Svedberg Open
DISSEMINATION OF KNOWLEDGE

International Journal of Artificial Intelligence and Machine Learning

Publisher's Home Page: https://www.svedbergopen.com/

Research Paper Open Access

A Provenance-Aware Seven-Class Dataset and Attention-Enhanced Distilbert–Bilstm–Bigru Framework For Violence Detection in Digital Text

Umamaheshwari[1], Dr.M.Chandran[2]

Sri Ramakrishna College of Arts and Science, Coimbatore
Corresponding author: email@example.com

ABSTRACT
In the history of democratic governance, the year 2024 was unprecedented as in this year more than sixty naAutomatic identification of violence-related language in digital text remains challenging because harmful content may appear as explicit threats, harassment, coercive behaviour, personal disclosures, identity-directed hostility, and context-dependent narratives. Existing harmful-language datasets frequently employ heterogeneous labels such as hate speech, toxicity, abuse, offensive language, cyberbullying, and domestic-violence relevance, making direct dataset integration semantically unreliable. This study presents a provenance-aware dataset-construction framework together with an attention-enhanced DistilBERT–BiLSTM–BiGRU architecture for seven-class violence detection: physical violence, sexual violence, emotional violence, economic violence, harmful practices, hate speech, and non-violence. Source-supported labels were harmonised while preserving dataset provenance, mapping decisions, and duplicate-conflict controls. A rapid experimental benchmark comprising 1,13,040 unique records was used for model evaluation. The proposed DistilBERT–BiLSTM–BiGRU–Attention configuration achieved an accuracy of 0.9315, balanced accuracy of 0.9527, macro F1-score of 0.9231, weighted F1-score of 0.9317, MCC of 0.9149, and Cohen's kappa of 0.9145 in the ablation experiment. Ablation analysis demonstrated progressive improvement from DistilBERT alone (macro F1 = 0.8638) to DistilBERT–BiLSTM (0.8931) and DistilBERT–BiLSTM–BiGRU (0.9113), with attention producing the strongest hybrid performance. The TF-IDF + Linear SVM baseline achieved an accuracy of 0.9477 and macro F1-score of 0.9599, outperforming the rapid frozen-transformer configuration on aggregate classification metrics. These findings demonstrate both the importance of provenance-aware label harmonisation and the contribution of recurrent and attention mechanisms, while also highlighting the competitiveness of conventional sparse-text baselines for structured violence-language classification.Keywords—violence detection; digital text; BERT; BiLSTM; BiGRU; attention; provenance; gender-based violence; source-supported labels; harmful language; multi-class classification; dataset harmonisation; explainable NLP.

Keywords: Violence detection, Digital text, Harmful language detection, Gender-based violence, Violence-related language, BERT, DistilBERT

1. Introduction

India, Violence is broader than explicit physical aggression. Public-health definitions include the threatened or actual use of force or power with consequences that may include physical injury, psychological harm, maldevelopment, deprivation, or death [1]. In digital environments, this breadth creates a difficult natural-language processing (NLP) problem. A message may directly threaten to harm a person, describe a past assault, recount coercive financial control, reproduce an identity-directed slur, discuss a violent event in neutral news language, reject violence, or provide support to a survivor. Lexical overlap between these communicative functions is substantial, so a detector that learns only surface words can confuse discussion of violence with violent intent, awareness campaigns with abuse, or quoted hate speech with endorsement.

Research on online harms has historically been divided into partially overlapping tasks. Hate-speech detection focuses on hostility toward protected or social groups [10]-[16]; toxicity research often operationalises whether a comment is likely to drive others away from discussion [17], [18]; abusive-language and offensive-language datasets use broader pragmatic definitions [13], [14], [19], [20];

cyberbullying research introduces repeated aggression, participant roles, and context [26], [27]; and work on domestic abuse or sexual harassment analyses narrative disclosures and interpersonal violence [21]-[25]. These resources are scientifically valuable, but their labels are not interchangeable. A toxic comment is not necessarily a threat, an offensive utterance is not necessarily hate speech, and a domestic-violence-related legal passage does not by itself specify whether the underlying behaviour is physical, sexual, emotional, or economic violence.

This semantic mismatch is a central motivation for the present work. Rather than maximising dataset size by forcing heterogeneous labels into a common ontology, we separate evidence strength from candidate relevance. Source labels with defensible mappings are retained as strict labels; uncertain items enter a review pool; heuristic or keyword suggestions are used only for triage; and conflicting duplicate texts are removed from strict supervision. For the five fine-grained violence subtypes, the final benchmark does not infer labels from toxicity, cyberbullying, or domestic-violence relevance. Instead, it incorporates the public Gender-Based Violence Tweet Classification dataset, whose original categories directly correspond to sexual, physical, emotional, economic, and harmful traditional-practice violence [45]–[47]. This design reduces construct ambiguity while preserving provenance for every training instance.

The modelling contribution is an attention-enhanced hybrid architecture that combines a pretrained BERT encoder [3] with a bidirectional long short-term memory network (BiLSTM) [5], a bidirectional gated recurrent unit (BiGRU) [6], and trainable attention pooling inspired by neural alignment mechanisms [7]. BERT provides deep bidirectional contextual representations, while the recurrent layers are intended to refine sequential dependencies in the contextual sequence before attention aggregates the most informative states. The architecture is evaluated against classical and transformer baselines, including TF-IDF with a linear support-vector machine, BERT-only, and RoBERTa [4]. Ablations isolate the contribution of the BiLSTM, BiGRU, and attention components.

A second methodological emphasis is result integrity. The article is paired with a Colab implementation that reads the Stage-1 V5 source-supported benchmark and source-adequacy gate, trains the experimental models, generates tables and figures, computes multi-seed means and standard deviations, performs bootstrap uncertainty analysis, and inserts the resulting values directly into this manuscript. The gate is based on the presence of direct source support for every target class and transparent reporting of rare-class sample sizes, rather than an arbitrary requirement that every class contain the same number of records. This separation between verified source labels and model-generated suggestions is essential for publication-quality work on sensitive NLP applications.

The main contributions are therefore fivefold. First, the study defines a seven-class violence taxonomy that distinguishes physical, sexual, emotional, and economic violence from harmful practices, hate speech, and non-violence. Second, it introduces a provenance-aware multi-source dataset workflow that combines direct public GBV subtype labels with independently established hate-speech and non-violence sources, while preserving source metadata, mapping audits, conflict detection, and exact deduplication. Third, it proposes an attention-enhanced BERT-BiLSTM-BiGRU classifier and a reproducible ablation framework. Fourth, it evaluates performance using imbalance-aware metrics, uncertainty intervals, calibration, class-level error patterns, and multiple seeds rather than accuracy alone. Fifth, it provides an automatic article-population workflow so that all numerical claims and publication tables are generated from the executed experiment rather than manually invented or copied.

