An Efficient Hybrid Deep Learning Approach for Detecting Online Abusive Language

Vuong M. Ngo[1,4], Cach N. Dang[2] ✉, Kien V. Nguyen[3], and Mark Roantree[4]

[1] Faculty of Information Technology, Ho Chi Minh City Open University, Ho Chi Minh City, Vietnam
[2] BRIDGE Research Group, Ho Chi Minh City University of Transport, Ho Chi Minh City, Vietnam
[3] Ho Chi Minh University of Banking, Ho Chi Minh City, Vietnam
[4] Insight Centre for Data Analytics, School of Computing, Dublin City University, Dublin, Ireland
vuong.nm@ou.edu.vn, cach@ut.edu.vn, kiennv_htttql@hub.edu.vn, mark.roantree@dcu.ie

Abstract. The digital age has expanded social media and online forums, allowing free expression for nearly 45% of the global population. Yet, it has also fueled online harassment, bullying, and harmful behaviors like hate speech and toxic comments across social networks, messaging apps, and gaming communities. Studies show 65% of parents notice hostile online behavior, and one-third of adolescents in mobile games experience bullying. A substantial volume of abusive content is generated and shared daily, not only on the surface web but also within dark web forums. Creators of abusive comments often employ specific words or coded phrases to evade detection and conceal their intentions. To address these challenges, we propose a hybrid deep learning model that integrates BERT, CNN, and LSTM architectures with a ReLU activation function to detect abusive language across multiple online platforms, including YouTube comments, online forum discussions, and dark web posts. The model demonstrates strong performance on a diverse and imbalanced dataset containing 77,620 abusive and 272,214 non-abusive text samples (ratio 1:3.5), achieving approximately 99% across evaluation metrics such as Precision, Recall, Accuracy, F1-score, and AUC. This approach effectively captures semantic, contextual, and sequential patterns in text, enabling robust detection of abusive content even in highly skewed datasets, as encountered in real-world scenarios.

Keywords: Supervised Learning · Artificial Intelligent · Child Abuse Comments · Social Media · Forums · Dark Web

1 Introduction

The rise of the digital age and the Internet has fueled widespread use of platforms like social media and forums, enabling free expression. Nearly 45% of the global

population uses social media, often becoming addicted. However, this has also led to harassment, bullying, and harmful behavior, such as hate speech, toxic comments, and sharing obscene content. Digital technologies make such behavior possible across various online platforms, including social media, messaging, and gaming sites [6]. Social media has become a major platform for hostile behavior, with 65% of parents worldwide acknowledging its prevalence. Additionally, one-third of adolescents who play mobile games have reported being victims of bullying [38]. According to [33], 44% of internet users in the United States have personally encountered various forms of online harassment, with 28% experiencing severe hostility.

Automatically detecting and analyzing online abuse text presents significant challenges due to the complexity of language, contextual ambiguity, the dynamic evolution of terminology, and the sheer volume of data. These challenges are further amplified in the detection of Abuse Material shared on the dark web, where privacy and anonymity are prioritized, making it difficult to trace perpetrators [23]. Additionally, offenders often employ sophisticated evasion techniques, such as the use of code words, slang, cryptic abbreviations, and other forms of linguistic obfuscation, to bypass detection mechanisms and conceal illicit activities [22]. Furthermore, legal and ethical constraints necessitate careful handling of sensitive data, limiting the extent to which automated systems can operate effectively. These factors collectively make the development of robust AI-driven detection methods both critical and highly challenging [30].

Machine learning (ML) and deep learning (DL) techniques have been widely adopted to develop abusive text detection models, including Support Vector Machines (SVM) [21], Random Forest [4], Decision Trees [35], Naïve Bayes (NB) [19], K-Nearest Neighbors (KNN) [36], Long Short-Term Memory (LSTM) [5], Bi-directional LSTM (Bi-LSTM) [16], Convolutional Neural Networks (CNN) [17], and Bidirectional Encoder Representations from Transformers (BERT) [29]. However, each individual ML/DL technique has limitations, such as sensitivity to noise, difficulty handling large-scale data, or an inability to capture complex relationships. Hybrid models that combine ML/DL techniques can leverage the strengths of multiple algorithms to address these weaknesses [8]. Therefore, we propose a hybrid DL model combining BERT, CNN, LSTM, and ReLU activation. This integration improves prediction accuracy, especially for complex or imbalanced datasets, enhances generalization, reduces overfitting, and adapts to a wider range of tasks. These benefits are particularly valuable in domains like abusive text detection, where language can be highly variable and nuanced.

The contributions of our research can be articulated as follows:

- Building an integrated dataset of 77,620 abusive and 272,214 non-abusive text samples from three sources to create a diverse, imbalanced collection that enables comprehensive model evaluation across real-world scenarios.
- Proposing a hybrid DL model that effectively captures abusive text in both English and Romanized scripts by integrating BERT, CNN, LSTM, and the ReLU activation function. The model achieves very high performance with

a Precision of 0.991, Recall of 0.986, Accuracy of 0.995, F1-score of 0.989, and AUC of 0.992, as evaluated using five-fold cross-validation.
– Conducting a comprehensive comparison of the proposed model with traditional ML and standalone DL baselines using a diverse benchmark dataset comprising YouTube comments, forum discussions, and dark web posts, evaluated through 5-fold cross-validation across multiple performance metrics.

