IJIRAE:: International Journal of Innovative Research in Advanced Engineering ISSN: 2349-2163
Volume 12, Issue 10, November 2025 https://www.ijirae.com/archives
https://doi.org/10.26562/ijirae.2025.v1210.13

A Chained Deep Learning Model for Fine-Grained Cyberbullying Detection with Bystander Dynamics

T K Pradeepkumar
Assistant Professor, Department of CSE
AMC Engineering College, Bengaluru, India
tkpradeep.it@gmail.com
https://orcid.org/0000-0001-8396-4283
Abhishek B, Abhishek K, Akhilesh AM, Achyuth Kumar GV
Department of CSE,
AMC Engineering College, Bengaluru, India
abhishekabhi09988@gmail.com, abhij4017@gmail.com
akhileshamakhil@gmail.com, achyuthgvmrb@gmail.com

Publication History
Manuscript Reference No: IJIRAE/RS/Vol.12/Issue10/OCAE10095
Research Article| Open Access | Double-Blind Peer-Reviewed | Article ID: IJIRAE/RS/Vol.12/Issue10/OCAE10095
Received:22,October 2025,Revised:28,October 2025, Accepted:31, October 2025, Published Online: 21, November 2025.
https://www.ijirae.com/volumes/Vol12/iss-10/13.OCAE10095.pdf
Citation: Pradeep,Abhishek,Abhishek,Akhilesh,Achyuth(2025), A Chained Deep Learning Model for Fine-Grained Cyber bullying Detection with Bystander Dynamics,IJIRAE: International Journal of Innovative Research in Advanced Engineering, Volume 12, Issue 10 of 2025 pages 424-428
doi:>https://doi.org/10.26562/ijirae.2025.v1210.13
BibTeX Key: Pradeep@2025Chained
IJIRAE papers should be cited as IJIRAE (International Journal of Innovative Research in Advanced Engineering, AM Publications, India 2025, ISSN 2349-2163, https://doi.org/10.26562/ijirae.2025.v1210.13 The journal's official abbreviation is IJIRAE. Orcid: https://orcid.org/0009-0004-9398-7488
Copyright©2025 copyright by the authors. This article is an open access article distributed under the terms and conditions of the Creative Commons Attribution (CC BY) license (https://creativecommons.org/licenses/by/4.0/).

Abstract: The complexity, concealment, and context-dependence of online harassment occurring through social networking sites have made it a growing concern. Conventional detection systems frequently ignore bystander participation and with respect to online conversations in favour of treating digital abuse as a straight forward binary problem. This paper introduces a chained neural architecture that combines a bidirectional transformer encoder known as BERT along with the Long Short-Term Memory (LSTM) network to detect fine-grained levels of cyber bullying with the incorporation of by standard dynamics. The model first identifies the part played by standers defenders, instigators, neutrals, or others and then uses this output to predict the aggression level of a cyberbullying incident. Using the publicly available CYBY23 dataset, the introduced model yields an F1-score of 89%, improving detection performance by25.35% compared to traditional binary classifiers. The results show that considering bystander involvement improves context understanding and reduces misclassification between aggression and bullying.
Keywords: Cyber bullying Detection, BERT, LSTM, By stander Roles, Deep Learning, Multi-Label Classification

I. INTRODUCTION

With the rapid expansion of social media, cyber bullying has emerged as a serious and ongoing online problem. It involves repeated and intentional digital attacks meant to hurt others who mayhave a hard time defending themselves. This issue extends beyond technical challenges to include effects on people's mental health, often causing stress, depression, or social withdrawal. Most current cyber bullying detection models mainly look at single posts or comments by themselves. But real cyberbullying happens within conversational threads that include multiple people. Some of these people are by standers who can act as defenders, instigators, or just observers. These social interactions have a big impact on how bullying spreads or stops. So, to detect bullying accurately, it's important to consider both how strong the aggression is and how by standers behaves. This helps create more realistic and context-aware detection methods. To address these short comings, the present work suggests a chained deep neural system that integrates BERT for understanding context and LSTM for analysing sequences. The model examines both the meaning of the text and how users interact to categorise cyber bullying as high aggression, low aggression, or non aggression. This method improves accuracy, makes results easier to understand, and enablesreal-time issue detection. It works better than older models that ignore the social and conversational context. To overcome these limitations, this work proposes a chained deep neural framework that brings together a bidirectional transformer encoder known as BERT for contextual generating feature representations and LSTM (Long Short-Term Memory) for sequential behaviour modelling. By including bystander roles and emotional cues as extra features, the proposed system captures both the meaning and social aspects of online aggression. This approach not only boosts detection accuracy but also provides clearer insight into how user interactions affect the spread or reduction of cyber bullying.