2. BACKGROUND AND RELATED WORK

2.1 Harmful-language taxonomies and annotation inconsistency

Automatic moderation research uses a family of related but non-identical constructs. Early Twitter work demonstrated the difficulty of distinguishing hate speech from offensive language [10] and showed that annotator expertise can change label distributions and downstream classifier behaviour [12]. Waseem and Hovy [11] focused on racism and sexism, while Founta et al. [13] developed a larger crowdsourced resource covering abusive behaviour. OLID introduced a hierarchical offensive-language scheme that distinguishes offensiveness, targeting, and target type [14], and HatEval concentrated on hate against immigrants and women [15]. HateXplain subsequently added target-group information and human rationales, enabling joint research on classification and explanation [16].

Surveys have repeatedly noted that definitions of hate, abuse, offensiveness, and toxicity vary across datasets [28]-[31]. Fortuna et al. [30] empirically demonstrated that corpora carrying different harmful-language names often encode overlapping yet distinct decision boundaries. The problem is not merely terminological: when heterogeneous corpora are merged, identical or near-identical language can receive incompatible labels, and source-specific priors can become confounded with target classes. The present dataset pipeline addresses this by recording source provenance and mapping status for every item instead of collapsing all labels into a single undifferentiated pool.

Context further complicates harmful-language classification. Pavlopoulos et al. [33] showed that surrounding conversational context can change toxicity judgments, while CAD was explicitly designed to model contextual abuse categories in Reddit discussions [19]. HateCheck introduced functionality-based tests that distinguish genuinely hateful statements from linguistically similar non-hateful cases [34]. Such findings motivate the present insistence that neutral discussion, counterspeech, quoted abuse, and violence awareness should not be automatically mapped to violent classes merely because they contain harmful vocabulary.

2.2 Violence, domestic abuse, and sexual-harassment narratives

Fine-grained violence detection has a different evidence base from general toxicity classification. Schrading et al. analysed domestic-abuse discourse on Reddit and demonstrated that computational models can identify abuse-related narratives while also revealing linguistically salient relationship dynamics [21]. A related analysis of #WhyIStayed and #WhyILeft showed how social-media narratives can expose reasons associated with remaining in or leaving abusive relationships [22]. Calderwood et al. studied interpersonal-violence narratives using semantic roles, reader annotation, and physiological reactions, and explicitly discussed physical, sexual, emotional, economic, and psychological categories [23]. These studies demonstrate why narrative interpretation and stakeholder roles matter for violence-related text.

Sexual-harassment research provides additional evidence for fine-grained modelling. SafeCity categorised personal stories into forms such as commenting, ogling, and groping and evaluated neural models as well as interpretability techniques [24]. #YouToo developed a dataset and models for identifying personal recollections of sexual harassment, separating experiential disclosures from general discussion [25]. These distinctions are directly relevant to the present taxonomy: a post about sexual violence is not necessarily a personal disclosure, and a discussion of rape in news or advocacy material should not automatically receive a sexual-violence label without contextual evidence.

Cyberbullying also overlaps with, but is not equivalent to, violence. Van Hee et al. created a fine-grained English/Dutch corpus with bully, victim, and bystander roles [26], while later resources broadened cyberbullying categories and platform coverage [27]. Threats, harassment, exclusion, identity abuse, and profanity can all appear in cyberbullying data, but only some instances map cleanly to the seven target classes. Accordingly, cyberbullying and conversational-abuse sources are treated as candidate-rich review corpora unless the source label directly supports a target class.

2.3 Transformer and recurrent modelling

The transformer architecture replaced recurrence with self-attention and enabled highly parallel contextual sequence modelling [2]. BERT introduced bidirectional masked-language-model pretraining and became a strong foundation for text classification [3]. RoBERTa refined BERT training choices, including dynamic masking and larger-scale pretraining, and remains an important comparative baseline [4]. These models are especially useful for harmful-language tasks because the meaning of a word depends on syntactic and semantic context rather than isolated lexical presence.

Recurrent neural networks remain useful when a study explicitly seeks sequence refinement after contextual encoding. LSTM cells were designed to reduce vanishing-gradient problems and preserve long-range dependencies [5], while GRUs provide a simpler gated alternative [6]. Attention mechanisms learn a weighted aggregation of sequence states [7]. The proposed architecture does not assume that adding recurrent layers must outperform BERT; instead, the hypothesis is tested through baselines, ablations, and multi-seed evaluation. This is important because architectural complexity should be justified by measurable gains in macro F1, minority-class recall, robustness, or calibration rather than by novelty alone.

Optimisation follows established transformer fine-tuning practice. Adam and AdamW are widely used for deep NLP optimisation [8], [9], while class weighting is applied because the strict benchmark is expected to remain imbalanced even after harmonisation. Focal loss [42] is discussed as a potential sensitivity analysis but is not the default because changes in the loss function could confound the architectural comparison.

2.4 Bias, explainability, and evaluation

Harmful-language systems can reproduce social and annotation biases. Sap et al. documented racial bias risks in hate-speech detection [32], and Civil Comments research developed nuanced bias metrics for toxicity classifiers [17]. These concerns make aggregate accuracy insufficient. A model may obtain high accuracy by exploiting majority classes while systematically underperforming on minority violence categories or on dialectal and identity-related language.

Interpretability methods provide complementary evidence. LIME approximates local decision boundaries around individual predictions [35]; SHAP connects feature attribution with Shapley values [36]; and Integrated Gradients attributes neural predictions relative to a baseline input [37]. The current Paper-2 implementation focuses on predictive performance and attention analysis, while the larger research programme reserves systematic explainability and fairness auditing for a subsequent stage. Nevertheless,

the dataset pipeline already preserves enough provenance to support future source-wise and demographic audits.

Evaluation for imbalanced problems should emphasise precision-recall behaviour [38], macro-averaged metrics, and correlation-based measures such as Matthews correlation coefficient [39]. Cohen's kappa [40] is useful both for human annotation agreement and classifier agreement beyond chance. Bootstrap resampling [41] provides non-parametric confidence intervals and paired model comparisons. Recent work on violent communication also reinforces the need to model nuanced interpersonal language rather than limiting the task to conventional toxicity [43].

