International Journal of Engineering Technology and Management Sciences
Website: ijetms.in Issue:3 Volume No.10 May - June – 2026
DOI:10.46647/ijetms.2026.v10i03.015 ISSN: 2581-4621

SafeTalk: A Hybrid CNN-RNN Framework for Automated Online Toxic Comment Detection

Mrs. Kavya[1], CH. Lakshana[2], G. Arusha[3], T. Mokshitha[4]
[1]Assistant Professor, Department of CSE(Data Science), VidyaJyothi Institute of Technology, Hyderabad
[2,3,4]UG Student,Dept of CSE(Data Science) , VidyaJyothi Institute of Technology, Hyderabad

Abstract
The widespread prevalence of toxic comments on social media platforms such as Facebook, Instagram, and YouTube has become a critical threat to healthy online discourse. Such content—encompassing insults, hate speech, threats, and identity-based harassment—discourages open participation and can cause measurable psychological harm to users. Automated detection systems are therefore essential for large-scale content moderation. This paper presents SafeTalk, a deep learning framework that integrates Convolutional Neural Networks (CNNs) and Long Short-Term Memory (LSTM) recurrent networks in a sequential hybrid architecture for binary toxic comment classification. CNNs are applied first to extract salient local n-gram features from token embeddings, while LSTM layers subsequently capture sequential context and long-range word dependencies. The proposed model is trained and evaluated on the Jigsaw Toxic Comment Classification Challenge dataset, comprising over 159,000 labelled comments. Experimental results demonstrate that the hybrid approach achieves superior classification performance compared to single-architecture baselines, with a weighted F1-score of approximately 0.95 and an AUC-ROC of 0.97. The system supports both binary and multi-label classification and is designed with deployment feasibility in mind.
Keywords: Online Safety, Toxic Comment Detection, Convolutional Neural Networks, LSTM, Hybrid Deep Learning, Natural Language Processing, Content Moderation, Jigsaw Dataset, Sentiment Analysis.

I. INTRODUCTION

The rapid expansion of user-generated content across social media platforms has introduced a pervasive challenge: the proliferation of toxic comments that undermine constructive online interaction. Toxic language encompasses a wide spectrum of harmful expression, including personal insults, threats of violence, hate speech directed at protected groups, and deliberately offensive remarks. Studies by the Pew Research Institute indicate that 73% of adult internet users have witnessed online harassment, while 40% have experienced it personally, with nearly half of reported cases involving severe forms [1], [10].

These harmful interactions carry tangible consequences. Targeted individuals frequently experience psychological distress, reduced willingness to participate in public discourse, and, in extreme cases, self-harm. Meanwhile, platform operators face increasing regulatory and reputational pressure to enforce community standards at scale. Conventional moderation approaches—relying on keyword blacklists, user reporting, or manual review—are inadequate given the volume and linguistic diversity of user-generated content [9].

Deep learning has demonstrated strong capabilities in text classification tasks, with Convolutional Neural Networks (CNNs) excelling at extracting local features and Recurrent Neural Networks (RNNs)—particularly Long Short-Term Memory (LSTM) variants—capturing sequential dependencies. Neither architecture alone, however, fully addresses the complexity of toxic language detection, which requires sensitivity to both local indicators (specific words or phrases) and broader contextual cues [3], [4].

This paper introduces SafeTalk, a hybrid CNN-LSTM model designed to leverage the complementary strengths of both architectures. Trained on the publicly available Jigsaw Toxic

DOI:10.46647/ijetms.2026.v10i03.015 ISSN: 2581-4621

Comment Classification dataset, the system classifies comments as toxic or non-toxic with high accuracy and is designed to be integrated into real-time moderation pipelines. The remainder of this paper is organised as follows: Section II reviews related literature; Section III states the problem; Section IV describes the proposed system; Section V presents the architecture; Section VI details implementation; Section VII reports results; Sections VIII and IX present future scope and conclusions [5], [8].

II. LITERATURE SURVEY

Background on Toxic Language Detection

Research into automated detection of harmful online content has evolved across three broad generations: rule-based systems, classical machine learning approaches, and deep learning methods. Early rule-based filters employed keyword blacklists and regular expressions, which proved brittle against evolving slang, deliberate obfuscation, and context-dependent meaning [9].

Pang and Lee [1] established foundational concepts in computational opinion mining and sentiment analysis, proposing evaluation frameworks that remain widely referenced. Their survey formalised the distinction between factual and subjective text and outlined feature engineering strategies for sentiment classification tasks.