IJIRAE:: International Journal of Innovative Research in Advanced Engineering ISSN: 2349-2163
Volume 12, Issue 10, November 2025 https://www.ijirae.com/archives
https://doi.org/10.26562/ijirae.2025.v1210.13

The model's results are tested on the CYBY23 dataset, showing clear improvements in precision, recall, and F1-score compared to basic and older models.

II. RELATED WORK

The initial research on cyber bullying detection utilized machine-learning approaches that included the SVM algorithm, the Naïve Bayes approach, and TF-IDF-based text analysis. The methods based on keywords proved successful for detection, but they struggled to recognise the complete meaning and semantic relations among words in dialogues. The advent of deep-learning methods and transformer-based architectures through a bidirectional transformer model known as BERT brought a break through to this field because models gained the capability to process language in both directions. Recent studies have introduced novel approaches that use extended versions of previous techniques to build multi-class cyberbullying detection systems that analyse sentiment, emotions, and conversational context for better results. Research studies have ignored bystander behaviour because they concentrated their analysis on the interactions between bullies and their victims. Research shows that by stander actions have a major impact on how online aggression spreads, so scientists need to study this phenomenon further. The HSAN and TG Bully models achieved better accuracy results, but theydo not include mechanisms to link participant roles with their aggression levels. The current research gap drives our development to chained classification system, which uses BERT for semantic role identification and LSTM for detailed bullying sequence analysis.

COMPARISON TABLE

Table1: Comparative Summary of Major Cyber bullying Detection Studies

Field	Haifa Saleh Alfurayj et al. (2024)	Ajanthaa Lakkshmanan et al. (2024)	Wan Noor HamizaWan Ali et al.(2018)	J.Sathya & F.MaryHarin Fernandez(2024)	Wachiraporn Tapaopong et al. (2024)
Title	A Chained Deep Learning Model For Fine-Grained Cyber bullying Detection with By stander Dynamics	Emotion-Driven Cyber bullying Detection: A Hybrid CNN-LSTM and BERT Approach	Cyber bullying Detection: An Overview	Effective Automatic Cyber bullying Detection Using a Hybrid SVM+NLP Approach	Enhancing Cyber bullying Detection on Social Media Using Transformer Models
Year	2024	2024	2018	2024	2024
Dataset Used	CYBY23(by stander-Labeled tweets)	Multilingual Sentiment Dataset	Public social Media datasets	Twitter & forum Text datasets	5-class labeled Social media dataset
Techniques / Algorithms	BERT+LSTM (Chained Classifier)	CNN-LSTM+BERT+ GloVe/FastText	SVM, Naïve Bayes, JRip, J48	SVM+NLP(TF-IDF, LIWC2,WordNet)	BERT, RoBERTa, ALBERT, DistilBERT (PyTorch fine-tuned)
Model Type	Transformer +Sequential Hybrid	Hybrid Deep Learning (Emotion-based)	Machine Learning (Classical)	HybridML+NLP	Transformer-basedMulti-class Classification
Key Features Used	Role based conversation context, aggression level	Emotion and sentiment embeddings	Content, sentiment, user	Lexical+ Psychological text features	Textual semantics and context
Accuracy / F1-Score	89% F1-Score	0.97F1-Score	High accuracy with SVM (~90%)	Accuracy 93.15%, Precision 93%, Recall 94.5%	Accuracy 94.36%, Precision 93.83%, Recall 93.91%
Strengths	Incorporates By stander roles; Multistage context understanding	Emotion-driven and multilingual; robust hybrid design	Broad survey; Foundational For future models	Strong performance using simple ML	Lightweight transformer; high efficiency
Limitations	Limited to CYBY23; cascading chain error	Text-only model; lacks multimodal input	Lacks deep Context and sarcasm detection	Not real-time; no multimedia integration	Limited interpretability; Context not deeply modelled
Gaps Identified	Fixed chain dependency; needs adaptive chaining	Requires multimodal data (image+text)	Needs deep Learning integration	Needs real time implementation	Requires better Explain ability and adaptability
Key Insight	Role-aware chaining Enhances fine-grained detection	Emotions improve detection precision	Traditional ML is less effective for context	Hybrid NLP is still relevant for simpler tasks	Transformers lead but need context integration