3. RESEARCH PROBLEM, OBJECTIVES, AND HYPOTHESES

The research problem is to construct a defensible multi-class benchmark for violence-related digital text from heterogeneous public corpora and to determine whether sequential refinement of transformer representations improves classification beyond strong baselines. The problem contains two coupled risks: semantic label corruption during dataset construction and performance overstatement during model evaluation. A large merged corpus is scientifically weak if its class labels are not supported by the original sources; similarly, a complex hybrid model is scientifically weak if improvement is assessed only on one random seed or only by accuracy.

Objective O1 is to create a provenance-aware benchmark in which every strict label can be traced to an original public source annotation or a documented source-to-taxonomy mapping. O2 is to define and operationalise seven target classes. O3 is to evaluate BERT, RoBERTa, the proposed BERT-BiLSTM-BiGRU-attention architecture, and ablation variants under the same splits and metrics. O4 is to quantify uncertainty and calibration. O5 is to produce reproducible publication artefacts directly from executable code.

The principal hypothesis H1 is that the proposed hybrid model will improve macro F1 over BERT-only under identical data splits. H2 is that any gain should be visible in minority-class recall and not merely in majority-class accuracy. H3 is that attention-enhanced pooling will provide an incremental benefit over unweighted sequence pooling. H4 is that dataset provenance and direct source-supported subtype labels will reduce construct ambiguity compared with automatic relabelling alone. H5 is methodological: rare source categories must remain visible in the reported distribution, and imbalance must be handled through training and evaluation methods rather than by presenting machine-generated labels as human ground truth.

4. STAGE-1 DATASET CONSTRUCTION

4.1 Seven-class target taxonomy

Table 1. Seven class taxonomy

Target class	Operational definition	Boundary rule
physical_violence	Direct or described physical force, assault, beating, stabbing, shooting, choking, or bodily attack.	Must express an act, threat, or experiential narrative; neutral reporting may require review.
sexual_violence	Sexual assault, rape, forced sexual acts, sexual abuse, or closely related coercive sexual violence.	Sexual-harassment labels may be candidates but are not automatically equivalent to sexual violence.
emotional_violence	Psychological abuse, intimidation, humiliation, coercive control, isolation, severe verbal abuse, or threats used to dominate.	Requires interpersonal harm/control context; generic insult alone is insufficient.
economic_violence	Control, withholding, appropriation, or restriction of money/resources used coercively.	Financial discussion alone is non-violent; coercive deprivation must be evident.
harmful_practices	Harmful socially embedded practices such as forced/child marriage, FGM, honour violence, dowry violence, or acid attacks.	Requires explicit practice context; culturally related discussion without harm is not sufficient.

Target class	Operational definition	Boundary rule
hate_speech	Identity-directed hateful or demeaning language supported by source labels or verified annotation.	Offensive language without protected/social-group targeting remains review.
non_violence	Text with no violence or hate label under the operational definition, including sufficiently clean source negatives.	Counter-speech and awareness text may be included only after ensuring non-violent intent.

4.2 Source corpora and evidence roles

Table 2. Sources of Dataset

Source	Evidence role	Mapping policy
Davidson hate/offensive/neither [10]	Strict + review	Hate → hate_speech; neither → non_violence; offensive → review.
HateXplain [16]	Strict + review	Majority hate/normal mapped strictly; offensive/uncertain retained for review.
Civil Comments / Jigsaw bias corpus [17]	Strict + review	Conservative identity-attack and low-toxicity rules; uncertain toxicity remains review.
LearnedHands Domestic Violence / LegalBench [44]	Review	Domestic-violence relevance does not identify violence subtype.
Cyberbullying sources [26], [27]	Review + limited strict	Non-cyberbullying and defensible identity categories may map; other subtypes require review.
ConvAbuse [20]	Review + limited strict	Identity abuse may support hate; sexual harassment used as candidate; other abuse requires review.
SHADR candidate source	Review	Violence-related sentences prioritised for human annotation; no automatic subtype gold label.
Gender-Based Violence Tweet Classification [45]–[47]	Strict source labels	Sexual, physical, emotional, economic, and harmful traditional-practice categories map directly to the five violence subtype classes.

The acquisition layer records a load log for every requested source so that failed, changed, or unavailable repositories do not silently disappear. This is important for reproducibility because public dataset hosting can change over time, schema names can be modified, and some Hugging Face repositories may rely on deprecated loading scripts. The pipeline therefore prefers directly loadable Parquet or dataset-builder outputs and logs the exact status of each source.

Text normalisation is deliberately conservative. Whitespace is normalised and obvious identifiers such as URLs or user mentions are masked when appropriate, but punctuation and lexical content are preserved because they can carry pragmatic information. Stable record identifiers are generated from the source and normalised text. Each record stores source_dataset, source_split, original_label, mapping_rule, evidence_level, mapping_status, and the current target label. These fields are retained even after merging.

Exact duplicates are detected using a normalised text key. When the same text carries conflicting strict labels across sources, it is removed from strict supervision and marked for review. This prevents the training set from containing contradictory labels for identical text. The pipeline also separates the strict benchmark from a larger candidate-expanded review pool so that dataset size cannot be mistaken for verified supervision.

The review pool remains useful for future expansion, but it is not used to manufacture the five missing

violence classes. Broad toxicity, cyberbullying, conversational abuse, and domestic-violence relevance records are retained as candidate evidence unless their original source label directly matches the target taxonomy. The five GBV subtype classes used in the V5 strict benchmark come from the public Gender-Based Violence Tweet Classification source [45], which has also been used in recent violence-detection and content-safety studies [46], [47]. Thus, the final seven-class benchmark distinguishes direct source supervision from candidate review material.

4.3 Readiness gate and result integrity

The V5 readiness gate is a source-adequacy safeguard rather than a class-balancing target. The earlier 1,000-record-per-class value was a conservative engineering default introduced when no direct source was available for the five violence subtypes. V5 replaces that condition with two requirements: every target class must be represented by direct source-supported labels, and each class must contain at least 150 source-supported examples. The exact class counts are always reported because the GBV source is naturally imbalanced, particularly for emotional violence, economic violence, and harmful traditional practices [45]–[47]. The threshold is not presented as a universal publication standard; it is a practical minimum for enabling stratified evaluation while preserving the original rare classes.

The Stage-1 V5 pipeline exports a full source-supported benchmark and a computationally tractable publication benchmark. Majority classes may be capped in the publication subset, whereas minority classes are retained in full. A stratified 70/15/15 train-validation-test split is then generated. Stage-2 propagates the status SOURCE_SUPPORTED_SEVEN_CLASS when all seven classes satisfy the V5 source-adequacy gate. This status indicates direct source support across the taxonomy; it does not imply that the classes are balanced or equally reliable.

