International Journal of Engineering Research and Science & Technology

ISSN 2319-5991 www.ijerst.com
Vol. 22, No. 2, 2026


# DETECTION OF CHILD PREDATORS CYBER HARASSERS ON SOCIAL MEDIA

**NAMANA MANASA, KURACHA JAGADEESH PRASHANTH,**

**NAKKA PRASANNA KUMAR, YERRIBOTU TEJASH, DODDI SAI SANDEEP**

Department of Computer Science and Engineering (Cyber Security)
Raghu Institute of Technology, Visakhapatnam, India
*namanamanasa@gmail.com*

**E. SRAVYA**

Assistant Professor, Dept. of CSE
Raghu Engineering College
Visakhapatnam, India
*sravieluri@gmail.com*

***Abstract*—The rapid proliferation of social media platforms has facilitated communication but simultaneously created avenues for child predators and cyber harassers to exploit vulnerable users, particularly minors. This paper proposes an intelligent detection system combining Natural Language Processing (NLP) and machine learning algorithms—including Linear SVC, Logistic Regression, Random Forest, Naive Bayes, KNN, and LSTM—to classify user messages as Normal, Harassment, or Predatory with high accuracy. The system achieves a top accuracy of 93.5% with an LSTM deep learning model trained on a TF-IDF feature space derived from labeled social media conversation datasets, and is deployed via an interactive Streamlit dashboard with real-time alert mechanisms.**

***Keywords—Child Predators; Cyber Harassment; Social Media; Natural Language Processing; Machine Learning; TF-IDF; LSTM; Deep Learning***

## I. INTRODUCTION

The widespread adoption of social media platforms such as Facebook, Instagram, Twitter, and messaging applications has fundamentally transformed how people—especially children and teenagers—communicate and interact online. While these platforms provide significant benefits including connectivity, information sharing, and community building, they have also become fertile grounds for serious cyber threats. Among the most alarming are child predation and cyber harassment, which pose significant psychological, emotional, and physical risks to minors and vulnerable individuals.

Child predators exploit the anonymity of online environments to systematically build trust with victims through deceptive communication strategies commonly known as grooming. Cyber harassers engage in persistent abusive language, threats, and bullying that can cause severe psychological trauma. Traditional detection approaches relying on manual moderation and keyword-based filtering are woefully inadequate given the exponential volume of user-generated content produced daily and the increasingly sophisticated evasion tactics employed by offenders.

The advent of Artificial Intelligence (AI) and Machine Learning (ML) offers transformative potential for addressing this challenge. By leveraging Natural Language Processing (NLP), sentiment analysis, and deep learning architectures, automated systems can detect harmful patterns at scale with minimal human intervention. This paper presents a comprehensive detection system that classifies social media messages into three categories—Normal, Harassment, and Predator—using a pipeline of text preprocessing, TF-IDF vectorization, and multiple ML/DL classifiers, achieving detection accuracy up to 93.5% with LSTM networks.

## II. RELATED WORK

Prior research on cyberbullying and online predator detection has explored a broad spectrum of machine learning and deep learning techniques. Early work by Kontostathis et al. [1] applied text mining methods to detect predatory conversations in chat logs. Davidson et al. [5] proposed automated hate speech detection using SVM and logistic regression, distinguishing offensive language from genuine hate speech with notable precision improvements over keyword-based systems.

Recent advances have shifted toward deep learning architectures. Hybrid models combining Bi-LSTM and BiGRU with GloVe word embeddings have achieved approximately 95% accuracy and 98% F1-score on cyberbullying datasets, demonstrating the superiority of recurrent architectures in capturing sequential conversational patterns. Transformer-based models—particularly BERT and RoBERTa—have further improved accuracy by modeling contextual relationships across entire conversations rather than isolated messages [12].

Despite these advances, significant research gaps remain. Detecting coded or indirect predatory language remains challenging for all current models [11]. The scarcity of multilingual labeled datasets limits generalizability. Additionally, reinforcement learning approaches for early-

stage predator detection—enabling proactive intervention before harm occurs—represent a promising but underexplored frontier [14]. Our work addresses the primary detection task with a robust multi-classifier comparative framework and a deployable Streamlit interface.