Yin et al. [2] addressed online harassment detection in Web 2.0 environments through supervised learning, combining content-level, sentiment-level, and contextual features. Their work demonstrated that harassment detection is feasible through machine learning, though dataset scarcity constrained generalisation.

Razavi et al. [3] proposed a multi-level flame detection system incorporating weighted pattern repositories and configurable sensitivity thresholds. The system offered a degree of adaptability to different deployment contexts and sensitivity requirements.

Deep Learning for Text Classification

Mikolov et al. [4] introduced Word2Vec, enabling dense semantic word representations through the Skip-gram and Continuous Bag-of-Words models. The use of Negative Sampling and subsampling of frequent words improved both efficiency and embedding quality, enabling richer input representations for downstream classifiers.

Dos Santos and Gatti [5] demonstrated that deep CNNs could capture semantic nuances in short texts without manual feature engineering, outperforming traditional methods on sentiment classification benchmarks. Kim [6] subsequently showed that even shallow CNNs with pre-trained word vectors achieved strong results across multiple sentence classification tasks with minimal tuning.

Nobata et al. [7] introduced an annotated corpus for abusive language detection on Yahoo platforms and showed that engineered machine learning features could outperform deep learning models when training data was limited, underscoring the importance of dataset quality.

Large-Scale Content Moderation Research

Wulczyn et al. [8], [9] conducted large-scale analyses of personal attacks on English Wikipedia using a combination of crowdsourcing and machine learning. Their findings challenged the assumption that harassment originates primarily from a small population of chronic offenders, revealing instead a broadly distributed behavioural pattern.

Uhls et al. [10] examined the developmental effects of social media on adolescents, highlighting the particular vulnerability of younger users to toxic online environments. Their work underscores the public health dimension of content moderation and motivates investment in effective automated solutions.

III. PROBLEM STATEMENT

Social media platforms generate vast volumes of user comments daily, making manual content moderation both impractical and prohibitively costly at scale. Existing automated detection approaches based on single-architecture models suffer from significant limitations: CNN-based

DOI:10.46647/ijetms.2026.v10i03.015 ISSN: 2581-4621

classifiers capture local lexical patterns effectively but lack the capacity to model long-range contextual dependencies, while RNN-based systems provide sequential context modelling but are computationally intensive and prone to overfitting on imbalanced datasets [3], [4].

Additionally, toxic language is inherently varied: a comment may contain no explicit profanity yet convey clear hostility through sarcasm, implication, or targeted framing. Detecting such nuanced toxicity requires a model capable of simultaneously extracting indicative local features and integrating them within a broader sequential context. Current single-model deployments fail to satisfy this dual requirement, resulting in elevated false-negative rates that allow harmful content to persist [9].

There is therefore a need for an integrated deep learning system that combines local feature extraction with sequential modelling, is trained on a large and representative labelled dataset, and can be deployed in near real-time moderation pipelines. This paper addresses this need through the SafeTalk hybrid architecture [5], [8].

IV. PROPOSED SYSTEM

SafeTalk is designed to provide automated, scalable toxic comment classification through a hybrid deep learning pipeline. The system accepts raw comment text as input and produces a toxicity probability score along with a binary or multi-label classification outcome. The architecture is composed of four principal stages: text preprocessing, embedding, feature extraction via CNN, sequential modelling via LSTM, and classification via a dense output layer.

In the preprocessing stage, comments undergo cleaning to remove HTML markup, URLs, punctuation, and special characters. Text is lowercased, tokenised using the Keras Tokenizer utility, stop words are removed, and words are lemmatised to their root forms. Sequences are padded or truncated to a fixed maximum length of 200 tokens [3], [4].

The embedding stage maps token indices to dense 128-dimensional vectors initialised using GloVe pretrained weights trained on six billion tokens. These representations capture semantic and syntactic relationships between words, providing a richer input signal than one-hot encoding and enabling effective transfer learning.

The CNN block applies 128 one-dimensional convolutional filters of kernel size five to the embedded sequences, followed by batch normalisation and max-pooling. This extracts the most salient local n-gram features while reducing sequence length. The resulting feature maps are then processed by a Bidirectional LSTM layer with 128 units, which models sequential dependencies in both forward and backward directions, capturing context that single-direction LSTM layers may miss [5], [8].

The output of the LSTM is passed through dropout regularisation layers and a 64-unit dense hidden layer before reaching the sigmoid output neuron. This final layer produces a toxicity probability score between 0 and 1, with values above 0.5 classified as toxic. Class-weighted binary cross-entropy loss accounts for the natural imbalance between toxic and non-toxic examples in the training data.