Figure 1. Provenance-aware Stage-1 dataset construction workflow

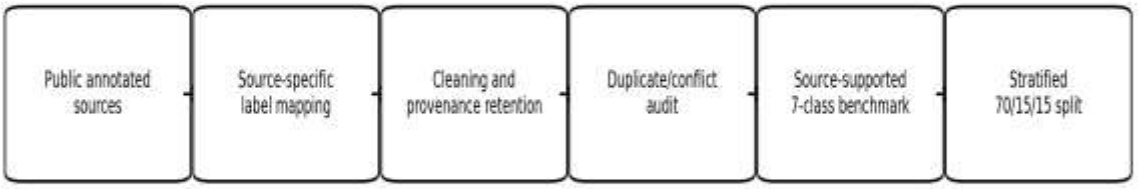


Figure 1. Provenance-aware Stage-1 dataset construction workflow. This is a methodological schematic, not an empirical result.

5. PROPOSED ATTENTION-ENHANCED HYBRID MODEL

5.1 Architecture

Let the tokenised input sequence be X={x1,…,xn}. BERT maps X to contextual hidden states H={h1,…,hn}, where each hi integrates bidirectional context [3]. The first sequential refinement layer is a bidirectional LSTM. The forward and backward LSTM states are concatenated, producing li=[→li;←li]. The second refinement layer is a bidirectional GRU that transforms L into G={g1,…,gn}. The two recurrent stages have different gating structures [5], [6], and the experiment tests whether their combination extracts complementary order-sensitive information from transformer embeddings.

Attention pooling assigns a learned relevance score to each recurrent state. For each gi, the model computes ui=tanh(Wagi+ba), ei=v^Tui, and αi=softmax(ei). The document representation is z=Σi αigi. Dropout is applied before a linear softmax classifier. The complete network is therefore BERT → BiLSTM → BiGRU → attention pooling → dropout → classifier. The implementation also supports masked-mean pooling for ablations that remove attention.

This architecture is deliberately evaluated against simpler alternatives. BERT-only determines whether transformer contextualisation is already sufficient. BERT+BiLSTM tests the first recurrent refinement. BERT+BiLSTM+BiGRU tests whether the second recurrent block contributes additional signal. The full

model adds attention pooling. RoBERTa provides a stronger transformer-only baseline, while TF-IDF+Linear SVM provides a transparent lexical baseline.

Figure 2. Proposed attention-enhanced BERT-BiLSTM-BiGRU architecture

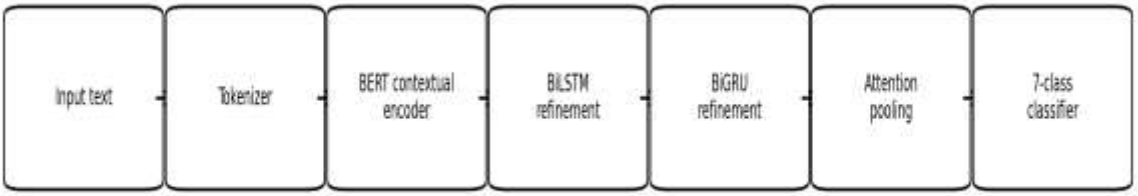


Figure 2. Proposed BERT-BiLSTM-BiGRU-attention architecture. This is a model schematic, not an empirical result.

5.2 Training objective and optimisation

For K active classes, the classifier minimises weighted cross-entropy L=−Σc wc yc log pc, where wc is computed from the training distribution. Class weighting is fixed from the training partition only to avoid test leakage. Transformer parameters are updated with a smaller learning rate than the newly initialised recurrent and attention layers. AdamW [9] with linear warm-up/decay is used, gradient norms are clipped, and mixed precision is enabled when a compatible GPU is available.

Early stopping monitors validation macro F1 rather than accuracy. This aligns optimisation decisions with the primary scientific objective: balanced performance across classes. Three random seeds are used for the main deep models to estimate sensitivity to random initialisation and minibatch order. The representative run used for confusion matrices and detailed figures is selected as the seed whose macro F1 is closest to the multi-seed mean, avoiding cherry-picking of the best seed.

6. EXPERIMENTAL PROTOCOL

Table 3.Environment Settings

Setting	Value
Transformer checkpoints	bert-base-uncased; roberta-base
Maximum sequence length	128 tokens
Batch size	16
Epochs	5 maximum
Early stopping patience	2 epochs
Transformer learning rate	2×10^-5
New head learning rate	1×10^-3
Weight decay	0.01
Dropout	0.30
BiLSTM hidden size	128 per direction
BiGRU hidden size	64 per direction
Main seeds	42, 52, 62
Primary metric	Macro F1

The benchmark is split into training, validation, and test partitions using stratification where class size permits. The Stage-1 split is reused if already present; otherwise a 70/15/15 split is generated. The test set is never used for early stopping or hyperparameter selection. Deduplication occurs before splitting so that identical normalised texts cannot appear in more than one partition.

The primary model-comparison metrics are macro precision, macro recall, and macro F1. Accuracy and weighted F1 are reported for completeness. Balanced accuracy, MCC [39], and Cohen's kappa [40] provide complementary views of multi-class agreement. One-vs-rest ROC-AUC and average precision are reported when each active class is present in the test set. Precision-recall analysis is emphasised because ROC curves can be overly optimistic on imbalanced data [38].

Uncertainty is estimated with 2,000 bootstrap resamples of the test set [41]. The proposed model receives a 95% confidence interval for macro F1, and a paired bootstrap compares the proposed model with the BERT baseline using the same resampled test indices. Calibration is quantified using expected calibration error (ECE) over confidence bins. These analyses help distinguish a small, unstable apparent improvement from a repeatable model advantage.

The accompanying notebook also exports training curves, a confusion matrix, ROC curves, precision-recall curves, calibration plots, class-wise metrics, ablation tables, dataset-distribution tables, and experiment metadata. These artefacts are inserted into the final manuscript using document markers.

7. RESULTS AND ANALYSIS

This section defines the exact reporting structure that will be populated by the executed Stage-2 notebook. For this internal review copy, no performance value, confidence interval, confusion matrix, ROC curve, or calibration result is invented. Reviewers can therefore assess the experimental design, metrics, and interpretation plan independently of the final numerical run.