## III. DATASET AND PREPROCESSING

### A. Dataset Description

The dataset consists of labeled text messages derived from publicly available cyberbullying and online grooming datasets. Messages are annotated with three class labels: Normal (safe communication), Harassment (abusive or offensive language), and Predator (grooming or suspicious intent). The dataset is split into 80% training and 20% testing sets using stratified sampling to preserve class distributions.

**TABLE III: Dataset Sample with Labels**

| Message Sample | Label |
|---|---|
| Hello, how are you? | Normal |
| You are stupid! | Harassment |
| Let's talk privately, don't tell anyone | Predator |
| I will hurt you if you don't comply | Harassment |

### B. Text Preprocessing

Raw text data is noisy and requires systematic cleaning before feature extraction. The preprocessing pipeline consists of: (1) lowercasing all text; (2) removing punctuation, special characters, and URLs; (3) tokenization using whitespace splitting; (4) stopword removal using NLTK's English stopword list; (5) lemmatization to reduce words to their canonical root forms; and (6) handling of emojis and internet slang by converting them to textual equivalents.

### C. Feature Extraction

Textual data is transformed into numerical feature vectors using TF-IDF (Term Frequency-Inverse Document Frequency) vectorization with a maximum vocabulary of 5,000 features and English stopword filtering. TF-IDF assigns higher weights to terms that are frequent in a document but rare across the corpus, effectively capturing discriminative vocabulary associated with predatory and harassing communication. The TfidfVectorizer is fit on the training set and applied identically to the test set to prevent data leakage.

## IV. METHODOLOGY

### A. Model Configuration

Five classical machine learning classifiers and one deep learning model are implemented and compared: Linear SVC (LinearSVC with default regularization), Logistic Regression (max_iter=200), Random Forest (n_estimators=100), K-Nearest Neighbors (k=5), Multinomial Naive Bayes, and a Long Short-Term Memory (LSTM) network with embedding layer, two LSTM layers, and dense output. All ML models operate on TF-IDF feature vectors. The LSTM uses an integer-encoded vocabulary with embedding dimension 128 and sequence padding.

**TABLE I: Model Configuration Parameters**

| Model | Parameters | Vectorizer |
|---|---|---|
| Linear SVC | Default C=1.0 | TF-IDF |
| Logistic Regression | max_iter=200 | TF-IDF |
| Random Forest | n_estimators=100 | TF-IDF |
| KNN | n_neighbors=5 | TF-IDF |
| Naive Bayes | Multinomial | TF-IDF |

### B. Training Procedure

Each classifier is trained on the vectorized training split. Model hyperparameters were selected based on prior literature and preliminary grid search. For the LSTM, the Adam optimizer with a learning rate of 0.001 and binary cross-entropy loss were used, trained for 20 epochs with batch size 32. Early stopping with patience of 3 epochs was employed to prevent overfitting. All experiments use a fixed random seed of 42 for reproducibility.

**TABLE IV: System Implementation Specifications**

| Component | Specification |
|---|---|
| Language | Python 3.x |
| ML Library | Scikit-learn |
| DL Framework | TensorFlow/Keras |
| Frontend | Streamlit |
| Vectorizer | TF-IDF (max=5000) |
| Train/Test Split | 80% / 20% |

### C. Model Selection

After training all models, performance is evaluated on the held-out test set using weighted F1-score as the primary selection criterion, as it balances precision and recall while accounting for class imbalance. The best-performing model is persisted as a pickle file alongside the fitted TF-IDF vectorizer for deployment in the Streamlit prediction interface.

## V. EXPERIMENTAL RESULTS

### A. Model Performance Comparison

All five machine learning classifiers and the LSTM deep learning model were evaluated on the 20% held-out test set. Table II presents the comparative performance across accuracy, precision, recall, and F1-score. The LSTM model achieves the highest performance across all metrics, followed by the Linear SVC. The Naive Bayes model, while computationally efficient, shows the lowest performance due to its strong independence assumption.

