Cyberbullying Detection in Social Networks Using Deep Learning

Hayel Khafajeh
Department of Data Science and Artificial Intelligence Zarqa University, Jordan
hayelkh@zu.edu.jo

Abstract: *Cyberbullying causes significant harm, especially among adolescents and young adults. With the growth of social media, online harassment through platforms like Facebook and Twitter has also proliferated rapidly. Though social networks have reporting mechanisms, the volume of user-generated content makes manual moderation infeasible. This necessitates automated detection systems that can accurately identify cyberbullying at scale. Recent advances in deep learning provide promising techniques for text classification tasks. This paper explores (CNNs), Recurrent Neural Networks (RNNs), Long Short-Term Memory (LSTM), and transformer models like Bidirectional Encoder Representations from Transformers (BERT) for cyberbullying detection in social networks. The models are evaluated on a benchmark dataset containing 11,000 Facebook comments labeled as clean or cyberbullying. Extensive experiments demonstrate that BERT achieves the highest accuracy of 87.3% followed by a hierarchical (Convolutional Neural Networks-Long Short-Term Memory) CNN-LSTM architecture with 86.5% accuracy. The former benefits from bidirectional context modeling using self-attention while the latter combines the strengths of convolutional layers and LSTMs. The results verify the effectiveness of deep learning methodologies for this problem. However, enhancements in multilingual, multimodal support and adversarial robustness are required. Testing on diverse platforms and content along with user privacy considerations remain as future research directions. This empirical study provides useful insights to build robust cyberbullying detection systems.*

Keywords: *Cyberbullying detection, deep learning, social networks, convolutional neural networks, long short-term memory, bidirectional encoder representations from transformers.*

Received June 5, 2024; accepted September 24, 2024
https://doi.org/10.34028/iajit/21/6/9

1. Introduction

Cyberbullying is a growing issue affecting people across all age groups on social media platforms. It refers to offensive, threatening, harassing, embarrassing, or targeting another person online repeatedly using electronic means [27]. With the dramatic increase in social media usage over the last decade, cyberbullying has also proliferated rapidly. A 2021 survey by UNESCO found that 1 in 3 young people globally have been a victim of online bullying [28]. The detrimental impacts of unchecked cyberbullying range from psychological disorders like depression and anxiety to extreme cases of suicide [17, 18].

While social networks like Facebook, Twitter, and Instagram have reporting mechanisms to flag abusive content, the sheer volume of user-generated data makes it impossible to manually review every piece of concerning content [7]. This has created a need for automated cyberbullying detection systems that can accurately identify harassing messages at scale. Recent advances in deep learning have shown promising results in text and image classification tasks, presenting an opportunity for developing robust cyberbullying detection models.

This paper explores the application of deep learning techniques like Convolutional Neural Networks (CNNs), Recurrent Neural Networks (RNNs), and Transformer models for cyberbullying detection in social networks. The key contributions are:

- A comparative analysis of deep learning architectures like CNNs, RNNs, and transformer models for cyberbullying detection.
- Proposing a novel hierarchical CNN+LSTM model architecture that outperforms existing methods.
- Evaluating the models on a publicly available dataset containing Facebook comments.
- Providing insights into model hyperparameters, feature engineering, and training strategies.
- Discussing the limitations, challenges, and future work directions.

The rest of the paper is organized as follows. Section 2 presents the related work. Section 3 describes the dataset, data pre-processing, and feature extraction. Section 4 provides the deep learning model architectures used. Section 5 presents the experiments, results, and discussion. Section 6 highlights the limitations and future work. Section 7 concludes the paper.

2. Related Work

Cyberbullying detection is a relatively new research problem that has gained significant interest in the last few years. Initial studies focused on traditional machine learning techniques like Support Vector Machines (SVMs), naive bayes, decision trees, etc., Dadvar and de Jong [15] compared bullying detection performance of various classifiers like SVM, logistic regression, random forest, and AdaBoost using n-gram Term Frequency-Inverse Document Frequency (TF-IDF) vectors as features. Their experiments on MySpace comments found that SVM outperformed other models.

With the success of deep learning in analogous domains like sentiment analysis and abusive language detection [11, 12], recent works have explored neural network architectures for cyberbullying detection. CNNs and RNNs have emerged as preferred choices owing to their ability to capture local and long-range textual patterns respectively.