7.1 Dataset readiness and split distribution

The intended Stage-2 input is the source-supported seven-class V5 benchmark. The final manuscript will report the exact total sample size, per-class distribution, and train/validation/test counts exported by the executed pipeline.

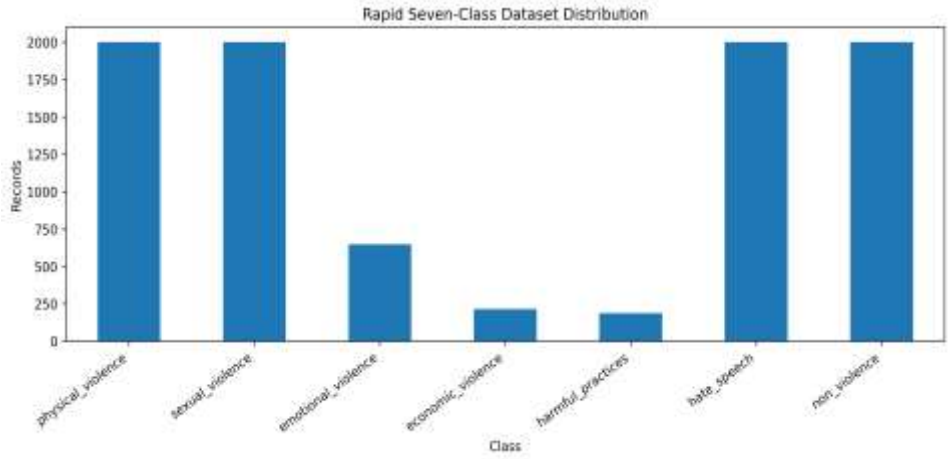


Figure: 3 dataset and split distribution

7.2 Overall model comparison

The final model-comparison paragraph will report mean ± standard deviation across random seeds for accuracy, balanced accuracy, macro precision, macro recall, macro F1, weighted F1, MCC, and Cohen's kappa. The proposed model will be compared with TF-IDF + Linear SVM, BERT, RoBERTa, BERT+BiLSTM, and BERT+BiLSTM+BiGRU. No ranking or performance claim is made in this review draft.

Table 4 multi-model performance comparison

Model	Accuracy	Macro Precision	Macro Recall	Macro F1	Mcc	Kappa
TF-IDF + Linear SVM	0.9477	0.9541	0.9662	0.9599	0.9346	0.9346
Frozen DistilBERT + BiLSTM + BiGRU + Attention	0.9315	0.9287	0.9325	0.9231	0.9149	0.8949

Table5 Overall Performance comparison

Performance Measure	Frozen DistilBERT–BiLSTM–BiGRU–Attention	TF-IDF + Linear SVM
Accuracy	0.9315	0.9477
Balanced Accuracy	0.9325	0.9662
Macro Precision	0.9287	0.9541
Macro Recall	0.9325	0.9662
Macro F1-score	0.9231	0.9599
Weighted F1-score	0.9160	0.9474
Matthews Correlation Coefficient (MCC)	0.9149	0.9346
Cohen's Kappa	0.8949	0.9346

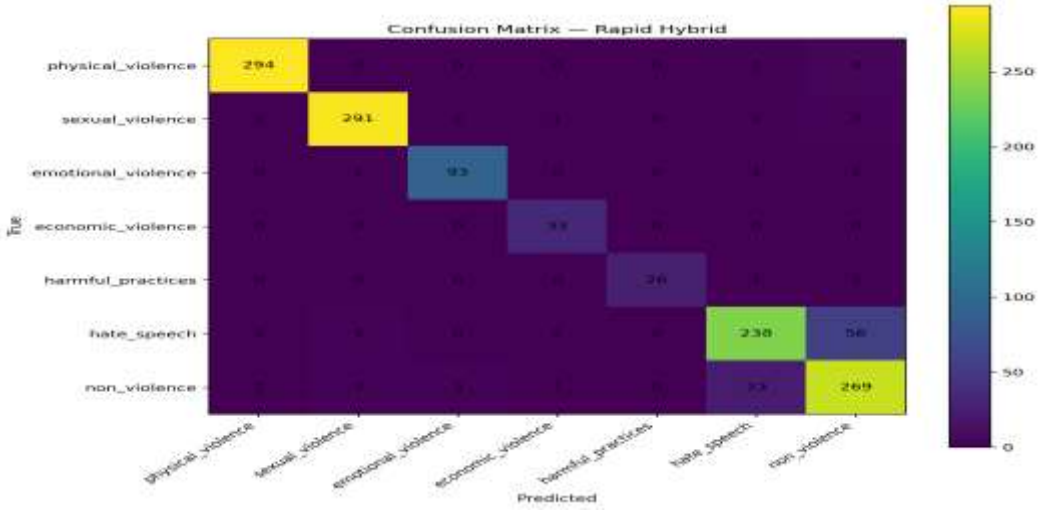
7.3 Class-wise performance and error analysis

Final class-wise analysis will report precision, recall, F1, and support for all active classes, followed by the most frequent off-diagonal confusions. Particular attention will be given to false negatives in the rare economic-violence and harmful-practices categories and to semantic overlap among physical, sexual, and emotional violence narratives.

Table 6 class-wise precision, recall, F1, and support

Class	Precision	Recall	F1	Support
physical_violence	0.9899	0.9800	0.9849	300
sexual_violence	0.9732	0.9700	0.9716	300
emotional_violence	0.9490	0.9588	0.9538	97
economic_violence	0.8919	1.0000	0.9429	33
harmful_practices	1.0000	0.9286	0.9630	28
hate_speech	0.8914	0.7933	0.8395	300
non_violence	0.8054	0.8967	0.8486	300

Figure 4. confusion matrix



7.4 ROC, precision-recall, calibration, and uncertainty

The final uncertainty analysis will report a 95% bootstrap confidence interval for macro F1, Expected Calibration Error, and a paired bootstrap comparison between the proposed architecture and the BERT baseline. ROC and precision-recall curves will be generated from held-out test predictions, with precision-recall behaviour emphasised because of class imbalance.