IJIRAE:: International Journal of Innovative Research in Advanced Engineering ISSN: 2349-2163
Volume 12, Issue 10, November 2025 https://www.ijirae.com/archives
https://doi.org/10.26562/ijirae.2025.v1210.13

The reviewed studies show a clear transition away from conventional machine-learning methods like SVM and Naïve Bayes to deep learning and transformer-based models such as CNN, LSTM, and BERT. As seen in Table 1, hybrid and transformer models achieved higher accuracy and F1-scores compared to classical approaches. However, the majority of studies concentrated solely on bully–victim interactions, ignoring the social influence of bystanders. The work by Haifa Saleh Alfurayj et al.(2024) introduced a chained BERT–LSTM model that included by stander roles, improving detection accuracy by 25%. Despite these improvements, challenges remain in handling multimodal data, real-time adaptability, and contextual understanding motivating the proposed fine-grained, role-aware detection model.

III. METHODOLOGY

The proposed system employs a two-layer chained architecture designed to exploit the correlation between by stander roles and bullying severity.

A. Dataset Description: We used the CYBY23 dataset, a Twitter-based corpus comprising 639 annotated tweets across 150 threads. Each tweet includes two labels

- By stander Role: Instigator, Defender, Neutral, or Other
- Fine-Grained Bullying Label: Low aggression, High aggression, or Aggression without Bullying

Data was collected using the Twitter API, filtered by hate-related keywords and hash tags (e.g., racism, immigrant, Nazi, slurs). Each thread was manually annotated by multiple domain experts, achieving a Fleiss'Kappa score of 0.60, indicating substantial inter annotator agreement. The dataset captures real conversational flow, enabling role-based Before model training, the dataset underwent text normalization, including cleaning the text by excluding URLs, mentions, and emojis, along with tokenization and lemmatization. Stop words were filtered, with hash tags also being retained as they often convey emotion and intent. The processed dataset was divided into 70 % training, 15 % validation, and 15 % testing subsets to ensure unbiased evaluation. Ethical compliance was maintained by anonymizing all usernames and sensitive content. Aligning with Twitter's data usage policy. context analysis instead of isolated message classification.

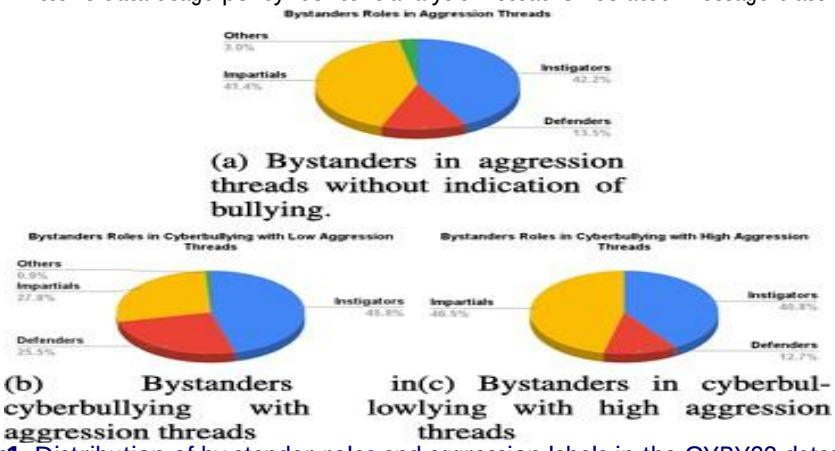

Fig1. Distribution of by stander roles and aggression labels in the CYBY23 dataset

