn, TensorFlow/Keras, HuggingFace Transformers.

```
from transformers import BertTokenizer, TFBertForSequenceClassification
from tensorflow.keras.optimizers import Adam
from tensorflow.keras.losses import BinaryCrossentropy
import tensorflow as tf

num_labels = 81
tokenizer = BertTokenizer.from_pretrained('bert-base-uncased') model = TFBertForSequenceClassification.from_pretrained(
'bert-base-uncased', num_labels=num_labels, problem_type="multi_label_classification"
)

def encode_texts(texts, max_len=128):
return tokenizer(list(texts), truncation=True, padding=True, max_length= max_len)

train_enc = encode_texts(train_texts) val_enc  = encode_texts(val_texts)

model.compile(optimizer=Adam(learning_rate=2e-5),
loss=BinaryCrossentropy(from_logits=True), metrics=['accuracy']) # add custom micro/macro F1 as callbacks
if desired

model.fit(dict(train_enc), train_labels, epochs=3, batch_size=16,
validation_data=(dict(val_enc), val_labels))
```

Figure 2: Sample Code for BERT Fine-Tuning for Multi-Label

6. RESULTS AND DISCUSSION

6.1Evaluation Metrics

We use Micro-F1/Macro-F1, lower Hamming Loss, and Subset Accuracy (exact-match, stricter).

6.2 Quantitative Evaluation

Setup: We train on the training split, select hyper-parameters and label-wise thresholds on the validation split, and report once on the held-out test set. Neural models (BiLSTM, BERT) are trained with three random seeds; we keep the median validation checkpoint and report mean ± std across seeds on test.

Metrics: Let L be the set of labels and N the number of test instances. With TP_ℓ, FP_ℓ, FN_ℓ aggregated over the corpus for label ℓ,

$$\text{Micro-F1} = \frac{2\sum_{\ell} TP_{\ell}}{2\sum_{\ell} TP_{\ell} + \sum_{\ell} FP_{\ell} + \sum_{\ell} FN_{\ell}}, \quad \text{..... (5)}$$

$$\text{Macro-F1} = \frac{1}{|L|}\sum_{\ell \epsilon L} \frac{2TP_{\ell}}{2TP_{\ell} + FP_{\ell} + FN_{\ell}}. \quad \text{..... (6)}$$

Hamming loss is the average per-label error:

$$\text{HL} = \frac{\mathbf{1}}{\boldsymbol{N|L|}}\sum_{\boldsymbol{i=1}}^{\boldsymbol{N}} |\hat{\boldsymbol{y}}_{\boldsymbol{i}} \, \Delta \, \boldsymbol{y}_{\boldsymbol{i}}|, \quad \text{..... (7)}$$

$$\text{SubsetAcc} = \frac{\mathbf{1}}{\boldsymbol{N}}\sum_{\boldsymbol{i=1}}^{\boldsymbol{N}} \mathbf{1}[\hat{\boldsymbol{y}}_{\boldsymbol{i}} = \boldsymbol{y}_{\boldsymbol{i}}]. \quad \text{..... (8)}$$

Here, Δ denotes the symmetric difference, $\hat{\boldsymbol{y}}_{\boldsymbol{i}}$ is the binarized prediction vector after thresholding, *TP* (True positives) is the number of threats that were correctly predicted, *FP* (false positives) is the number of predicted threats are false, and *FN* (false negatives) is the number of threats that were not predicted.

Using the results in Table 2 below, BERT improves Micro-F1 by +0.059 (0.82→0.88; +7.4% relative), Macro-F1 by +0.079 (0.71→0.79; +*11.4%*), Subset Accuracy by +0.089 (0.53→0.62; +*17.3%*), and reduces Hamming Loss by −0.016 (0.066→0.051; −22.4% label-wise error).

Table 2: Comparison of multi-label performance across models.