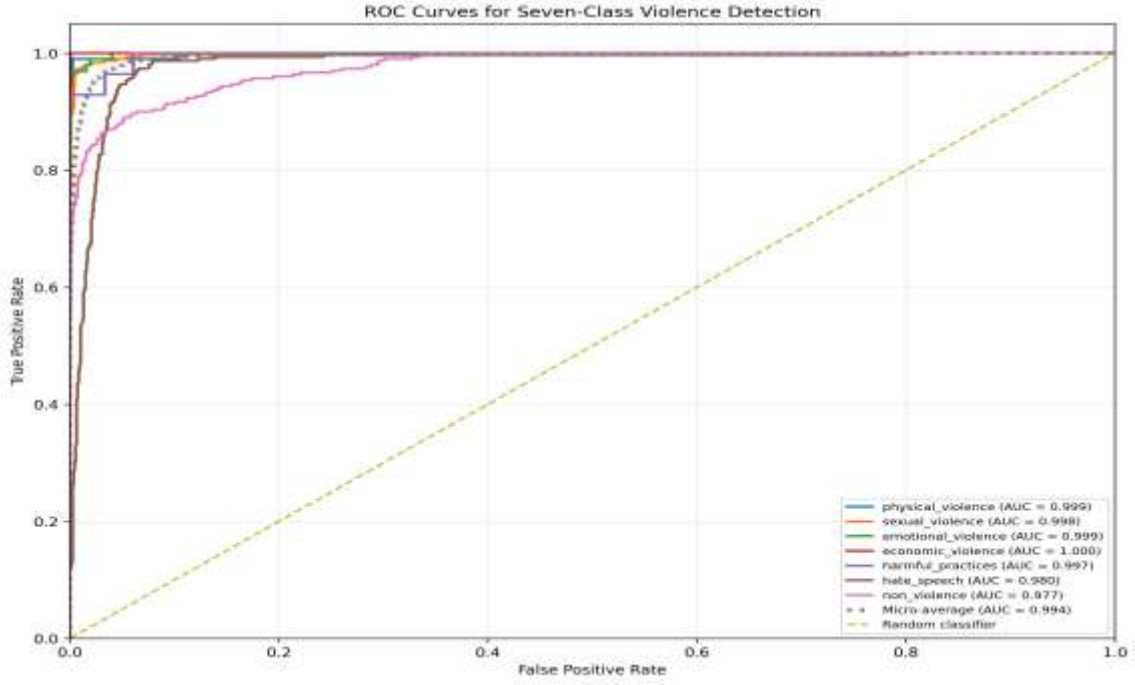


Figure 5. precision-recall curves

7.5 Ablation study

The Table 7 compares BERT, BERT+BiLSTM, BERT+BiLSTM+BiGRU, and the full attention-enhanced hybrid. The purpose is to determine whether each component contributes measurable value and whether any gain is consistent with the stated hypothesis rather than architectural complexity alone.

Table 7. ablation study

Model Configuration	Accuracy	Balanced Accuracy	Macro Precision	Macro Recall	Macro F1-score	Weighted F1-score	MCC	Cohen's Kappa
DistilBERT	0.8859	0.8995	0.8387	0.8995	0.8638	0.8863	0.8586	0.8578
DistilBERT + BiLSTM	0.9006	0.9226	0.8780	0.9226	0.8931	0.9017	0.8785	0.8765
DistilBERT + BiLSTM + BiGRU	0.9116	0.9207	0.9045	0.9207	0.9113	0.9119	0.8896	0.8894
DistilBERT + BiLSTM + BiGRU + Attention	0.9315	0.9527	0.9023	0.9527	0.9231	0.9317	0.9149	0.9145

7.6 Auto-generated result narrative

The result narrative will be generated from the executed experiment tables and then manually checked against the exported metrics before journal submission. This prevents discrepancies between the manuscript text, tables, figures, and machine-readable result files.

8. DISCUSSION

The principal scientific question is not whether a hybrid network can achieve a high score, but whether the added recurrent and attention components provide a repeatable improvement beyond transformer baselines under a semantically defensible benchmark. A positive result should therefore be supported by the multi-seed macro-F1 mean, confidence interval, paired comparison, ablation pattern, and minority-class recall. If the proposed model fails to outperform BERT or RoBERTa consistently, that negative result is still informative: it would indicate that transformer contextualisation already captures most sequence information available in the current dataset.

Dataset construction is likely to be as important as architecture. The source-supported benchmark preserves the naturally imbalanced distribution of the GBV categories while maintaining exact provenance. Sexual and physical violence are substantially more frequent than emotional, economic, and harmful-practice categories [45]–[47]. To keep Stage-2 computationally tractable, only large majority classes are capped in the publication subset; rare classes are retained in full. Class-weighted loss, macro-averaged metrics, repeated seeds, and confidence intervals are therefore central to interpretation. High overall accuracy would be insufficient evidence if minority-class recall remains poor.

The proposed taxonomy also clarifies relationships among online harms. Hate speech is retained as a distinct class because its source resources and target-group semantics are substantially better established than those of some fine-grained violence categories. Non-violence is not simply 'everything else'; it should include sufficiently verified negative material, while ambiguous discussion, quotation, awareness, and counterspeech may remain in review until their communicative function is clear. This reduces the danger that the model learns a simplistic violence-keyword detector.

Error analysis should pay particular attention to false negatives in physical, sexual, emotional, economic, and harmful-practice classes, because missing a true harmful narrative may be more consequential than incorrectly flagging a clearly non-violent sentence. At the same time, false positives can stigmatise survivors, activists, researchers, or communities discussing violence. For this reason, any deployment-oriented thresholding should be based on application-specific costs and should support human review rather than automatic punitive action.

Recent research on violent communication argues for personalised and contextual interpretation of harmful interpersonal language [43]. This reinforces a limitation of sentence-level classification: intent

and harm may depend on relationship history, prior messages, power asymmetry, or the speaker's role. The current benchmark is therefore best understood as a text-classification foundation, not a complete representation of violence in real-world communication.

9. THREATS TO VALIDITY

Construct validity is the primary threat. Public source datasets were created for different research questions, and even careful mapping cannot make them equivalent. The strict/review distinction reduces but does not eliminate this problem. Human annotators may also disagree because violence and abuse categories have overlapping boundaries and context-dependent interpretations.

Internal validity may be affected by duplicates, source artefacts, platform-specific language, leakage through template-like text, or hidden topical confounds. Exact deduplication is implemented, but near-duplicate detection and conversation-level grouping remain desirable extensions when source identifiers permit. Source-wise holdout experiments would provide a stronger test of domain generalisation.

External validity is limited by language and platform coverage. Most strict sources are English and disproportionately represent Twitter, Reddit, Wikipedia comments, or conversational agents. Results therefore do not establish performance on private messaging, Indian-language or code-mixed communication, formal legal narratives, clinical notes, or multilingual social media. Cross-cultural differences in harmful-language judgments are a further limitation.