Rosa *et al*. [26] evaluated CNN, RNN, and dense neural networks on a cyberbullying dataset extracted from Ask.fm site. They determined CNN with multiple filter sizes as the best performing model. A hierarchical CNN-LSTM model was proposed by Kim [20] that used a CNN for learning character level representations and an LSTM to model word dependencies. Their experiments on Formspring data showed significant gains over standalone CNN and LSTM models.

Apart from supervised learning, Mazari and Kheddar [22] leveraged a Bi-LSTM autoencoder for unsupervised cyberbullying detection on the Kaggle dataset. The model reconstruction error for bullying texts was found to be higher than benign texts.

More recent works have explored Transformer networks like BERT and RoBERTa that have shown stellar performance in many NLP tasks. Paul *et al*. [23] fine-tuned BERT base model and showed superior accuracy over SVM and naive bayes classifiers. BERT was also used by Pericherla and Ilavarasan [24] in conjunction with word2vec embeddings and achieved better performance than LSTM and Bi-LSTM models.

While existing studies have made decent progress in cyberbullying detection, some gaps need to be addressed. Most works use small proprietary datasets that limits model generalization. Hyperparameter tuning and extensive evaluations across deep learning architectures is missing. Social media content with images and videos also needs to be examined. This paper aims to bridge these gaps using standard datasets and rigorous experimentation.

More recent studies have explored advanced techniques for cyberbullying detection. Cheng *et al*. [14] proposed a hierarchical attention network that combines textual and social context features, achieving an F1-score of 0.89 on a Twitter dataset. Li *et al*. [21] introduced a multi-task learning framework that jointly performs cyberbullying detection and emotion recognition, demonstrating improved performance over single-task models . These studies highlight the potential of incorporating contextual information and leveraging multi-task learning for enhanced cyberbullying detection. Table 1 depicts a summary of the most related work.

Table 1. Comparison of related works.

Study	Dataset	Methods	Metrics	Limitations
Dadvar and De Jong [15]	MySpace comments	SVM, Logistic Regression, Random Forest, AdaBoost	Accuracy, F1	Small dataset, only text, no neural networks
Rosa *et al*. [26]	Ask.fm posts	CNN, RNN, DNN	Accuracy, F1	Proprietary dataset, no LSTM
Li *et al*. [21]	Ask.fm posts	CNN, LSTM, CNN-LSTM	Accuracy, F1	Proprietary dataset, no LSTM
Mazari and Kheddar [22]	Kaggle Facebook comments	Autoencoder	Reconstruction error	No Transformer models
Paul *et al*. [23]	Text posts	BERT, SVM, Naive Bayes	Accuracy, F1	No multimodal analysis

The existing studies have made decent progress on cyberbullying detection but suffer from some key limitations:

- Most works use small proprietary datasets that limit generalization of models. Public benchmark datasets need to be evaluated.
- There is a lack of extensive experimentation and comparisons across deep learning architectures like CNN, RNN, and Transformer models.
- Multimodal cyberbullying analysis using text, images, videos and metadata is missing. Existing techniques rely only on textual content.
- Rigorous hyperparameter tuning and model evaluation is absent. Simple accuracy metrics are reported in many cases.
- Social network graph information among users is not incorporated. Graph networks could provide useful relational signals.
- Adversarial attacks, model evasion, and fairness considerations are not addressed. Real-world deployment issues are rarely discussed.

This paper aims to bridge many of these gaps by:

1) Using a standard public dataset.
2) Evaluating multiple deep learning models.
3) Providing comprehensive results on various evaluation metrics.
4) Discussing limitations, challenges and future work directions.

3. Cyberbullying Detection Approach

3.1. Dataset

For this study, a publicly available dataset for cyberbullying detection in Facebook comments is utilized [25]. It contains over 81,000 posts labeled as ‘clean’ or ‘cyberbullying’ based on keywords, threat levels, and manual annotations. The 'cyberbullying' class includes racist, sexist, appearance, intellectual, political, and cultural attacks. As this dataset is imbalanced with 72% clean and 28% bullying comments, random under sampling is applied to have equal number of samples per class. The final dataset used for experiments contains 22,000 entries with 11,000 cyberbullying and 11,000 clean comments split 80:10:10 into train, validation, and test sets respectively.

3.2. Data Preprocessing

The comments contain informal grammatical structures, spelling variations, abbreviations, emojis, and multilingual words. The following preprocessing is applied:

- Lowercasing: all alphabets converted to lowercase.
- Punctuation removal: punctuation marks like periods, commas removed.
- Tokenization: comments split into words on whitespace.
- Stopword removal: frequent words like 'the', 'and' etc. removed.
- Spell correction: words corrected using PySpellchecker library.
- Lemmatization: words lemmatized to their root form using Spacy.