As illustrated in Fig. 2, the role distribution varies notably across aggression levels. In threads without explicit bullying, Instigators represent approximately 42% of participants, while Defenders and Impartials form smaller portions. In low-aggression threads, the percentage of Defenders increases, indicating stronger victim support. Conversely, in high-aggression threads, Instigators dominate nearly 49% of the interactions, while Defenders decline to around 12%, suggesting that the intensity of aggression reduces defensive intervention. These distributions confirm that bystander behaviour is context-dependent and significantly influences the overall tone of online discussions.

B. Proposed Chained Architecture

The two-stage pipeline integrates a BERT model followed by an LSTM model

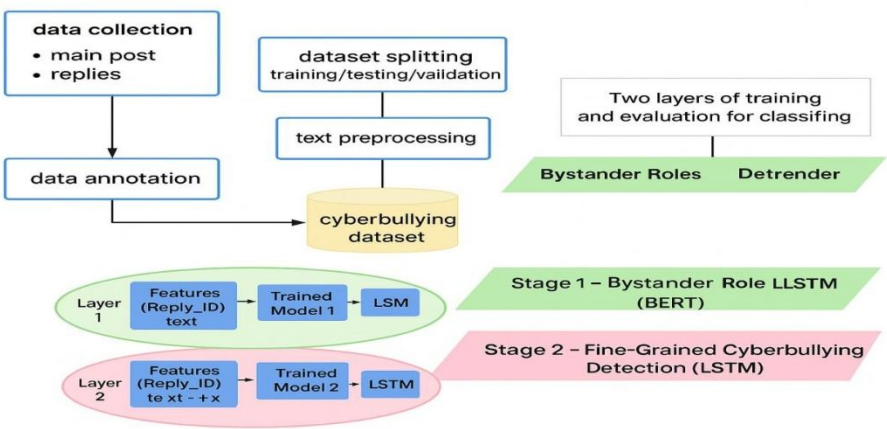

Fig2. Flowchart of the chained fine-grained cyber bullying detection model combining BERT and LSTM

If you look at the Fig.1, you can see the whole set up, which is like a chain, and it has two parts, one after the other, bringing together the BERT and LSTM models.

Stage 1– By stander Role Identification (BERT):

First, we have the part where we figure out whatrole people play, and we use BERT for this. In this first step, the BERT neural encoder, known as Bidirectional Encoder Representations from Transformers, gets adjusted. It uses the clean tweet messages and extra information about replies, so it can determine what each person in a chat is doing. The system then knows if someone is starting trouble, defending someone, staying neutral, or just not involved at all. The BERT model is picked because it can understand words based on the words around them, and this lets it catch even small differences in how people sound or what they mean. Each tweet is turned into a 768-number code using BERT's encoder, and then a softmax layer is used to sort out how likely each role is. These output numbers show how probable it is that a person's message fits into each role group. You can receive a clear picture from this. To make things work better and more steadily, the model is trained using something called the Adam optimiser, at a learning rate of 2e-5. It also includes something called batch normalisation, which helps control how the information flows. This makes the model adapt better to all the different ways people write and talk on social media. It'sa pretty serious joke, the amount of effort required for this, and it helps a lot.

Stage 2– Fine-Grained Cyber bullying Detection (LSTM):

These cond step involves fine-grained cyber bullying detection, employing a Long Short-Term Memory (LSTM) network. This model predicts the degree of aggression in each post using the by stander role probabilities from Stage1 and the tweet embeddings that were previously generated. Because it learns how messages evolve gradually with in a thread, the LSTM proves effective in this task since it can detect when someone is defending the victim or when aggression is rising. The output layer divides the final prediction into three groups: Aggression without Bullying, High Aggression, and Low Aggression. To prevent overfitting and guarantee that the model works well on unseen data, dropout regularization (rate=0.3) and ReLU activation functions are employed. By linking the two phases, the system captures user interactions and social relationships along with increasing accuracy. This combined method provides the model with a better understanding of online bullying trends and how bystander behaviour affects aggression levels.