V. SYSTEM ARCHITECTURE

The SafeTalk architecture follows a unified pipeline in which each processing stage feeds directly into the next, enabling end-to-end gradient flow during training. The input layer accepts padded integer token sequences of fixed length 200. These are projected into the 128-dimensional embedding space, initialised with GloVe weights to exploit pretrained semantic knowledge.

The first convolutional layer applies 128 filters with a kernel width of five, scanning across the embedded token sequences to detect n-gram patterns indicative of toxicity. Batch normalisation stabilises the activations following convolution, and a max-pooling layer with pool size two halves the sequence length, reducing computational cost while retaining the strongest detected features.

The pooled CNN output is processed by a Bidirectional LSTM layer. Bidirectional processing allows the network to incorporate both past and future token context when encoding each position,

DOI:10.46647/ijetms.2026.v10i03.015 ISSN: 2581-4621

improving sensitivity to grammatical and semantic structures that span word boundaries. The LSTM output is regularised with a dropout rate of 0.4, applied both within recurrent connections and to the final hidden state [5].

Two dense layers follow: a hidden layer with 64 units and ReLU activation that performs non-linear transformation of the combined CNN-LSTM representation, and an output layer with a single sigmoid unit that produces the final toxicity probability. For multi-label classification, the output layer is extended to six sigmoid units corresponding to the Jigsaw label categories: toxic, severe toxic, obscene, threat, insult, and identity hate [3], [4].

VI. IMPLEMENTATION DETAILS

Programming Language and Environment

Python 3.9 serves as the primary implementation language, selected for its extensive ecosystem of machine learning and natural language processing libraries. All experiments are conducted in Google Colaboratory with GPU acceleration (NVIDIA T4), enabling efficient training of deep learning models without local hardware requirements [3], [4].

Libraries and Frameworks

TensorFlow 2.x and Keras provide the deep learning infrastructure, including the model API, layer primitives, and training utilities. NumPy and Pandas handle numerical operations and dataset management respectively. Matplotlib is used for training visualisation and diagnostic plotting. Scikit-learn supplies evaluation metric computation including the classification report, confusion matrix, and ROC-AUC functions [7].

Data Handling and Preprocessing

The Jigsaw Toxic Comment Classification Challenge dataset is loaded from Kaggle and split into training, validation, and test subsets. Comments are cleaned using regular expressions, tokenised using the Keras Tokenizer with a vocabulary cap of 20,000 terms, and padded to a maximum sequence length of 200 tokens. GloVe 200-dimensional pretrained embeddings are loaded and used to initialise the embedding matrix [1].

Model Training and Optimisation

The model is compiled with binary cross-entropy loss and the Adam optimiser with an initial learning rate of 0.001. Class weights inversely proportional to class frequencies are applied to the loss function to address label imbalance. Training employs early stopping with a patience of three epochs based on validation loss, ReduceLROnPlateau scheduling with a factor of 0.5, and ModelCheckpoint saving the best-performing weights by validation AUC [5], [8].

VII. RESULTS

Training Performance

The SafeTalk hybrid model converges within six to eight epochs across all experimental runs. Training and validation accuracy curves align closely throughout training, with validation accuracy stabilising at approximately 95.5% and training accuracy reaching 96.2%. The close correspondence between training and validation metrics confirms that the regularisation strategy effectively mitigates overfitting despite the class imbalance present in the dataset.

Classification Report

Evaluation on the held-out test set yields the following classification metrics: for the non-toxic class, precision 0.97, recall 0.98, and F1-score 0.97; for the toxic class, precision 0.88, recall 0.80, and F1-score 0.84. The overall weighted average F1-score is 0.95 and the ROC-AUC score is 0.97, indicating strong discriminative ability across both classes [3], [4].

Comparison with Baselines

The hybrid CNN-LSTM architecture outperforms both the standalone CNN baseline (weighted F1 0.91, AUC 0.94) and the standalone LSTM baseline (weighted F1 0.93, AUC 0.95). The performance improvement is most pronounced for medium-to-long comments, where sequential context plays a more significant role in determining toxicity. On short comments of fewer than ten

DOI:10.46647/ijetms.2026.v10i03.015 ISSN: 2581-4621

tokens, all three models perform comparably, as local lexical features dominate in that regime [5], [9].

Interactive Classification Demo

A command-line interactive demo allows users to submit arbitrary comment text and receive a real-time toxicity probability and binary label. Testing with a range of comments including borderline cases and sarcastic expressions confirms that the model generalises reasonably beyond the training distribution, though context-dependent ambiguity remains a challenge that future transformer-based extensions may address [8].