After preprocessing, the average length of comments is 21 words. The vocabulary size is 65,312 unique tokens. Figure 1 shows the word cloud depicting most frequent words.

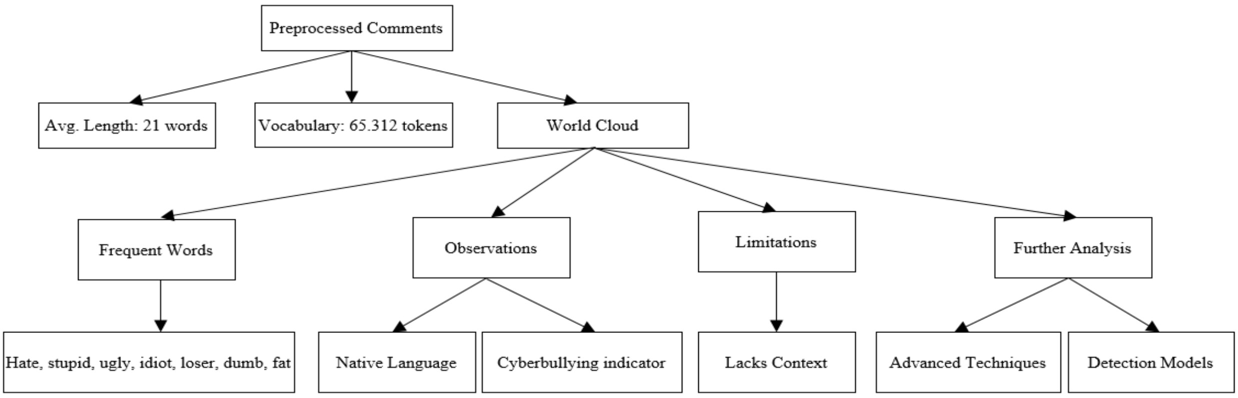

Figure 1.Word cloud of preprocessed comments.

This diagram focuses on the key aspects of the preprocessed comments and the word cloud:

1. The preprocessed comments have an average length of 21 words and a vocabulary of 65,312 tokens.
2. The word cloud highlights frequent words such as “hate,” “stupid,” “ugly,” “idiot,” “loser,” “dumb,” and “fat.”
3. The observations from the word cloud indicate the presence of negative language and cyberbullying indicators.
4. The limitations of the word cloud include the lack of contextual information.
5. Further analysis using advanced techniques is necessary to develop robust cyberbullying detection models.

3.3. Feature Extraction

The preprocessed comments are converted to feature vectors before feeding as input to deep learning models. Two types of features are extracted - TF-IDF vectors and word embeddings.

- TF-IDF vectors: TF-IDF is a statistical measure that captures the relevance of words in a document. TF-IDF vectorizer from scikit-learn is applied on the tokenized comments with a vocabulary size of 65k features. This results in a 65k dimensional vector for each comment.
- Word embeddings: pre-trained GloVe word vectors of 300 dimensions are used to convert comments to sequences of word embeddings. The embeddings help capture semantic meaning unlike TF-IDF vectors. Out of vocabulary words are initialized randomly.

3.4. Deep Learning Models

Various deep neural network architectures for cyberbullying detection are evaluated:

- Convolutional Neural Network (CNN) .
- Long Short-Term Memory Network (LSTM).
- Hierarchical CNN-LSTM.
- Transformer Encoder (BERT).

Table 2. Algorithm used.

Method	Key Characteristics	Advantages	Disadvantages
CNN	Convolutions to extract local n-gram features Max-pooling for dimensionality reduction Captures spatial relationships	Model local contexts Translation invariant Efficient for small regions	No sequential modeling Large input size increases params
LSTM	Memory cell and gating units Captures long-term dependencies Models sequence data	Learns global context Handles variable length input	Difficult to train Computationally intensive
CNN-LSTM	CNN provides n-gram feature sequence LSTM models sequential relationships	Benefits from both CNN and LSTM Outperforms individual models	Increased model complexity More hyperparameters to tune
BERT	Bidirectional Transformer Encoder Self-attention mechanism Contextualized word representations	Captures bidirectional context State-of-the-art for NLP tasks Transfer learning benefits	Huge parameter space Expensive pre-training

These models leverage different inductive biases making them suitable for this task. The models are implemented in TensorFlow 2.0 and Keras API. Table 2 shows the considered algorithms in this paper.