The structure of this paper is as follows. In Section 2, we review relevant research. Section 3 outlines our methodology, providing our hybrid DL model and its comprising modules in detail. In Section 4, we describe the dataset and evaluate both our model and selected baseline models using various metrics, followed by a detailed performance analysis. Finally, Section 5 concludes the paper and offers insights for future research directions.

2 Related Work

Statistical analysis of large datasets and CSV files helps uncover patterns and trends, enabling unbiased insights from raw data. Studies such as [18] and [31] utilized statistical methods to analyze abuse-related information. In [18], data analysis began with SPSS, followed by hierarchical logistic regression to assess three key outcomes: (1) sexual revictimization, (2) psychological dating violence, and (3) physical dating violence. Similarly, [31] examined the prevalence and contributing factors of sexual abuse among adolescent girls during the COVID-19 pandemic. While, natural language processing techniques can be used to extract features and content from abuse chats and conversations. In [1], the authors used Latent Dirichlet Allocation to detect tweets about child abuse and applied Conjunctive Analysis of Case Configurations, finding that longer tweets from users with smaller accounts, lacking URLs or images, were more likely to disclose abuse. However, these studies did not leverage the advantages of ML/DL techniques in detecting abuse texts.

Some studies have applied ML/DL techniques to classify abusive posts, such as [11], [14], [15], [3] and [12]. However, [11], [14] and [15] primarily focused on self-figure drawings or pornographic images, utilizing object categories, visual attention mechanisms, gender-related visual features (e.g., long hair, dresses), and image metrics (e.g., luminance, sharpness). In contrast, [3] and [12] worked with videos, combining facial recognition with other biometric modalities, such as speaker recognition and age estimation.

Similar to our task, several studies have applied ML/DL techniques to detect abusive text, such as [7], [34], [5], [13], [20] and [16]. In [7], the authors applied natural language inference alongside the BERT model to detect harmful communication strategies. They used a dataset of 6,771 chat messages sent by child sex offenders, sourced from platforms such as MySpace and Yahoo Instant Messenger. In [34], the authors proposed a model based on BERT, combined with a text tokenization method, to classify complaints across multiple dimensions and provide deeper insights into the dynamics of abuse. They analyzed 1,196 reports collected from the Colombian Child Hotline, covering topics

such as grooming, sexual content disclosure, and cyberbullying. Meanwhile, [5] employed explainable artificial intelligence (XAI) techniques to identify early warning signs, aiming to raise awareness among individuals with limited prior knowledge of child sexual abuse. The study was based on responses to survey questions collected from 3,002 participants. In [13], Kaur et al. developed a robust abusive language detection model using a dataset of 14,200 English tweets. The approach leveraged DL architectures, specifically LSTM and Gated Recurrent Unit (GRU) networks, to capture contextual dependencies and sequential patterns for accurate classification.

Nearly, [20] introduced a hybrid feature selection approach combining the Enhanced Non-Dominated Sorting Genetic Algorithm II with XGBoost, aiming to optimize classification performance while reducing the number of selected features. The model was evaluated on the Levantine Hate Speech and Abusive dataset, consisting of 5,846 political Arabic tweets from Syria and Lebanon, labeled as normal, abusive, or hate speech. Finally, in [16], the authors applied feature selection techniques such as XGBoost and Multilayer Perceptron to enhance a BiLSTM model for detecting cyberbullying texts and identifying specific bullying types. The dataset comprised 15,294 microblogs containing personal cyberbullying attacks, categorized accordingly, along with 17,826 non-cyberbullying texts for comparison.

However, the ML/DL models for abusive text detection proposed in the aforementioned studies (i.e., [7], [34], [5], [13], [20], and [16]) differ from ours in both scope and methodological approach.

3 Methodology

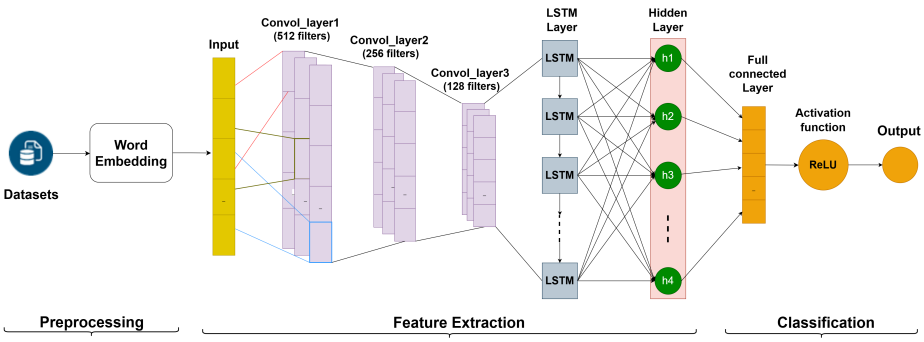


Fig. 1: Process of methodology for hybrid model

In this paper, we propose a hybrid DL model for detecting online abusive comments. The model architecture, illustrated in Figure 1, is designed to predict the abuse polarity of a text and classify it accordingly. It consists of three main modules: `Preprocessing`, `Feature Extraction` and `Classification`. In the

`Preprocessing` module, BERT is employed as the word embedding model to generate feature vectors. The `Feature Extraction` module integrates CNN and LSTM models to capture both local and sequential features of the text. Finally, the `Classification` module applies a ReLU activation function to produce the final output.