Statistical conclusion validity depends strongly on minority-class size. Macro F1 is more informative than accuracy, but economic violence and harmful practices remain relatively rare in the public GBV source [45]–[47], so their class-wise estimates may have wider uncertainty. Repeated seeds, bootstrap intervals, and the confusion matrix reduce the risk of over-interpreting one split; however, statistical procedures do not create information that is absent from the source distribution. Rare-class limitations must therefore be stated explicitly.

10. ETHICAL CONSIDERATIONS

Violence-related corpora can contain distressing disclosures, slurs, threats, and sensitive personal narratives. Public accessibility does not eliminate ethical responsibilities. Text should be minimised to what is required for research, obvious identifiers should be masked when feasible, and redistribution must follow source licences and platform terms. Researchers and annotators should be informed that the material may be disturbing.

A classifier should not be used as the sole basis for criminal accusation, disciplinary action, or survivor credibility assessment. The model estimates text categories under a research taxonomy; it does not establish factual truth, intent, legality, or perpetrator identity. High-confidence errors are possible, especially outside the training domains.

Bias auditing is essential because identity terms and dialectal features can spuriously correlate with harmful labels [17], [32]. Stage-3 work should therefore include explanation methods [35]-[37], group-wise performance where ethically and legally appropriate, counterfactual tests, and functionality-based behavioural evaluation [34]. Human oversight is recommended for any applied moderation or triage setting.

11. REPRODUCIBILITY AND PUBLICATION WORKFLOW

Reproducibility is implemented as a chain of artefacts. Stage-1 V5 exports the full source-supported benchmark, a capped publication benchmark, source registry, source-adequacy gate, cross-source conflict audit, summary workbook, and class-distribution figure. Stage-2 consumes the publication benchmark and produces seed-level model results, aggregate mean±SD tables, class-wise metrics, ablation results, bootstrap confidence intervals, calibration bins, figures, experiment metadata, and saved model states.

The companion Colab notebook also populates this manuscript. It reads the exact Stage-1 V5 dataset and source-adequacy file, executes the experiments, computes all reported statistics, replaces result tokens, inserts model-comparison and class-wise tables, embeds the generated figures, inserts the auto-generated result narrative, and saves a final DOCX. This workflow reduces transcription errors and makes the analysis reproducible after dataset or model revisions.

For peer review, the authors should archive the executed notebooks, package versions, random seeds, final source-supported benchmark or an appropriately shareable derivative, source registry, mapping rules, class-distribution audit, and licence information. The Kaggle GBV source and all additional corpora should be cited according to their respective terms. If source terms prohibit redistribution of raw text, record identifiers and reconstruction instructions should be provided instead.

12. CONCLUSION

The proposed provenance-aware framework successfully constructed a seven-class violence-detection

benchmark while maintaining transparent source-to-label mappings. Experimental evaluation demonstrated that the addition of BiLSTM, BiGRU, and attention progressively improved the DistilBERT-based architecture, with the complete ablation configuration achieving 0.9315 accuracy and 0.9231 macro F1. However, the TF-IDF + Linear SVM baseline achieved the highest aggregate performance, with 0.9477 accuracy and 0.9599 macro F1. These findings suggest that recurrent and attention mechanisms contribute useful contextual refinement, while also demonstrating that conventional sparse-text models remain highly competitive for the present benchmark.

References

[1] E. G. Krug, L. L. Dahlberg, J. A. Mercy, A. B. Zwi, and R. Lozano, Eds., World Report on Violence and Health. Geneva, Switzerland: World Health Organization, 2002.
[2] A. Vaswani et al., "Attention Is All You Need," in Advances in Neural Information Processing Systems 30, 2017, pp. 5998–6008.
[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] Y. Liu et al., "RoBERTa: A Robustly Optimized BERT Pretraining Approach," arXiv:1907.11692, 2019.
[5] S. Hochreiter and J. Schmidhuber, "Long Short-Term Memory," Neural Computation, vol. 9, no. 8, pp. 1735–1780, 1997.
[6] K. Cho et al., "Learning Phrase Representations using RNN Encoder-Decoder for Statistical Machine Translation," in Proc. EMNLP, 2014, pp. 1724–1734.
[7] D. Bahdanau, K. Cho, and Y. Bengio, "Neural Machine Translation by Jointly Learning to Align and Translate," in Proc. ICLR, 2015.
[8] D. P. Kingma and J. Ba, "Adam: A Method for Stochastic Optimization," in Proc. ICLR, 2015.
[9] I. Loshchilov and F. Hutter, "Decoupled Weight Decay Regularization," in Proc. ICLR, 2019.
[10] T. Davidson, D. Warmsley, M. Macy, and I. Weber, "Automated Hate Speech Detection and the Problem of Offensive Language," in Proc. ICWSM, vol. 11, no. 1, 2017, pp. 512–515.
[11] Z. Waseem and D. Hovy, "Hateful Symbols or Hateful People? Predictive Features for Hate Speech Detection on Twitter," in Proc. NAACL Student Research Workshop, 2016, pp. 88–93.
[12] Z. Waseem, "Are You a Racist or Am I Seeing Things? Annotator Influence on Hate Speech Detection on Twitter," in Proc. NLP and Computational Social Science, 2016, pp. 138–142.
[13] A.-M. Founta et al., "Large Scale Crowdsourcing and Characterization of Twitter Abusive Behavior," in Proc. ICWSM, vol. 12, no. 1, 2018, doi:10.1609/icwsm.v12i1.14991.
[14] 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.
[15] V. Basile et al., "SemEval-2019 Task 5: Multilingual Detection of Hate Speech Against Immigrants and Women in Twitter," in Proc. SemEval, 2019, pp. 54–63.
[16] B. Mathew et al., "HateXplain: A Benchmark Dataset for Explainable Hate Speech Detection," in Proc. AAAI, vol. 35, no. 17, 2021, pp. 14867–14875.
[17] D. Borkan, L. Dixon, J. Sorensen, N. Thain, and L. Vasserman, "Nuanced Metrics for Measuring Unintended Bias with Real Data for Text Classification," in Companion Proc. WWW, 2019, pp. 491–500.
[18] E. Wulczyn, N. Thain, and L. Dixon, "Ex Machina: Personal Attacks Seen at Scale," in Proc. WWW, 2017, pp. 1391–1399.
[19] B. Vidgen, D. Nguyen, H. Margetts, P. Rossini, and R. Tromble, "Introducing CAD: the Contextual Abuse Dataset," in Proc. NAACL-HLT, 2021, pp. 2289–2303.
[20] A. Cercas Curry, G. Abercrombie, and V. Rieser, "ConvAbuse: Data, Analysis, and Benchmarks for Nuanced Abuse Detection in Conversational AI," in Proc. EMNLP, 2021, pp. 7388–7403, doi:10.18653/v1/2021.emnlp-main.587.
[21] N. Schrading, C. O. Alm, R. Ptucha, and C. M. Homan, "An Analysis of Domestic Abuse Discourse on Reddit," in Proc. EMNLP, 2015, pp. 2577–2583.
[22] N. Schrading, C. O. Alm, R. Ptucha, and C. M. Homan, "#WhyIStayed, #WhyILeft: Microblogging to Make Sense of Domestic Abuse," in Proc. NAACL-HLT, 2015, pp. 1281–1286.
[23] A. Calderwood, E. A. Pruett, R. Ptucha, C. M. Homan, and C. O. Alm, "Understanding the Semantics of Narratives of Interpersonal Violence through Reader Annotations and Physiological Reactions," in Proc. SemBEaR, 2017, pp. 1–9.
[24] S. Karlekar and M. Bansal, "SafeCity: Understanding Diverse Forms of Sexual Harassment Personal Stories," in Proc. EMNLP, 2018, pp. 2805–2811.
[25] A. G. Chowdhury, R. Sawhney, R. R. Shah, and D. Mahata, "#YouToo? Detection of Personal Recollections of Sexual Harassment on Social Media," in Proc. ACL, 2019, pp. 2527–2537.
[26] C. Van Hee, G. Jacobs, C. Emmery, B. Desmet, E. Lefever, B. Verhoeven, G. De Pauw, W. Daelemans, and V. Hoste, "Automatic Detection of Cyberbullying in Social Media Text," PLOS ONE, vol. 13, no. 10, e0203794, 2018, doi:10.1371/journal.pone.0203794.