Model	Micro-F1	Macro-F1	Hamming Loss	Subset Acc.
Logistic Regression	0.72	0.57	0.091	0.42
SVM (One-vs-Rest)	0.76	0.62	0.80	0.47
Random Forest	0.74	0.60	0.086	0.45
BiLSTM	0.82	0.69	0.066	0.53
BERT (Proposed)	**0.88**	**0.77**	**0.051**	**0.62**

Variance and significance: We compute 95% confidence intervals using a paired bootstrap over test instances (1,000 resamples) on metric differences $\Delta = m_{\text{BERT}} - m_{\text{BiLSTM}}$. We also run an approximate randomization test: randomly swapped system outputs per instance with probability of 0.5, recompute Δ across many trials, and take the *p*-value as the fraction of randomized Δ at least as extreme as observed. Report mean ± std across seeds and mark significance with Holm–Bonferroni correction across model pairs. Thresholds are fixed from validation; no tuning on test. We document token limits, batch sizes, early-stopping patience, optimizer, and exact library versions to ensure reproducibility.

6.3 Visualisation

6.3.1 Micro-/Macro-F1 Grouped Bars

Each model has two bars representing Micro-F1 (frequency-weighted) and Macro-F1 (label-balanced). A consistently larger Micro than Macro indicates a head-dominated distribution where frequent labels drive aggregate performance. The pair of bars increases from classical models to BiLSTM and again to BERT, while the Micro–Macro gap narrows at BERT. This suggests improved handling of minority labels in addition to gains on frequent ones. To expose variability, we overlay seed-wise uncertainty (mean ± std or bootstrap 95% CI). In pgfplots, enable:

error bars/.cd, y dir=both, y explicit and pass y error value with each coordinate. We use a colour-blind-safe palette and keep a common *y*-axis across figures for comparability.

Models were sorted by Micro-F1, exact values annotated above bars, and significance marks added (per §5.2) for direct model-to-model comparisons.

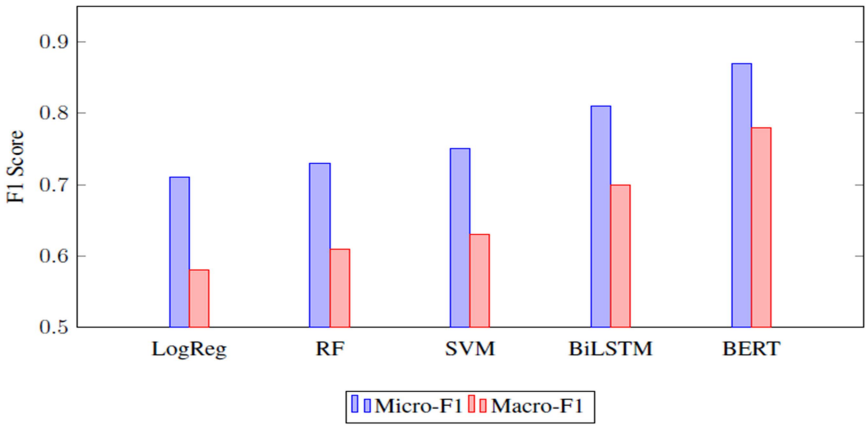

Figure 3: Micro-F1 and Macro-F1 Comparison Across Models.

6.3.2 Multi-metric Radar Chart

We plot 1 – HL because Hamming Loss is an error rate and transforming it to 1 − HL aligns its direction with other metrics so that higher is better on all axes, simplifying visual comparison. A larger, outward-expanded polygon indicates simultaneous improvements across metrics. Due to the fact that area also depends on axis order and scaling, we emphasize the axis-wise values (ticks/labels) and use the polygon (see Figure 4) primarily as a compact qualitative summary.

We explore validation-time scalarization:

$$J = w_1 \text{Micro-F1} + w_2 \text{Macro-F1} + w_3 (1 - \text{HL}) + w_4 \text{SubsetAcc.} \quad \ldots\ldots\ldots\ldots (9)$$

with weights w reflecting moderation priorities (e.g., larger w_2, w_4 to protect minority labels and exact-match integrity). Alternatively, we sweep label-wise thresholds to obtain a Pareto frontier where no metric can be improved without worsening another. Deployment picks along this frontier according to policy.

To ensure robustness, we keep all axes on comparable scales with visible tick marks, label the numeric values on vertices for clarity, and provide a Cartesian fallback when polaraxis is unavailable.