The purpose of this module combination is to enhance detection accuracy compared to single-model approaches when applied to complex data from diverse sources, although it requires longer computation time. The hybrid architecture leverages the complementary strengths of BERT, CNN, and LSTM: BERT captures the full contextual meaning of words through bidirectional analysis, CNN effectively extracts salient textual features, and LSTM retains past information via its cell states, enhancing the model's ability to learn sequential dependencies.

3.1 Word Embedding Layer

The initial phase of our model involves transforming the input text into a sequence of word embeddings—dense vector representations that encode semantic relationships among words. In this work, we deployed a pre-trained BERT model tailored for the Online Abusive Comments (OAC) dataset. After fine-tuning its parameters, BERT serves as a feature extractor, generating contextualized embeddings for the proposed hybrid framework. The OAC dataset is passed through the BERT model to generate feature vectors, which are then fed into the subsequent layers of the hybrid architecture. The BERT-base-uncased architecture was used to produce contextualized embeddings with a hidden size of 768.

3.2 Convolutional Neural Networks model

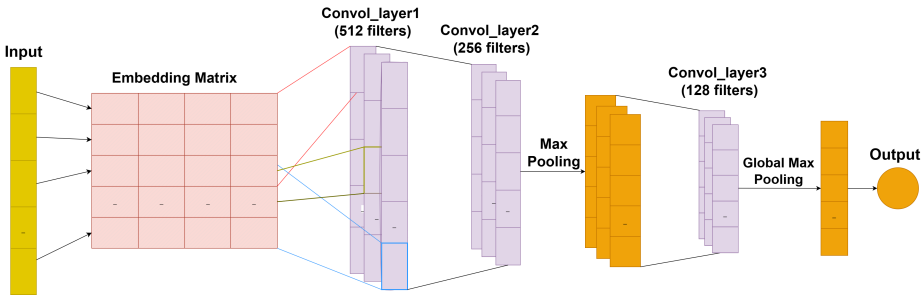


Fig. 2: A convolutional neural network

A Convolutional Neural Network (CNN) is a form of feedforward neural network that processes data sequentially from input to output, without any feedback loops. It employs a deep layered architecture [37], usually starting with convolutional and pooling (or subsampling) layers that extract and transform features

from the input data. These are then followed by one or more fully connected layers that perform the final classification.

While the general architecture of CNNs is independent of the dimensionality of the input data, their implementation depends on it. The dimensionality determines how subsampling filters slide across the data. In natural language processing, a one-dimensional convolutional layer (1D CNN) with m filters is commonly applied, producing an m-dimensional feature vector for each document n-gram. These feature vectors are then aggregated using max-pooling, followed by a ReLU (Rectified Linear Unit) activation function. The resulting output is then forwarded to a fully connected (linear) layer, which performs the ultimate classification step, as illustrated in Figure 2. To capture local contextual patterns in BERT embeddings, a sequence of three 1D convolutional layers with 512, 256, and 128 filters, respectively, and a kernel size of 3 is applied.

Let $w_{i:n} \in \mathbb{R}^d$ denote the input text consisting of n words, where each word is represented as a d-dimensional embedding vector. The sequence of embeddings forms a $d \times n$ matrix that serves as the input to the convolutional layer, where filters are applied across the text.

For each l-word n-gram, we define:

$$c_i = [w_i, \ldots, w_{i+l-1}] \in \mathbb{R}^{(d \times l)}, \quad 0 \leq i \leq n - l \tag{1}$$

For each filter $f_j \in \mathbb{R}^{d \times l}$, the convolution operation is computed as the inner product $\langle c_i, f_j \rangle$. The resulting convolutional feature maps are collected into a matrix $F \in \mathbb{R}^{n \times m}$, where m is the number of filters.

A max-pooling operation is then applied across the n-gram dimension:

$$p_j = \max_i (F_{ij}) \tag{2}$$

The pooled features are passed through a ReLU non-linearity to introduce activation.

Finally, the extracted feature representation is passed through a fully connected layer that produces a probability distribution across the target classes. The class corresponding to the highest probability is then chosen as the final prediction.

3.3 Long short-term memory model

The Long Short-Term Memory (LSTM) network is a specialized variant of the Recurrent Neural Network architecture [32]. Each LSTM unit comprises three main gates—the `forget gate`, `input gate`, and `output gate`—as well as input/output components and a memory cell that allows the model to learn and retain long-term dependencies. The structure of an LSTM block is illustrated in Figure 3. In our model, the extracted features are passed through an LSTM layer with a hidden size of 500 to effectively capture sequential dependencies.