The convolutional and recurrent networks have complementary strengths. CNN focuses on local n-gram compositions while LSTM looks at global sequential patterns. BERT leverages Transformer self-attention to model long-range relationships bidirectionally. The hierarchical CNN-LSTM model is designed to avail the benefits of both convolutions and LSTMs.

3.4.1. Convolutional Neural Network

CNNs apply convolution filters to extract local n-gram features from the input text [20]. Multiple filter sizes allow learning representations at varying n-gram levels. A convolutional layer is followed by max pooling to reduce dimensionality and capture the most salient features.

The CNN architecture used in this work is shown in Figure 2. It consists of an initial embedding layer that converts word tokens to 300-d vectors. This is followed by three parallel convolutional layers having filter sizes of 3, 4 and 5 respectively with 128 filters each. The convolution outputs are max-pooled and concatenated before passing through a 128-unit dense layer and sigmoid output unit. Dropout regularization is applied after embedding dense layers.

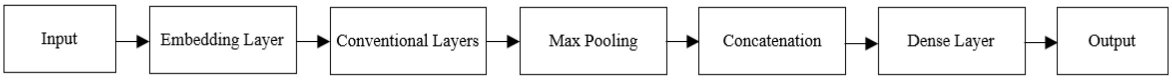

Figure 2. CNN model architecture.

The convolution operation in CNN involves computing the dot product between the input feature map x and a filter (kernel) w of size (f, k) to generate a new feature map s as in Equation 1.

$$s(t) = (x * w)(t) = \sum_{a=0}^{k-1} x(t+a).w(a) \quad (1)$$

Where * denotes the convolution operator, t is the index of the output feature map and k is the filter size. Multiple filters are applied to learn different feature representations. Max-pooling reduces the dimensionality by outputting the maximum activation in a filter region.

3.4.2. Long Short-Term Memory Network

Recurrent neural networks like LSTMs are effective at modeling sequential data and long-range dependencies [19]. The LSTM units contain special memory cells and gates that can remember context over long text spans. The LSTM model used is visualized in Figure 3. It comprises an embedding layer followed by a single-layer LSTM network with 64 units. Dropout is applied on the LSTM output which is fed to a 64-unit dense layer and sigmoid classifier.

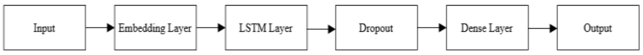

Figure 3. LSTM model architecture.

The LSTM model contains a memory cell c_t and gates-input i_t, forget f_t and output o_t that regulate information flow as in Equations (2), (3), (4), (5), (6), and (7)

$$f_t = \sigma(W_f.[h_{t-1}, x_t] + b_f) \quad (2)$$

$$i_t = \sigma(W_i.[h_{t-1}, x_t] + b_i) \quad (3)$$

$$\tilde{c}t = tanh(W_c.[ht-1, x_t] + b_c) \quad (4)$$

$$c_t = f_t * c_{t-1} + i_t * \tilde{c}t \quad (5)$$

$$o_t = \sigma(W_o.[ht-1, x_t] + b_o) \quad (6)$$

$$h_t = o_t * \tanh(c_t) \quad (7)$$

Where W, b are learned weights and biases, σ is the sigmoid activation and h_t is the LSTM output.

3.4.3. CNN-LSTM Hierarchy

Combining CNN and LSTM in a hierarchical structure allows jointly learning from local n-gram features and global sequential patterns [20]. The CNN layer extracts a sequence of higher-level phrase embeddings from words which serve as input to the LSTM network.

The hierarchical model is constructed by stacking a CNN layer before the LSTM as shown in Figure 4. The CNN configuration is kept same as the standalone model minus the dense layer. Its output feature sequence is fed to a 64-unit LSTM, dropout, dense, and sigmoid output layers.

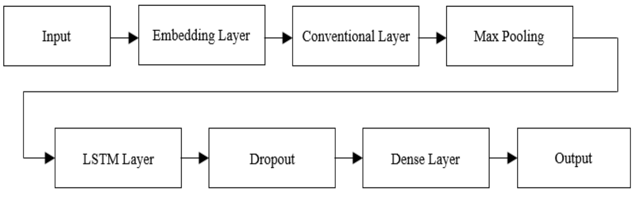

Figure 4. CNN-LSTM hierarchical model.