C. Algorithmic Overview

The chained process can be represented as:

For each tweetXi with labels Yi = [by stander_role, aggression_label]:

1. Trainf1using BERT(Xi)→predict by stander_role
2. Append predicted role to feature space
3. Trainf2 using LSTM (Xi+predicted_role) →predict aggression_label Return combined predictions

This sequential prediction enables label dependency learning, where by stander roles influence the subsequent aggression classification outcome.

IV. RESULTS AND DISCUSSION

Google Colab with a T4 GPUwas employed for the experiments. 60% of the dataset was allocated for training, 20% for validation, and 20% for testing.

A. Model Performance

The chained model was contrasted with a base line model that ignored by stander roles to evaluate the effectiveness of the suggested strategy. TheCYBY23 dataset was used to train and test both models in the same way.Weighted precision, recall, and F1-score were used to gauge performance.

Table 2. Comparison of baseline and proposed models evaluated on the CYBY23 dataset

Model	Weighted-Precision	Weighted-Recall	Weighted-F1-Score
Baseline (Without By standers)	0.70	0.71	0.72
Proposed(With Bystanders)	0.84	0.83	0.89

The suggested chained model beat the base line system in every metric, as indicated in Table1.The inclusion of by stander roles improved classification accuracy by approximately 25%, proving that contextual and social-role information enhances fine-grained cyber bullying detection. The improvement in F1-score from 0.72 to 0.89 demonstrates the effectiveness of integrating behavioural dynamics within the learning process. Including bystander features improved detection accuracy by approximately 25%, confirming that conversational context significantly impacts model reliability.

B. Optimization Analysis

The Adam optimizer produced the best convergence at a learning rate of 0.001 and 20 epochs. Dropout layers and early stopping techniques were applied to prevent over fitting. To analyze the effect of different optimization algorithms, we compared Adam, RMS Prop, and SGD on the CYBY23 dataset. Validation loss was monitored across epochs for each optimizer to determine the most stable and efficient configuration. As illustrated in Fig.3, the Adam optimizer achieved the lowest validation loss and stable convergence throughout training, outperforming RMS Prop and SGD. This confirms that Adam offers superior adaptive learning capability for the chained BERT–LSTM framework, making it ideal for fine-grained cyberbullying detection.

C. Observations

The proposed chained BERT–LSTM framework was evaluated on multiple parameters, including role detection accuracy, aggression classification, and overall interpretability. The following findings emerged during analysis. Tweets involving instigator roles were often correlated with high aggression.

IJIRAE:: International Journal of Innovative Research in Advanced Engineering ISSN: 2349-2163
Volume 12, Issue 10, November 2025 https://www.ijirae.com/archives
https://doi.org/10.26562/ijirae.2025.v1210.13

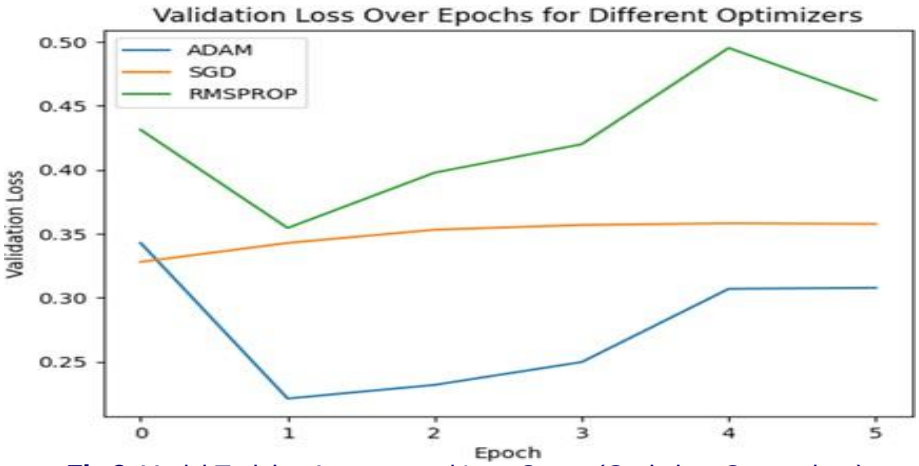

Fig.3–Model Training Accuracy and Loss Curves(Optimizer Comparison)