The `forget gate` regulates the extent to which information from the previous cell state c_{t-1} is preserved:

$$f_t = \sigma \left(W_f \cdot [h_{t-1}, x_t] + b_f \right) \tag{3}$$

The `input gate` determines the amount of new information to incorporate into the cell state, while the candidate cell state $\tilde{c}_t$ is calculated as follows:

$$i_t = \sigma(W_i[h_{t-1}, x_t] + b_i), \quad \tilde{c}_t = \tanh(W_c[h_{t-1}, x_t] + b_c) \tag{4}$$

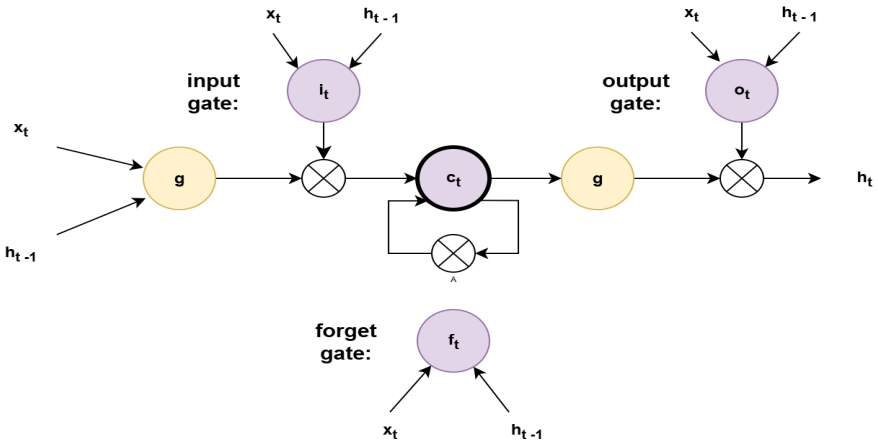


Fig. 3: Illustration of a LSTM block [32]

The updated cell state is then:

$$c_t = f_t \cdot c_{t-1} + i_t \cdot \tilde{c}_t \tag{5}$$

Finally, the `output gate` controls the proportion of the cell state that is exposed to the hidden output:

$$o_t = \sigma(W_o[h_{t-1}, x_t] + b_o), \quad h_t = o_t \cdot \tanh(c_t) \tag{6}$$

Each gate has its own learnable weights and biases, enabling the network to determine how much past information to retain, how much new input to incorporate, and how much of the internal state to expose at each time step.

3.4 Fully Connected Layer and Output

The context vector produced by the LSTM layer with an attention mechanism is fed into a fully connected (dense) layer utilizing a ReLU activation function. This layer enhances the high-level features learned by the CNN and LSTM modules, enabling the model to capture more intricate and non-linear relationships within the data.

Next, the output from the dense layer is passed to the final output layer, which employs an appropriate activation function (e.g., softmax) to generate the abuse classification probabilities. The model is optimized using the categorical cross-entropy loss function, and the class label with the highest probability is selected as the final prediction.

4 Model Evaluation

4.1 Our Dataset

To evaluate model performance, our dataset includes 77,620 abusive and 272,214 non-abusive text samples, integrated from three datasets. This diverse and imbalanced collection reflects real-world conditions, where abusive content is considerably less frequent than normal content. It enhances the model's generalization across different contexts, supports robust evaluation under realistic class distributions, and enables fair comparison with other imbalanced benchmark datasets. The combined datasets include:

1. Darkweb Dataset: We created the Darkweb dataset comprising 4,600 samples extracted from 352,000 dark web forum posts collected in 2022. It includes 2,500 child sexual abuse (CSA)-related and 2,100 non-CSA samples. Among them, 2,000 CSA and 100 non-CSA samples contain at least one sexual abuse phrase, see more details in [22]. Figure 4 presents blurred word clouds depicting single words and two-word phrases associated with sexual abuse, which were extracted from the post contents.
2. PAN12 Dataset[5]: Developed for the CLEF 2012 competition, this dataset aims to detect predatory behavior in online chats. It includes 198,054 conversations, consisting of 4,029 abusive and 194,025 non-abusive samples.
3. Roman Urdu Dataset: Contains 147,180 YouTube comments, evenly divided into abusive (73,590) and non-abusive (73,590) classes [2].

Fig. 4: Phrases related to sexual abuse on the dark web

4.2 Experimental Design and Metrics

To assess the model's performance, we use accuracy (ACC), precision (P), recall (R), and F1-score (F1), which are computed from the confusion matrix values: true positives (TP), false positives (FP), true negatives (TN), and false negatives (FN). These metrics are defined as follows: accuracy is computed by $ACC = \frac{TP+TN}{TP+FP+TN+FN}$, precision by $P = \frac{TP}{TP+FP}$, recall by $R = \frac{TP}{TP+FN}$, and the F1-score by $F1 = \frac{2 \cdot P \cdot R}{P+R}$ [27].