VIII. FUTURE SCOPE

Several directions can extend SafeTalk's capabilities and broaden its applicability. First, incorporating transformer-based encoders such as BERT or DistilBERT as the embedding and feature extraction backbone is expected to yield substantial accuracy improvements, given their deeper contextual representations and pretraining on large corpora [4].

Second, extending multilingual support through multilingual pretrained models such as XLM-R would enable the system to classify toxic content across languages, addressing the global scope of the problem. Third, real-time deployment as a REST API microservice would allow integration with social media platform moderation pipelines, with latency-optimised inference through model quantisation and ONNX export [5], [8].

Fourth, the system can be enhanced with explainability techniques such as attention visualisation and LIME, enabling moderators to understand the specific words or phrases driving toxicity predictions. Fifth, data augmentation strategies including back-translation and synonym substitution could be applied to improve model robustness and address the underrepresentation of rare toxicity types such as threats and identity-based hate speech.

Finally, integration with federated learning frameworks would allow privacy-preserving model training on decentralised platform data without centralising sensitive user content, facilitating adoption by privacy-conscious operators in jurisdictions subject to stringent data protection regulations [10].

IX. CONCLUSION

This paper has presented SafeTalk, a hybrid deep learning framework combining Convolutional Neural Networks and Long Short-Term Memory networks for automated toxic comment classification. By applying CNN layers to extract local n-gram features from token embeddings and LSTM layers to model sequential contextual dependencies, the system addresses the complementary limitations of single-architecture approaches.

Trained and evaluated on the Jigsaw Toxic Comment Classification Challenge dataset, SafeTalk achieves a weighted F1-score of 0.95 and a ROC-AUC of 0.97, outperforming standalone CNN and LSTM baselines. The system supports binary and multi-label classification and includes a preprocessing pipeline, class-weighted training, and an interactive classification interface [3], [4].

The modular architecture of SafeTalk facilitates future enhancements, including integration of transformer-based encoders, multilingual extension, and real-time deployment. This work demonstrates that hybrid deep learning architectures represent a practical and effective approach to automated online content moderation, contributing to the broader goal of creating safer and more inclusive digital environments for all users [5], [8].

REFERENCES

[1] Pang, B. and Lee, L., "Opinion mining and sentiment analysis," Foundations and Trends in Information Retrieval, vol. 2, no. 1-2, pp. 1-135, 2008.

[2] Yin, D., Xue, Z., Hong, L., Davison, B. D., Kontostathis, A. and Edwards, L., "Detection of harassment on web 2.0," in Proc. Content Analysis in the WEB 2.0 Workshop (CAW2.0), WWW, Madrid, Spain, 2009.

DOI:10.46647/ijetms.2026.v10i03.015 ISSN: 2581-4621

[3] Razavi, A. H., Inkpen, D., Uritsky, S. and Matwin, S., "Offensive language detection using multi-level classification," in Proc. 23rd Canadian Conf. Advances in Artificial Intelligence, 2010, pp. 16-27.
[4] Mikolov, T., Sutskever, I., Chen, K., Corrado, G. S. and Dean, J., "Distributed representations of words and phrases and their compositionality," in Advances in Neural Information Processing Systems, vol. 26, 2013, pp. 3111-3119.
[5] Dos Santos, C. N. and Gatti, M., "Deep convolutional neural networks for sentiment analysis of short texts," in Proc. COLING, 2014.
[6] Kim, Y., "Convolutional neural networks for sentence classification," in Proc. Conference on Empirical Methods in Natural Language Processing (EMNLP), 2014, pp. 1746-1751.
[7] Nobata, C., Tetreault, J., Thomas, A., Mehdad, Y. and Chang, Y., "Abusive language detection in online user content," in Proc. 25th International Conference on World Wide Web (WWW), 2016, pp. 145-153.
[8] Wulczyn, E., Thain, N. and Dixon, L., "Wikipedia Detox," Figshare, 2016. [Online]. Available: https://doi.org/10.6084/m9.figshare.4054689
[9] Wulczyn, E., Thain, N. and Dixon, L., "Ex machina: Personal attacks seen at scale," in Proc. 26th International Conference on World Wide Web (WWW), 2017.
[10] Uhls, Y. T., Ellison, N. B. and Subrahmanyam, K., "Benefits and costs of social media in adolescence," Pediatrics, vol. 140, no. S2, pp. S67-S70, 2017.