- Defenders frequently appeared in threads with low aggression or supportive responses.
- The model effectively differentiated sarcasticor in direct bullying, where aggression was subtle.
- In addition to these behavioural in sights, the model's quantitative performance was benchmarked against other algorithms and prior research works to validate its effectiveness.

Table 3. Benchmarking the proposed model against existing deep-learning algorithms

Model	Precision	Recall	F1-Score
SVM	0.65	0.66	0.67
CNN	0.72	0.70	0.71
BiLSTM	0.75	0.74	0.76
BERT	0.81	0.80	0.82
Proposed Chained BERT–LSTM	0.84	0.83	0.89

As illustrated in Table3, the comparative results as presented in Table1 show that traditional models like SVM and CNN achieved moderate accuracy but struggled with contextual understanding. The ability of the BiLSTM model to recognise sequential patterns in text led to its superior performance, and BERT's comprehension of bidirectional semantics further increased precision and recall. Outperforming all baseline models, the Proposed Chained BERT–LSTM model obtained the highest F1-score of 0.89. This improvement is attributed to its two-phase learning methodology, in which the LSTM layer categorises aggression levels according to both textual and social context, while the BERT layer determines bystander roles. By stander feature integration improved interpretability and contextual depth, facilitating the identification of covert and indirect forms of cyberbullying.

V. CONCLUSION

This study offers a chained deep learning frame work that combines LSTM and BERT for explicit by stander role modelling and fine-grained cyberbullying detection. When compared to traditional binary models, the suggested system shows a notable improvement in classification performance. By stander features improve interpretability and accuracy while providing a more thorough understanding of social dynamics and conversational flow during online harassment. In addition to its technical contribution, this project builds a solid social and legal basis. It aligns detected cyberbullying incidents with relevant provisions under the Information Technology Act,2000 (Sections66C, 66D, 66E, 67, 67A),the Indian Penal Code,1860(Sections 354D, 500, 507, 509) ,and the POCSO Act, 2012, providing victims with a clear legal pathway for redress. The system also generates structured digital evidence reports containing offensive messages, timestamps, sender details, and classification results, which can be directly submitted to Cyber Crime Cells or the National Cyber Crime Reporting Portal. Furthermore, the model promotes digital safety and awareness for all user groups, including teenagers, professionals, and vulnerable communities. By combining AI-driven detection with legal awareness and law-enforcement support, the system not only helps prevent online abuse but also strengthens cyber law enforcement and victim empowerment. Future efforts will be dedicated to expanding datasets, supporting multilingual content, and in corporating graph-based role modeling to enhance detection reliability and legal traceability across platforms.

REFERENCE

1. H.S.Alfurayj, S.L.Lutfi, and R.Perumal,"A Chained Deep Learning Model for Fine-Grained Cyber bullying Detection with Bystander Dynamics," IEEE Access, vol. 12, pp. 105588–105602, 2024.
2. Ziems,C.etal.,"Cyber bullying Detection Across Threads Using Contextual Representations,"Social Computing Conf., 2023.
3. P.R.Kumar and S.Kumar,"Temporal Graph Models for Online Harassment Detection," IEEE Transactions on Affective Computing, 2022.
4. L.Fengetal.,"Hierarchical Squashing Attention Networks for Multi-Class Cyber bullying Detection," Journal of Intelligent Systems, 2022.
5. J.Devlinetal.,"BERT: Pre-training of Deep Bi directional Transformers for Language Understanding," NAACL,2019