The AUC (Area Under the Curve) metric is also used to assess a model's overall classification performance. It represents the area beneath the ROC curve,

[5] https://pan.webis.de/clef12/pan12-web/

which graphs the True Positive Rate against the False Positive Rate at various threshold settings. Higher AUC values, approaching 1, signify greater classification accuracy and model effectiveness.

The experiments were carried out in a Kaggle Notebook environment featuring an NVIDIA Tesla T4 GPU with 16 GB of memory, of which up to 15 GB was available per session, alongside an Intel(R) Xeon(R) CPU @ 2.00 GHz with four logical cores. The setup used Python 3.10, with the PyTorch 2.5 framework and Hugging Face's Transformers library for implementing and fine-tuning the BERT-based model. The environment also provided up to 30 GB of RAM and 57.6 GB of temporary storage, which proved sufficient for training, validating, and evaluating the model's performance efficiently. We use 5-fold cross-validation during model development where the reported classification performance and execution times correspond to the averaged scores across the five runs.

4.3 Results and Discussion

The average Precision, F1-score, and AUC of the NB, LR, SVM, CNN, and LSTM models using TF-IDF ($tf.idf$) and Word2Vec ($w2v$) feature representations are presented in Figure 5, averaged across five-fold cross-validation. Overall, models leveraging Word2Vec embeddings generally outperform their TF-IDF counterparts across most evaluation metrics, reflecting the benefits of semantic representations in capturing contextual meaning. The only exception is the NB model, where $\text{NB}_{tf.idf}$ achieves higher Precision (0.696) and F1-score (0.783) than NB_{w2v}, although the latter records a slightly higher AUC (0.906 vs. 0.892), indicating a marginally better discriminative ability.

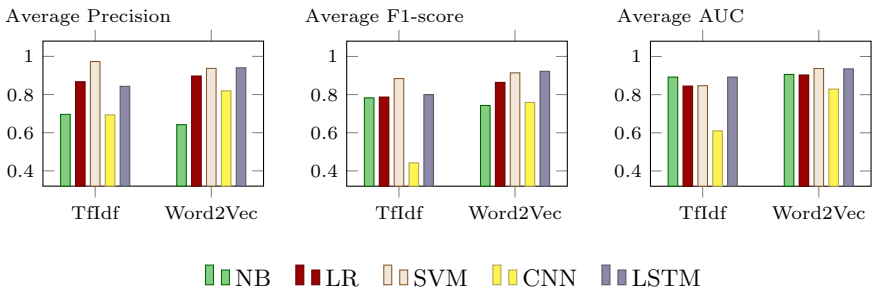
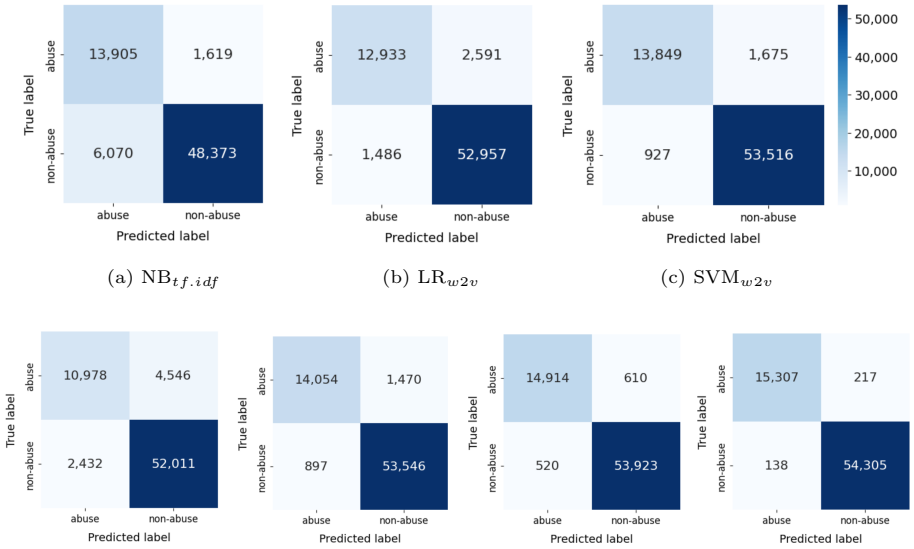


Fig. 5: Average Precision, F1-score and AUC of baseline ML and DL models

Among the remaining models, LR_{w2v}, SVM_{w2v}, CNN_{w2v}, and LSTM_{w2v} consistently achieve superior results. Specifically, LR_{w2v} outperforms LR_{tfidf} in Precision (0.897 vs. 0.867), F1-score (0.864 vs. 0.787), and Accuracy (0.903 vs. 0.845). SVM_{w2v} also shows improvements over SVM$tfidf$ in F1-score (0.914 vs. 0.884) and AUC (0.937 vs. 0.847). CNN_{w2v} also surpasses CNN_{tfidf} across all three evaluation metrics (Precision: 0.819 vs. 0.693, F1-score: 0.759 vs. 0.442, AUC: 0.829 vs. 0.610), demonstrating the advantage of semantic embeddings.