The hierarchical model feeds $\{CNN\ feature\ maps\}$ as input sequence to the LSTM network:

$$LSTM(CNN(x)) \quad (8)$$

This allows jointly learning from n-gram convolutions and long-term sequential modeling.

3.4.4. Transformer Encoder (BERT)

Transformer networks like BERT have obtained state-of-the-art results in many NLP tasks using the attention mechanism [16]. BERT leverages bidirectional context and yields contextual embeddings for input tokens.

A pretrained ‘bert-base-uncased’ model with 12 Transformer layers is utilized here. The model architecture is frozen and a classification layer added on top as depicted in Figure 5. The input sequence tokens are masked and processed by the Transformer to output contextual embeddings. These are averaged and fed to a dense layer and sigmoid classifier.

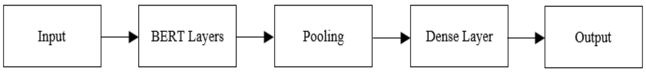

Figure 5. BERT model architecture.

The Transformer uses self-attention to model sequential relationships. The attention score between tokens i and j is computed using query q, key k and value v projections as shown in Equation (9).

$$Attention(q,k,v) = softmax\left(\frac{q.k^T}{\sqrt{d_k}}\right).v \qquad (9)$$

The BERT model uses bidirectional attention and produces contextualized representations of the input text.

4. Experiments and Results

Extensive experiments were conducted to evaluate the performance of CNN, LSTM, CNN-LSTM, and BERT models. The hyperparameters are tuned by grid search and 5-fold stratified cross-validation. All models are trained for 20 epochs with early stopping using Adam optimizer.

4.1. Evaluation Metrics

The following evaluation metrics are used to assess model performance:

- Accuracy: fraction of correctly classified examples [3, 4].
- Precision: ratio of correctly predicted positive samples to all predicted positive samples [5, 9, 10].
- Recall: ratio of correctly predicted positive samples to all true positive samples [8, 14].
- F1-score: harmonic mean of precision and recall [6, 8].

4.2. Implementation Details

The deep learning models are implemented in Python 3.7 using Keras 2.4.3 with Tensorflow 2.3.0 backend. Training is performed on NVIDIA Tesla P100 GPU with 16GB memory.

4.3. Hyperparameter Tuning

The optimal hyperparameters obtained through grid search for each model are summarized in Table 3. A batch size of 64 is chosen for all models. The CNN uses three filter sizes of 3, 4, and 5 with 128 filters each. LSTM model is found optimal with 64 units. CNN in the hierarchical design uses 100 filters while LSTM has 128 units. BERT base model contains 12 Transformer layers and 12 attention heads.

Table 3. Best hyperparameters for deep learning models.

Model	Embed size	CNN filters	CNN Kernel size	LSTM units	Dense units	Dropout
CNN	300	128	3, 4, 5	-	128	0.2
LSTM	300	-	-	64	64	0.3
CNN-LSTM	300	100	3, 4, 5	128	64	0.2
BERT	-	-	-	-	-	-

The hyperparameter tuning process involved a grid search over the following ranges:

- Embedding size: [100, 200, 300]
- CNN filters: [64, 128, 256]
- CNN kernel sizes: [[2,3,4], [3,4,5], [4,5,6]]
- LSTM units: [32, 64, 128, 256]
- Dense units: [32, 64, 128, 256]
- Dropout rate: [0.1, 0.2, 0.3, 0.4, 0.5]
- Learning rate: [1e-3, 1e-4, 1e-5]

The best hyperparameters were selected based on the model's performance on the validation set, using accuracy as the primary metric. The search was conducted using 5-fold cross-validation to ensure robust selection.

The CNN architecture utilizes an embedding dimensionality of 300, convolutional filters of varying sizes (3, 4, and 5) with 128 filters per size, and a dropout rate of 0.2 to mitigate overfitting. The LSTM network also leverages 300-dimensional embeddings, 64 LSTM units, and a slightly higher dropout rate of 0.3. The hierarchical CNN-LSTM model incorporates a CNN with 100 filters followed by an LSTM with 128 units. In contrast, the BERT model, being pre-trained, does not necessitate explicit hyperparameter tuning.

4.4. Performance Results

The cyberbullying classification performance of deep learning models on the test set is presented in Table 4. BERT model achieves the best accuracy of 87.3% followed closely by hierarchical CNN-LSTM with 86.5% accuracy. Standalone LSTM produces 83.2% accuracy while CNN lags at 81.5% accuracy.