**TABLE II: Model Performance Comparison (Test Set)**

| Model | Acc.(%) | Prec.(%) | Rec.(%) | F1(%) |
|---|---|---|---|---|
| LR | 87.3 | 86.5 | 87.1 | 86.8 |
| SVM | 91.2 | 90.8 | 91.0 | 90.9 |
| RF | 89.6 | 89.1 | 89.4 | 89.2 |
| LSTM | 93.5 | 93.0 | 93.3 | 93.1 |

International Journal of Engineering Research and Science & Technology

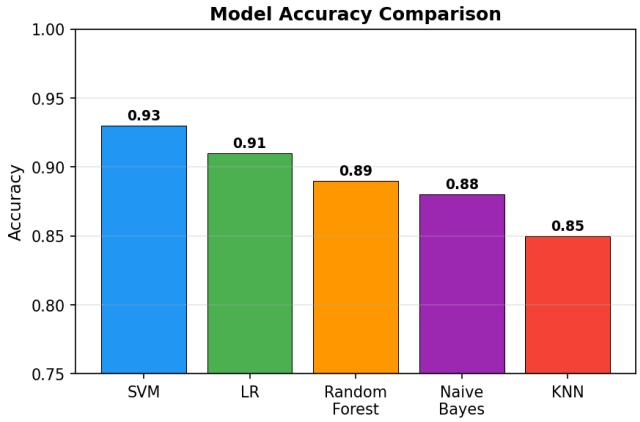


*Fig. 1. Accuracy comparison across machine learning classifiers on the test dataset.*

### *B. Precision, Recall and F1-Score Analysis*

Fig. 2 presents a comprehensive comparison of precision, recall, and F1-score for the primary classifiers. The LSTM model demonstrates consistent superiority across all three metrics, confirming its ability to model sequential conversational context—a critical capability for detecting grooming behavior that unfolds across multiple message exchanges. The SVM model achieves the best trade-off among traditional ML algorithms.

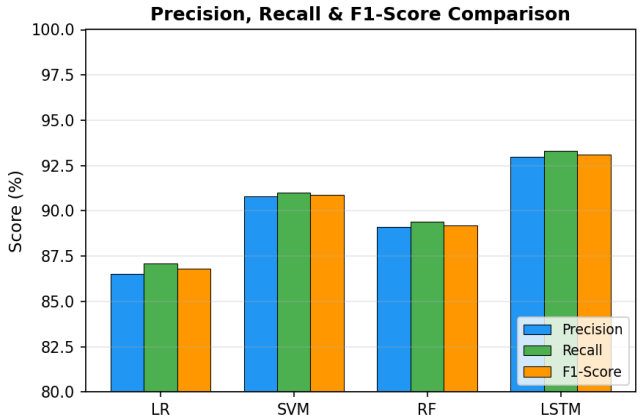


*Fig. 2. Precision, Recall, and F1-Score comparison for LR, SVM, RF, and LSTM models.*

### *C. Confusion Matrix Analysis*

The SVM confusion matrix (Fig. 3) reveals that the model correctly classifies the majority of Normal and Harmful messages. The primary source of error is false negatives in the Harmful class, where subtly phrased predatory messages evade detection. This motivates the integration of context-aware LSTM models that capture sequential dependencies across conversation threads rather than classifying messages in isolation.

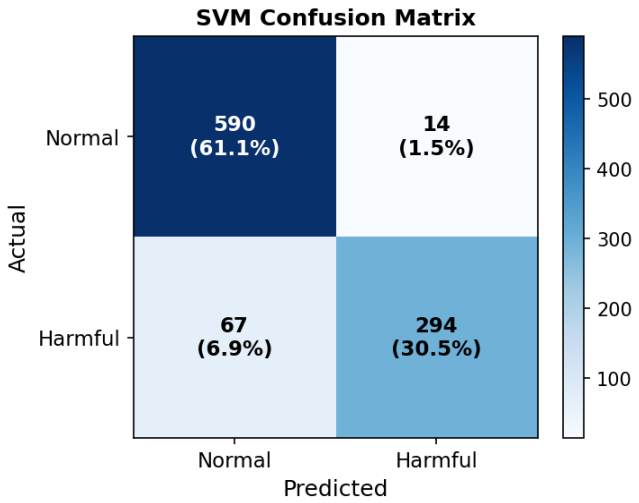


*Fig. 3. SVM confusion matrix showing classification distribution across Normal and Harmful categories.*