Similarly, LSTM_{w2v} achieves the best overall performance among all models, with a Precision of 0.94, F1-score of 0.922, and AUC of 0.935. These results highlight the effectiveness of DL and kernel-based models when combined with Word2Vec embeddings. Therefore, the best-performing models in this comparison are $\text{NB}_{tf.idf}$, LR_{w2v}, SVM_{w2v}, CNN_{w2v}, and LSTM_{w2v}.

Figure 6 presents and conpares the confusion matrix components across our proposed hybrid model, BERT and the the best-performing ML/DL models based on the $tf.idf$ or $w2v$. The proposed hybrid model achieved the highest number of TP (15,307) and TN (54,305), indicating its superior ability to correctly identify both abusive and non-abusive comments. It also recorded the lowest FP (138) and FN (217), demonstrating a strong balance between precision and recall. In contrast, traditional models such as NB and LR showed weaker performance, with significantly higher FP and FN. For example, $\text{NB}_{tf.idf}$ produced 6,070 FP and 1,619 FN, while CNN_{w2v} also struggled with 4,546 FN. BERT and LSTM_{w2v} performed considerably better, reducing misclassifications and achieving stronger overall accuracy; however, our model consistently outperformed them in all four metrics.



(a) $\text{NB}_{tf.idf}$ (b) LR_{w2v} (c) SVM_{w2v}

(d) CNN_{w2v} (e) LSTM_{w2v} (f) BERT (g) Our Model

Fig. 6: Confusion Matrices for models

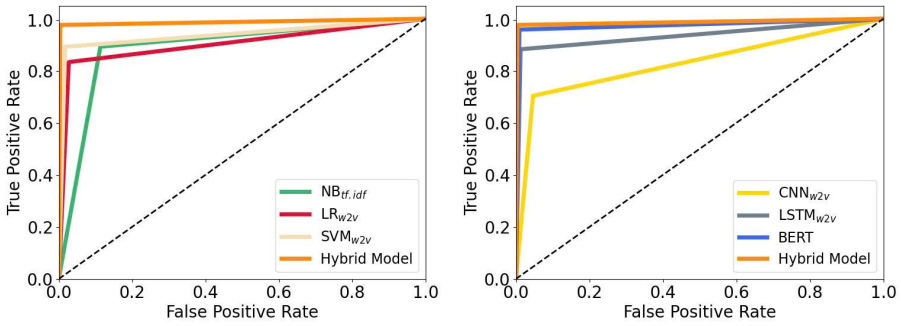
Table 1 summarizes the experimental results across all models. The proposed hybrid model achieved the best overall performance, with the highest Precision (0.991), Recall (0.986), Accuracy (0.995), F1-score (0.989), and AUC (0.992). This indicates its strong ability to correctly identify both abusive and non-abusive comments with minimal errors. BERT and LSTM_{w2v} also performed competitively, achieving high precision and recall but slightly lower accuracy

and F1-score compared to the hybrid model. Traditional ML models, including $NB_{tf.idf}$, LR_{w2v}, and SVM_{w2v}, delivered moderate performance, with NB showing the lowest precision (0.696) and CNN_{w2v} yielding the weakest overall results. In terms of computational efficiency, $NB_{tf.idf}$ and LR_{w2v} had the fastest training and prediction times, while DL models, particularly BERT and the hybrid model, required substantially longer processing times due to their complexity. Overall, the hybrid model offers the best trade-off between classification performance and reliability, albeit at the cost of increased computational time. These results highlight the effectiveness of integrating DL architectures in a hybrid framework to enhance both sensitivity and specificity in online abusive comment detection.

Table 1: Average model performance across 5-fold cross-validation

Classification	**Machine Learning**			**Deep Learning**			**Our Hybrid**
Performance	**$NB_{tf.idf}$**	**LR_{w2v}**	**SVM_{w2v}**	**CNN_{w2v}**	**$LSTM_{w2v}$**	**BERT**	**Model**
Average Precision	0.696	0.897	0.937	0.819	0.940	0.966	**0.991**
Average Recall	0.896	0.833	0.892	0.707	0.905	0.961	**0.986**
Average Accuracy	0.890	0.942	0.963	0.900	0.966	0.984	**0.995**
Average F1-score	0.783	0.864	0.914	0.759	0.922	0.964	**0.989**
Average AUC	0.892	0.903	0.937	0.829	0.935	0.976	**0.992**
Avg Training time (sec)	3.7	26.7	340.9	192.1	233.9	4,672.8	5,076.8
Avg Pred. time (sec)	1.75	0.03	42.65	1.80	1.38	203.60	459.78

Additionally, the Hybrid model also achieves the highest performance on the ROC curve, as shown in Figure 7, with an averaged AUC of 99.2%. The curve almost touches the top-left corner, indicating an optimal balance between the true positive rate and the false positive rate.