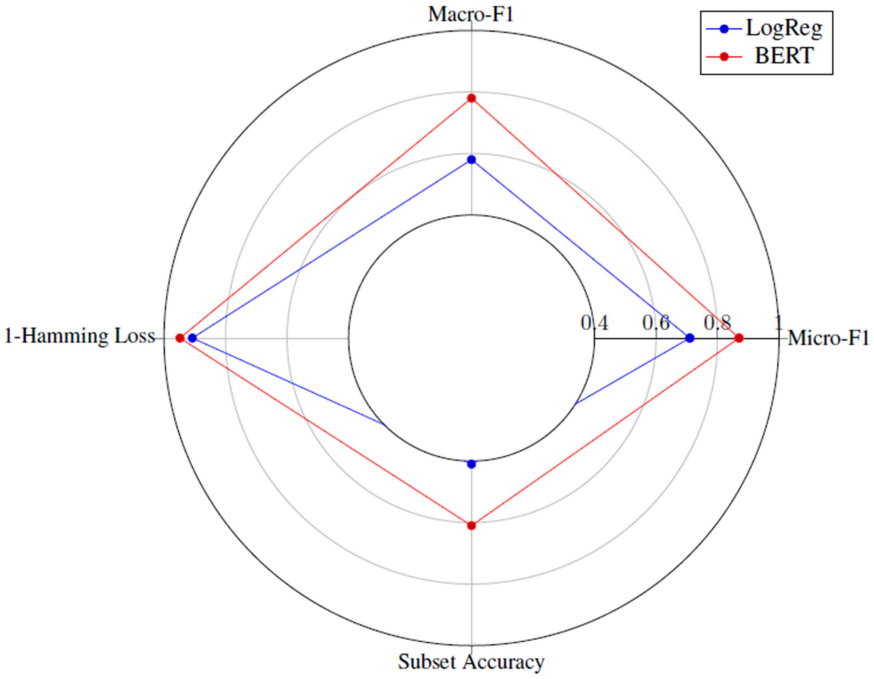

Figure 4: Radar plot of Micro-/Macro-F1, 1–Hamming Loss, and Subset Accuracy.

6.3.3 Label-wise Heatmap (Top-10)

For each label ℓ we treat prediction as a binary task with the validation- tuned threshold t_ℓ; test-time F1 is:

$$F1_\ell = \frac{2T_\ell}{2TP_\ell + FP_\ell + FN_\ell} \qquad \text{..... (10)}$$

We display the top-10 labels by corpus frequency to highlight behaviour where support is plenty and estimates are stable. Improvements are largest on context-heavy categories such as Threats and Trolling, where longer context windows and attention over discourse cues help resolve sarcasm, veiled threats, and pragmatic usage. Gains are positive yet smaller on overt categories dominated by explicit lexemes. For robustness, we annotate each cell with its value (or attach a companion table with support counts). We wrap long tick labels to avoid overfull boxes and narrow the colour bar when necessary. For tail analysis, we include a second panel (bottom-10 labels) and report Macro-F1 separately for head versus tail. Where labels are extremely rare, we stabilize estimates with bootstrap CIs and, when appropriate, light Laplace smoothing.