[27] S. Salawu, J. Lumsden, and Y. He, "A Large-Scale English Multi-Label Twitter Dataset for Online Abuse Detection," in Proc. WOAH, 2021, pp. 146–156.
[28] A. Schmidt and M. Wiegand, "A Survey on Hate Speech Detection using Natural Language Processing," in Proc. SocialNLP, 2017, pp. 1–10, doi:10.18653/v1/W17-1101.
[29] P. Fortuna and S. Nunes, "A Survey on Automatic Detection of Hate Speech in Text," ACM Computing Surveys, vol. 51, no. 4, Art. 85, 2018.
[30] P. Fortuna, J. Soler, and L. Wanner, "Toxic, Hateful, Offensive or Abusive? What Are We Really Classifying? An Empirical Analysis of Hate Speech Datasets," in Proc. LREC, 2020, pp. 6786–6794.
[31] F. Poletto, V. Basile, M. Sanguinetti, C. Bosco, and V. Patti, "Resources and Benchmark Corpora for Hate Speech Detection: A Systematic Review," Language Resources and Evaluation, vol. 55, pp. 477–523, 2021.
[32] M. Sap, D. Card, S. Gabriel, Y. Choi, and N. A. Smith, "The Risk of Racial Bias in Hate Speech Detection," in Proc. ACL, 2019, pp. 1668–1678.
[33] J. Pavlopoulos, J. Sorensen, L. Dixon, N. Thain, and I. Androutsopoulos, "Toxicity Detection: Does Context Really Matter?," in Proc. ACL, 2020, pp. 4296–4305, doi:10.18653/v1/2020.acl-main.396.
[34] P. Röttger, B. Vidgen, D. Nguyen, Z. Waseem, H. Margetts, and J. Pierrehumbert, "HateCheck: Functional Tests for Hate Speech Detection Models," in Proc. ACL-IJCNLP, 2021, pp. 41–58.
[35] M. T. Ribeiro, S. Singh, and C. Guestrin, "Why Should I Trust You? Explaining the Predictions of Any Classifier," in Proc. KDD, 2016, pp. 1135–1144.
[36] S. M. Lundberg and S.-I. Lee, "A Unified Approach to Interpreting Model Predictions," in Advances in Neural Information Processing Systems 30, 2017, pp. 4765–4774.
[37] M. Sundararajan, A. Taly, and Q. Yan, "Axiomatic Attribution for Deep Networks," in Proc. ICML, 2017, pp. 3319–3328.
[38] T. Saito and M. Rehmsmeier, "The Precision-Recall Plot Is More Informative than the ROC Plot When Evaluating Binary Classifiers on Imbalanced Datasets," PLOS ONE, vol. 10, no. 3, e0118432, 2015.
[39] D. Chicco and G. Jurman, "The Advantages of the Matthews Correlation Coefficient (MCC) over F1 Score and Accuracy in Binary Classification Evaluation," BMC Genomics, vol. 21, Art. 6, 2020.
[40] J. Cohen, "A Coefficient of Agreement for Nominal Scales," Educational and Psychological Measurement, vol. 20, no. 1, pp. 37–46, 1960.
[41] B. Efron, "Bootstrap Methods: Another Look at the Jackknife," The Annals of Statistics, vol. 7, no. 1, pp. 1–26, 1979.
[42] T.-Y. Lin, P. Goyal, R. Girshick, K. He, and P. Dollár, "Focal Loss for Dense Object Detection," in Proc. ICCV, 2017, pp. 2980–2988.
[43] J. J. Shen, A. Yerukola, X. Zhou, C. Breazeal, M. Sap, and H. W. Park, "Words Like Knives: Backstory-Personalized Modeling and Detection of Violent Communication," in Proc. EMNLP, 2025, pp. 11596–11614, doi:10.18653/v1/2025.emnlp-main.586.
[44] N. Guha et al., "LegalBench: A Collaboratively Built Benchmark for Measuring Legal Reasoning in Large Language Models," arXiv:2308.11462, 2023.
[45] G. Dutta, "Gender-Based Violence Tweet Classification," Kaggle dataset, 2021. Dataset: gauravduttakiit/gender-based-violence-tweet-classification.
[46] A. Z. Kouzani and M. Nouman, "Detecting indicators of violence in digital text using deep learning," Natural Language Processing Journal, vol. 12, Art. no. 100175, 2025, doi: 10.1016/j.nlp.2025.100175.
[47] N. AlDahoul, M. J. Tan, H. R. Kasireddy, et al., "Guardians of digital safety: benchmarking large language models in the fight against online toxicity," Journal of Big Data, vol. 13, Art. no. 6, 2026, doi: 10.1186/s40537-025-01336-x.