(a) ML models and our model (b) DL models and our model

Fig. 7: ROC curves for models

5 Conclusion and Future Work

This paper presents an efficient hybrid deep learning model that combines BERT, CNN, and LSTM architectures with a ReLU activation function to detect abusive language across multiple online platforms. By effectively integrating these components, the model captures semantic, contextual, and sequential dependencies in text, enabling robust classification of abusive and non-abusive content. Evaluation on a large, diverse, and imbalanced dataset—including YouTube comments, online forum posts, and dark web content—demonstrated its effectiveness under realistic conditions.

Experimental results showed that the proposed hybrid model outperformed traditional ML and standalone DL methods. It achieved strong results with a Precision of 0.991, Recall of 0.986, Accuracy of 0.995, F1-score of 0.989, and AUC of 0.992, confirming its high discriminative ability and robustness. Although computationally more demanding than simpler models, this trade-off remains acceptable for real-world applications requiring reliable performance. Overall, the hybrid model provides an effective and flexible approach to abuse detection, serving as a strong foundation for future research on content moderation and digital safety.

Future work will focus on extending this framework to multilingual and code-mixed text, improving interpretability through XAI methods, applying Ontology [25], [28], exploiting Knowledge Graph [26], [24] and incorporating multimodal [9] features such as images or user metadata. Additionally, optimizing computational efficiency via model compression or pruning will support real-time deployment in large-scale social media monitoring systems [10].

Acknowledgement

This publication has emanated from research conducted with the financial support of Taighde Éireann – Research Ireland under Grant number 12/RC/2289_P2.

References

1. Aguerri, J.C., Molnar, L., Miró-Llinares, F.: Old crimes reported in new bottles: the disclosure of child sexual abuse on twitter through the case# metooinceste. Social Network Analysis and Mining **13**(1), 27 (2023)
2. Akhter, M.P., Jiangbin, Z., Naqvi, I.R., Abdelmajeed, M., Sadiq, M.T.: Automatic detection of offensive language for urdu and roman urdu. vol. 8, pp. 91213 – 91226. IEEE (2020)
3. Alam, I., Basit, A., Ziar, R.A.: Utilizing age-adaptive deep learning approaches for detecting inappropriate video content. Human Behavior and Emerging Technologies **2024**(1), 7004031 (2024)
4. Amrit, C., Paauw, T., Aly, R., Lavric, M.: Identifying child abuse through text mining and machine learning. Expert Systems with Applications **88**, 402–418 (2017)
5. Chadaga, K., et al.: An explainable framework to predict child sexual abuse awareness in people using supervised machine learning models. Journal of Technology in Behavioral Science pp. 1–17 (2023)

6. Chinivar, S., et al.: Online offensive behaviour in socialmedia: Detection approaches, comprehensive review and future directions. Entertainment Computing **45**, 100544 (2023)
7. Cook, D., Zilka, M., DeSandre, H., Giles, S., Maskell, S.: Protecting children from online exploitation: Can a trained model detect harmful communication strategies? In: Proceedings of the 2023 AAAI/ACM Conference on AI, Ethics, and Society. p. 5–14. AIES '23, ACM (2023)
8. Dang, C.N., Moreno-García, M.N., De la Prieta, F.: Hybrid deep learning models for sentiment analysis. Complexity **2021**(1), 9986920 (2021)
9. Dao, P.Q., Roantree, M., Ngo, V.M.: Enhanced dual transformer contrastive network for multimodal sentiment analysis. In: Proceedings of the 17th International Conference on Management of Digital EcoSystems (MEDES'25). pp. 1–14. CCIS, Springer (Feb 2026)
10. Dao, P.Q., Ngo, V.M.: Exploring, investigating and exploiting sentiment analysis systems. Preprints.org pp. 1–23 (Nov 2025). https://doi.org/10.20944/preprints202510.0194
11. Gangwar, A., González-Castro, V., Alegre, E., Fidalgo, E.: Attm-cnn: Attention and metric learning based cnn for pornography, age and child sexual abuse (csa) detection in images. Neurocomputing **445**, 81–104 (2021)
12. Hole, M., Frank, R., Logos, K., Westlake, B., Michalski, D., Bright, D., Afana, E., Brewer, R., Ross, A., Swearingen, T., et al.: Developing automated methods to detect and match face and voice biometrics in child sexual abuse videos. Trends and Issues in Crime and Criminal Justice (648), 1–15 (2022)
13. Kaur, S., Singh, S., Kaushal, S.: Deep learning-based approaches for abusive content detection and classification for multi-class online user-generated data. International Journal of Cognitive Computing in Engineering **5**, 104–122 (2024)
14. Kissos, L., Goldner, L., Butman, M., Eliyahu, N., Lev-Wiesel, R.: Can artificial intelligence achieve human-level performance? a pilot study of childhood sexual abuse detection in self-figure drawings. Child Abuse & Neglect **109**, 104755 (2020)
15. Laranjeira, C., Macedo, J., Avila, S., Santos, J.: Seeing without looking: Analysis pipeline for child sexual abuse datasets. In: Proc. of 2022 ACM Conf. on Fairness, Accountability, and Transparency (FAccT'22). p. 2189–2205. ACM (2022)
16. Li, T., Zeng, Z., Sun, S.: A two-stage cyberbullying detection based on multi-view features and decision fusion strategy. Applied Intelligence **55**(4), 294 (2025)
17. Marshan, A., Nizar, F.N.M., Ioannou, A., Spanaki, K.: Comparing machine learning and deep learning techniques for text analytics: Detecting the severity of hate comments online. Information Systems Frontiers (2023)
18. Mazzarello, O., Gagné, M.E., Langevin, R.: Risk factors for sexual revictimization and dating violence in young adults with a history of child sexual abuse. Journal of Child & Adolescent Trauma **15**(4), 1113–1125 (2022)
19. Mckeever, S., Thorpe, C., Ngo, V.M.: Determining child sexual abuse posts based on artificial intelligence. In: 2023 International Society for the Prevention of Child Abuse & Neglect Congress (ISPCAN-2023). pp. 24–27. Edinburgh, UK (2023)
20. Mosa, M.A.: Optimizing text classification accuracy: a hybrid strategy incorporating enhanced nsga-ii and xgboost techniques for feature selection. Progress in Artificial Intelligence (2025)
21. Muneer, A., Fati, S.M.: A comparative analysis of machine learning techniques for cyberbullying detection on twitter. Future Internet **12**(11) (2020)
22. Ngo, V.M., Mckeever, S., Thorpe, C.: Identifying online child sexual texts in dark web through machine learning and deep learning algorithms. In: the APWG.EU