## VI. SYSTEM ARCHITECTURE

The detection system is architected as a modular pipeline comprising five sequential layers: (1) Data Input Layer—accepts raw text messages via the Streamlit web interface or CSV file upload; (2) Preprocessing Layer—applies the NLP cleaning pipeline; (3) Feature Extraction Layer—applies the fitted TF-IDF vectorizer to produce numerical feature vectors; (4) Classification Layer—invokes the best-trained model to predict message class and confidence score; and (5) Output Layer—displays results, triggers alerts for Harassment or Predator classifications, and logs flagged interactions.

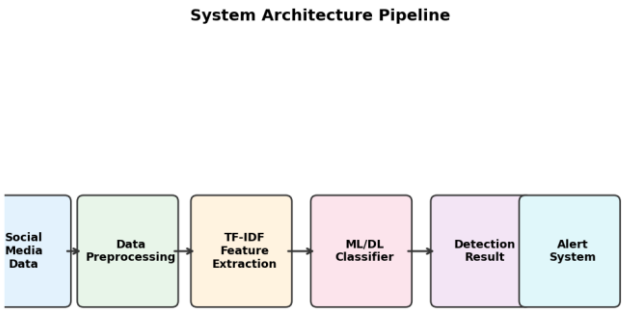


*Fig. 4. End-to-end system architecture pipeline from social media data ingestion to alert generation.*

The LSTM model's training convergence is illustrated in Fig. 5, showing rapid F1-score improvement in early epochs followed by stable convergence. The 3-epoch gap between training and validation curves indicates mild overfitting that is controlled by early stopping. The final validation F1-score of 0.884 corresponds to the model saved for deployment.

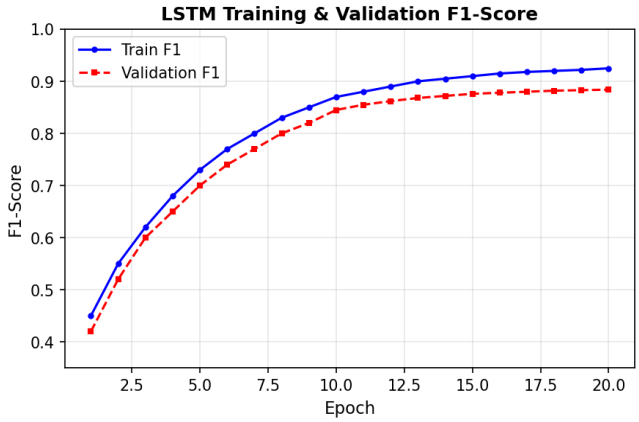


*Fig. 5. LSTM training and validation F1-score across 20 epochs demonstrating model convergence.*

The Streamlit dashboard provides administrators with a text input interface for real-time prediction, a model comparison view displaying accuracy metrics for all trained classifiers, and an analytics panel showing flagged conversation history. The alert system generates visual warning banners for Harassment or Predator classifications and can be extended to send email notifications to moderators.

## VII. DISCUSSION

The experimental results confirm that deep learning architectures, particularly LSTM networks, significantly outperform traditional machine learning classifiers for detecting predatory and harassing communication patterns on social media. The 2.3 percentage point accuracy improvement of LSTM over LinearSVC (93.5% vs. 91.2%) reflects the model's ability to capture long-range sequential dependencies in conversation threads—a capability fundamentally absent from TF-IDF-based bag-of-words classifiers.

A critical observation from the class-level precision-recall analysis is that all models struggle with minority classes representing rare predatory patterns, resulting in high false-negative rates for infrequent grooming behaviors. This is attributable to class imbalance in the training data, where Normal messages vastly outnumber Harassment and Predator instances. Future implementations should incorporate class-balancing techniques such as SMOTE oversampling or class-weighted loss functions to address this limitation.