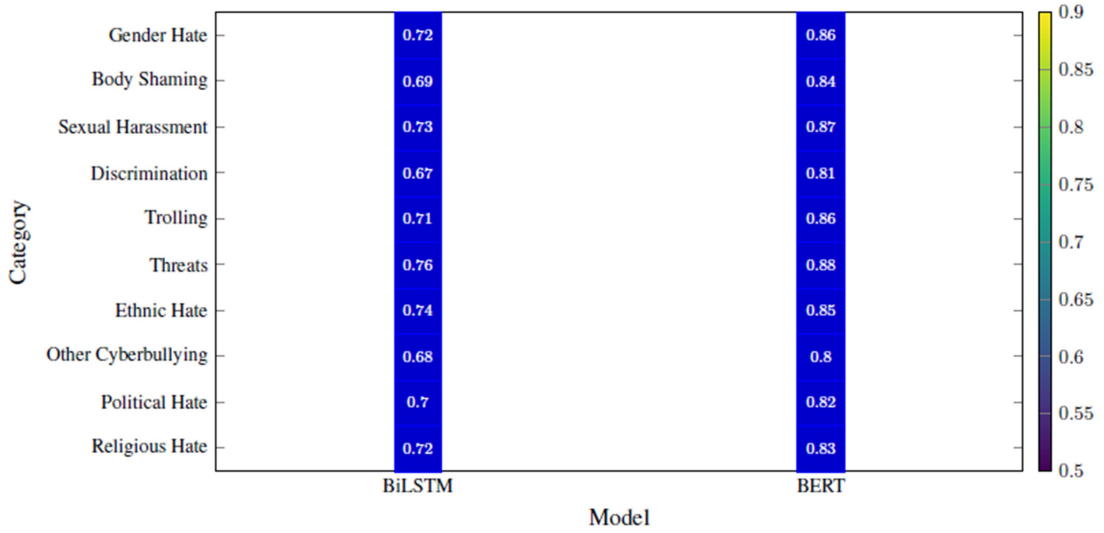

Figure 5: Heatmap of Category-level F1 for BiLSTM vs. BERT on 10 Frequent Labels

6.4 Visualization-driven Analysis

Thresholding impact: Multi-label decisions arise from label-wise thresholds on sigmoid scores. Optimizing t_ℓ for F1 yields balanced precision–recall, but safety-critical labels, like threats, may prefer recall-constrained thresholds. We therefore store and document t_ℓ selected on validation and keep them fixed on test and deployment.

Minority-label strategy: To mitigate class imbalance we (i) use inverse-frequency class weights or focal loss ($\gamma \in \{1, 2\}$), (ii) augment data via paraphrasing/back-translation and counterfactual swaps of identity terms, (iii) consider classifier-chain or hierarchical heads to share signal from parent to child labels, and (iv) monitor Macro-F1 (and head/tail Macro-F1) for early stopping rather than Micro-F1 alone. We assess calibration with ECE/MCE, Brier score, and reliability diagrams by label group. When miscalibration is detected, we fit Platt scaling or isotonic regression on validation predictions and then re-apply the fixed thresholds to calibrated scores. Calibration is audited by stratum (topic, length, domain) and re-evaluated periodically to detect shift. We deploy with per-label thresholds and audit logging, track Subset Accuracy and false-positive hot spots by category, mine hard negatives, and schedule periodic threshold retuning as base rates shift.

Micro Versus Macro Performance (Figure 3). Micro-F1 rises from classical models to BiLSTM and peaks with BERT (0.88), reflecting improvements on frequent labels. Macro-F1 increases to 0.79, indicating better minority-label performance; the persistent Micro–Macro gap signals remaining imbalance.

Multi-metric Profile (Figure 4). BERT shows balanced gains across Micro-F1, Macro-F1, Subset Accuracy, and 1–Hamming Loss. Higher Subset Accuracy suggests more coherent sample-wise label sets; higher 1–Hamming Loss (i.e., lower Hamming Loss) indicates fewer label-wise mistakes.

Label-wise Behaviour (Figure 5). BERT consistently outperforms BiLSTM across frequent categories, with the largest margins on *Threats* and *Trolling*, which benefit from contextual modeling. Minority categories (not shown) typically lag. We recommend label-wise threshold tuning, class-weighting or focal loss, and targeted data augmentation. Transformer-based models, especially BERT, outperform traditional machine learning (ML) and Recurrent Neural Network (RNN)-based approaches by leveraging contextual embeddings that capture sarcasm, threats, and implicit abuse (Sahoo et al., 2022; Bell et al., 2024). Remaining challenges include fairness and bias, calibration and thresholding across labels, and computational cost for real-time systems.

7. CONCLUSION

This study developed a multi-label machine learning architecture for detecting trolling and cyberbullying across 81 overlapping abuse categories using a large Kaggle dataset. It demonstrated that fine-tuned BERT models advance multi-label cyberbullying prediction and outperform classical and recurrent models in capturing context-dependent online abuse. BERT achieved the strong performance across all metrics (Micro-F1 = 0.88, Macro-F1 = 0.79, Subset Accuracy = 0.62, Hamming Loss = 0.051).

The model's sigmoid outputs, label-wise thresholds, and calibration improved interpretability and multi-label performance. Class weighting and data augmentation enhanced detection of minority and rare abuse categories. Fairness auditing and visualization techniques helped reduce bias and improve reliability. The open and modular implementation ensures reproducibility and supports real-world moderation tools. Scientifically, the study advances fairness-aware Natural Language Processing for safer cyberspace. It also proposes future extensions in federated learning, explainable AI, bias mitigation, hierarchical modeling, and incorporation of human-in-the-loop (HITL) paradigm.