Technical Summit and Researchers Sync-Up (APWG.EU-Tech 2023). pp. 1–6. CEUR Workshop Proceedings (2024)
23. Ngo, V.M., Thorpe, C., McKeever, S.: Analysing child sexual abuse activities in the dark web based on an efficient csam detection algorithm. In: The 2nd Annual Trust and Safety Research Conference. Stanford, USA (September 2023)
24. Ngo, V.M.: Discovering latent information by spreading activation algorithm for document retrieval. International Journal of Artificial Intelligence & Applications **5**(1), 23–34 (2014)
25. Ngo, V.M., Cao, T.H.: A generalized vector space model for ontology-based information retrieval. Vietnamese Journal on Information Technologies and Communications **22**(2), 43–53 (2009)
26. Ngo, V.M., Cao, T.H.: Discovering latent concepts and exploiting ontological features for semantic text search. In: Proceedings of the 5th Int. Joint Conf. on Natural Language Processing (IJCNLP 2011). pp. 571–579. ACL (2011)
27. Ngo, V.M., Cao, T.H.: Semantic search by latent ontological features. New Generation Computing **30**(1), 53–71 (2012)
28. Ngo, V.M., Cao, T.H., Le, T.: Wordnet-based information retrieval using common hypernyms and combined features. In: Proceedings of the 5th International Conference on Intelligent Computing and Information Systems (ICICIS 2011). pp. 1–6. ACM (2011)
29. Ngo, V.M., Gajula, R., Thorpe, C., Mckeever, S.: Discovering child sexual abuse material creators' behaviors and preferences on the dark web. Child Abuse & Neglect **147**, 106558 (2024)
30. Ngo, V.M., Thorpe, C., Dang, C.N., Mckeever, S.: Investigation, detection and prevention of online child sexual abuse materials: A comprehensive survey. In: 2022 RIVF Int. Conf. on Computing and Communication Technologies (RIVF). pp. 707–713 (2022)
31. Owusu-Addo, E., et al.: Prevalence and determinants of sexual abuse among adolescent girls during the covid-19 lockdown and school closures in ghana: A mixed method study. Child Abuse & Neglect **135**, 105997 (2023)
32. Palangi, H., et al.: Deep sentence embedding using long short-term memory networks: Analysis and application to information retrieval. IEEE/ACM transactions on audio, speech, and language processing **24**(4), 694–707 (2016)
33. Petrosyan, A.: Impact of online hate and harassment in the u.s. 2020 (2025), https://www.statista.com/statistics/971876/societal-impact-of-online-hate-harassment-usa/, accessed: 2025-09-02
34. Puentes, J., et al.: Guarding the guardians: Automated analysis of online child sexual abuse. In: 2023 IEEE/CVF International Conference on Computer Vision Workshops (ICCVW). pp. 3730–3734 (2023)
35. Talpur, B.A., O'Sullivan, D.: Cyberbullying severity detection: A machine learning approach. PLOS ONE **15**(10), 1–19 (2020)
36. Wadud, M.A.H., et al.: How can we manage offensive text in social media - a text classification approach using lstm-boost. International Journal of Information Management Data Insights **2**(2), 100095 (2022)
37. Yamashita, R., Nishio, M., Do, R.K.G., Togashi, K.: Convolutional neural networks: an overview and application in radiology. Insights into imaging **9**, 611–629 (2018)
38. Zuckerman, A.: 60 cyberbullying statistics: 2020/2021 data, insights & predictions, https://comparecamp.com/cyberbullying-statistics, accessed: 2025-04-02