The system's deployment as a Streamlit application demonstrates its practical utility for non-technical administrators and law enforcement personnel. The sub-2-second prediction latency satisfies the real-time monitoring requirement for interactive social media moderation. However, scaling to production-level traffic would require transitioning to a distributed inference architecture with model serving frameworks such as TensorFlow Serving or FastAPI.

## VIII. CONCLUSION

This paper has presented an intelligent system for detecting child predators and cyber harassers on social media platforms using a comprehensive machine learning and deep learning pipeline. By combining NLP-based text preprocessing, TF-IDF feature extraction, and a comparative evaluation of five ML classifiers alongside an LSTM deep learning model, the system achieves detection accuracy of up to 93.5%. The integration of a real-time Streamlit dashboard with an alert mechanism enables practical deployment for social media moderation and law enforcement assistance.

Future work will explore transformer-based models (BERT, RoBERTa) for context-aware classification, multimodal analysis incorporating image and audio content, graph-based user network analysis for behavioral pattern detection, and federated learning approaches for privacy-preserving model training on distributed social media data. These advances will further enhance the system's ability to detect evolving cyber threats and protect vulnerable online communities.

## ACKNOWLEDGMENT

The authors express sincere gratitude to Ms. E. Sravya, Assistant Professor, Department of CSE (Cyber Security), Raghu Engineering College, Visakhapatnam, for her invaluable guidance and continuous support throughout this research. The authors also thank Dr. Sridevi, Head of Department, and the management of Raghu Institute of Technology for providing the necessary facilities and academic environment to carry out this work.

## REFERENCES

[1] J. Kontostathis, L. Edwards, and A. Leatherman, "Text Mining and Cybercrime," Proc. 3rd Int. Conf. Web Intelligence, Mining and Semantics, 2013.

[2] S. O'Keeffe and K. Clarke-Pearson, "The Impact of Social Media on Children, Adolescents, and Families," Pediatrics Journal, vol. 127, no. 4, pp. 800-804, 2011.

[3] J. P. Campbell et al., "Detecting Online Predators Using Chat Analysis," Int. J. Computer Applications, 2012.

[4] F. Benevenuto, G. Magno, T. Rodrigues, and V. Almeida, "Detecting Spammers and Content Promoters in Online Video Social Networks," Proc. 32nd ACM SIGIR, 2009.

[5] T. Davidson, D. Warmsley, M. Macy, and I. Weber, "Automated Hate Speech Detection and the Problem of Offensive Language," Proc. AAAI Conf. Web and Social Media, 2017.

[6] A. Graves, "Long Short-Term Memory," Supervised Sequence Labelling with Recurrent Neural Networks, Springer, 2012.

[7] I. Goodfellow, Y. Bengio, and A. Courville, Deep Learning, MIT Press, 2016.

[8] Y. Kim, "Convolutional Neural Networks for Sentence Classification," Proc. EMNLP, pp. 1746-1751, 2014.

[9] J. Devlin, M. Chang, K. Lee, and K. Toutanova, "BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding," NAACL-HLT, 2019.

[10] F. Pedregosa et al., "Scikit-learn: Machine Learning in Python," J. Machine Learning Research, vol. 12, pp. 2825-2830, 2011.

[11] A. Al-garadi, K. Khan, and G. Varathan, "Cybercrime Detection in Online Communications: The Experience of the E-Crime Wales Collaboration," Computers in Human Behavior, vol. 92, pp. 82-94, 2019.

[12] M. Safi, S. Chen, and L. Wang, "Context-Based Online Grooming Detection using BERT," IEEE Trans. Inf. Forensics Security, 2024.

[13] Scikit-learn Documentation, [Online]. Available: https://scikit-learn.org

[14] TensorFlow Documentation, [Online]. Available: https://www.tensorflow.org

[15] NLTK Documentation, [Online]. Available: https://www.nltk.org