REFERENCES

1. Badjatiya, P., Gupta, S., Gupta, M. and Varma, V. (2017). Deep Learning for Hate Speech Detection in Tweets. *Proceedings of the 26th International Conference on World Wide Web Companion*. DOI:10.1145/3041021.3054223
2. Bäumler, J., Blöcher, L., Frey, L., Chen, X., Bayer, M. and Reuter, C. (2025). A Survey of Machine Learning Models and Datasets for the Multi-label Classification of Textual Hate Speech in English. *ArXiv, abs/2504.08609*. Available at: https://arxiv.org/abs/2504.08609.
3. Bell, S. J., Meglioli, M. C., Richards, M., Sánchez, E., Ropers, C., Wang, S., ..., Costa-jussà, M. R. (2024). On the Role of Speech Data in Reducing Toxicity Detection Bias. In Proceedings of the 2025 Conference of the Nations of the Americas Chapter of the Association for Computational Linguistics: Human Language Technologies, Albuquerque, New Mexico, 1, 1454-1468. Available at: https://aclanthology.org/2025.naacl-long.67/
4. Chen, Y., Zhou, Y., Zhu, S. and Xu, H. (2012). Detecting offensive language in social media to protect adolescent online safety. *2012 International Conference on Privacy, Security, Risk and Trust and 2012 International Confernece on Social Computing*, Amsterdam, Netherlands, 2012,. 71-80, doi:10.1109/SocialCom-PASSAT.2012.55.

5. Davidson, T., Warmsley, D., Macy, M. and 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). https://doi.org/10.1609/icwsm.v11i1.14955
6. Devlin, J., Chang, M. W., Lee, K., and Toutanova, K. (2018). BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. *North American Chapter of the Association for Computational Linguistics*, 4171–4186. https://doi.org/10.18653/V1/N19-1423.
7. Esackimuthu, S., and Balasundaram, P. (2023). VerbaVisor@Multimodal Hate Speech Event Detection 2023: Hate Speech Detection using Transformer Model. In *Proceedings of the 6th Workshop on Challenges and Applications of Automated Extraction of Socio-political Events from Text*, Varna, Bulgaria. INCOMA Ltd., Shoumen, Bulgaria, 79–83. Available at: https://aclanthology.org/2023.case-1.11/.
8. Fallah, H., Bellot, P., Bruno, E., and Murisasco, E. (2022). Adapting Transformers for Multi-Label Text Classification. CIRCLE (Joint Conference of the Information Retrieval Communities in Europe) 2022, Jul 2022, Samatan, France. ⟨hal-03727927⟩
9. Fortuna, P., and Nunes, S. (2018). A survey on automatic detection of hate speech in text. Association for Computing Machinery (ACM*), 51*(4), 1 – 30. https://doi.org/10.1145/3232676
10. Gutiérrez-Batista, K., Gómez-Sánchez, J. and Fernandez-Basso, C. (2024). Improving automatic cyberbullying detection in social network environments by fine-tuning a pre-trained sentence transformer language model. *Soc. Netw. Anal. Min.* 14, 136 https://doi.org/10.1007/s13278-024-01291-0.
11. Jahan, M. S., and Oussalah, M. (2023). A systematic review of hate speech automatic detection Natural Language Processing. *Neurocomputing*, 546:126232. https://doi.org/10.1016/j.neucom.2023.126232.
12. Kaggle. (2025). *Cyberbullying Dataset (CSV Version)*. Available at: https://www.kaggle.com/datasets (accessed on August 28, 2025).
13. Kaur, S., Singh, S., and Kaushal, S. (2021). Abusive Content Detection in Online User-Generated Data: A survey, *Procedia Computer Science*, 189, 274-281. https://doi.org/10.1016/j.procs.2021.05.098
14. Liu, Y., Ott, M., Goyal, N., Du, J., Joshi, M., Chen, D., Levy, O., Lewis, M., Zettlemoyer, L. and Stoyanov, V. (2019). RoBERTa: A Robustly Optimized BERT Pretraining Approach. *ArXiv, abs/1907.11692*.
15. Lu, G., Liu, Y., Wang, J., and Wu, H., (2023). CNN-BiLSTM-Attention: A multi-label neural classifier for short texts with a small set of labels, *Information Processing & Management*, 60(3). https://doi.org/10.1016/j.ipm.2023.103320
16. Mathew, B., Saha, P., Yimam, S. M., Biemann, C., Goyal, P. and Mukherjee, A. (2021). HateX-plain: A benchmark dataset for explainable hate speech detection. In *Proceedings of the AAAI Conference on Artificial Intelligence* 35(17):14867-14875. https://doi.org/10.1609/aaai.v35i17.17745
17. Nabiilah, G. Z., Prasetyo, S. Y. and Izdihar, Z. N. (2023). BERT base model for toxic comment analysis on Indonesian language (multi-label). *Procedia Computer Science, 216,* 714 - 721. https://doi.org/10.1016/j.procs.2022.12.188
18. Prabhu, R. and Seethalakshmi, V. (2025). A comprehensive framework for multi-modal hate speech detection (multi-label framing). *Scientific Reports*, 15, 13020. Available at: https://doi.org/10.1038/s41598-025-94069-z

19. Raffel, C., Shazeer, N., Roberts, A., Lee, K., Narang, S., Matena, M., Zhou, Y., Li, W., and Liu, P. J. (2020). Exploring the limits of transfer learning with a unified text-to-text transformer. *The Journal of Machine Learning Research (JMLR)*, 21(1), 5485 - 5551.
20. Rajaraman S., Ganesan P. andAntani S (2022). Deep learning model calibration for improving performance in class-imbalanced medical image classification tasks. PLoS ONE 17(1): e0262838. https://doi.org/10.1371/journal.pone.0262838.
21. Sahoo, N., Gupta, H. and Bhattacharyya, P. (2022). Detecting Unintended Social Bias in Toxic Language Datasets. In *Proceedings of the 26th Conference on Computational Natural Language Learning (CoNLL)* Abu Dhabi, United Arab Emirates (Hybrid), 132–143. Available at: https://aclanthology.org/2022.conll-1.10/.
22. Sap, M., Card, D., Gabriel, S., Choi, Y. and Smith, N. A. (2019). The risk of racial bias in hate speech detection. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, Florence, Ital, 1668–1678.
23. Shamatrin, D. (2025). Adaptive Thresholding for Multi-Label Classification via Global- Local Signal Fusion. Available at: https://doi.org/10.48550/arXiv.2505.03118
24. Ullah, N., Schultheis, E. Zhang, J. and Babbar, R. (2025). How Well Calibrated are Extreme Multi-label Classifiers? An Empirical Analysis. In Proceedings of the 31st ACM SIGKDD Conference on Knowledge Discovery and Data Mining V.1 (KDD '25). Association for Computing Machinery, New York, NY, USA, 1397–1408. https://doi.org/10.1145/3690624.3709333
25. Waseem, Z. and Hovy, D. (2016). Hateful symbols or hateful people? Predictive features for hate speech detection on Twitter. In Proceedings of the {NAACL} Student Research Workshop, 88 – 93. doi: 10.18653/v1/N16-2013
26. Yigezu, M. G., Kolesnikova, O., Sidorov, G. and Gelbukh, A. (2023). Transformer-Based Hate Speech Detection for Multi-Class and Multi-Label Classification. In *CEUR Workshop Proc. 3496 (HOMOMEX)*. Available at: https://ceur-ws.org/Vol-3496/homomex-paper5.pdf.
27. Yuan L., Xu X., Sun P., Yu, H. P., Wei, Y. Z. and Zhou, J. J. (2024) Research of multi-label text classification based on label attention and correlation networks. PLoS One, 19(9). Available at: https://journals.plos.org/plosone/article?id=10.1371/journal.pone.0311305.
28. Zampieri, M., Malmasi, S., Nakov, P., Rosenthal, S., Farra, N. and Kumar, R. (2019). Predicting the Type and Target of Offensive Posts in Social Media. DOI:10.18653/v1/N19-1144
29. Zhang, W., Wang, X., Wang, Q., Deng, T. and Wu, X. (2024). From Text Segmentation to Enhanced Representation Learning: A Novel Approach to Multi-Label Classification for Long Texts. In *Findings of the Association for Computational Linguistics: EMNLP 2024*, Miami, Florida, USA 6864–6873. Available at: https://aclanthology.org/2024.findings- emnlp.402.pdf.
30. Zhang, Z., Robinson, D. and Tepper, J. (2018). Detecting hate speech on Twitter using a convolution-GRU based deep neural network. In: Gangemi, A., *et al.* The Semantic Web. ESWC 2018. Lecture Notes in Computer Science(), 10843, 745 – 760. Springer, Cham. https://doi.org/10.1007/978-3-319-93417-4_48