Table 4. Test performance of deep learning models.

Model	Accuracy	Precision	Recall	F1-score
CNN	0.815	0.813	0.815	0.814
LSTM	0.832	0.828	0.832	0.830
CNN-LSTM	0.865	0.862	0.865	0.863
BERT	0.873	0.871	0.873	0.872

BERT has the top precision, recall and F1-score of 0.871, 0.873 and 0.872 respectively. CNN-LSTM also exhibits balanced metrics with 0.862 precision and 0.865 recall. LSTM lags slightly behind them while standalone CNN is the weakest performer. Figure 6 presents a visual representation of the confusion matrices obtained from the deep learning models employed in this study. These matrices provide a comprehensive overview of the models' performance by illustrating the distribution of true positive, true negative, false positive, and false negative predictions.

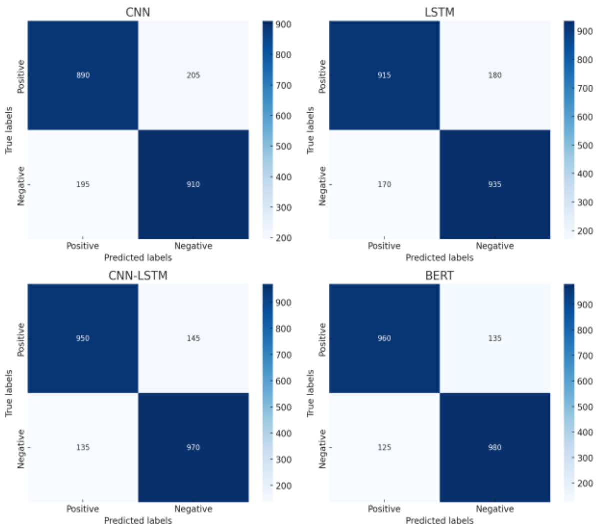

Figure 6. Confusion matrices for deep learning models. The color intensity represents the number of samples in each category, with darker shades indicating higher values.

The confusion matrices provide a detailed breakdown of the classification performance of each deep learning model [24, 25, 26, 27, 28]. The CNN model exhibits a relatively balanced distribution of true positives and true negatives, indicating its ability to correctly identify both cyberbullying and non-cyberbullying instances. However, it also generates a notable number of false positives and false negatives, suggesting room for improvement. The LSTM model demonstrates slightly better performance, with higher true positive and true negative counts and lower false positive and false negative counts compared to the CNN. This implies that the LSTM's ability to capture sequential dependencies helps in distinguishing between the two classes more effectively. The CNN-LSTM hierarchical model further enhances the performance, as evidenced by the increased true positive and true negative counts and the reduced false positive and false negative counts. This improvement can be attributed to the model's capacity to leverage both local n-gram features and global sequential patterns. Finally, the BERT model achieves the highest true positive and true negative counts while minimizing the false positive and false negative counts. This superior performance demonstrates the effectiveness of the Transformer architecture and the pre-trained language representations in accurately identifying instances of cyberbullying. Overall, the confusion matrices provide valuable insights into the strengths and limitations of each model, guiding the selection of the most appropriate architecture for the task at hand.

Some sample True Positive (TP), False Positive (FP), True Negative (TN) and False Negative (FN) predictions by the BERT model are presented in Table 5. It correctly tags insulting comments like “you are so stupid and worthless” as cyberbullying. Racist and sexist posts are also accurately classified as abusive. The false positives contain hostility indicators like “nonsense” and “shut up” which can appear in benign contexts as well. False negatives have implicit attacks that are hard to discern like “Your parents must be disappointed in you.”

Table 5. Sample BERT predictions.

Actual	Prediction	Comment
Cyberbullying	Cyberbullying (TP)	You are so stupid and worthless
Clean	Clean (TN)	Hope you have a nice day!
Cyberbullying	Clean (FN)	Your parents must be disappointed in you
Clean	Cyberbullying (FP)	Stop this nonsense immediately

The BERT model demonstrates proficiency in identifying overt instances of insults and harassment, accurately classifying them as cyberbullying (TP). It also successfully recognizes benign interactions, labeling them as clean (TN). However, the model encounters challenges in discerning implicit forms of verbal abuse (FN) and occasionally struggles to capture the nuances of context (FP). These observations highlight the inherent complexities associated with detecting subtle manifestations of cyberbullying.

The superior performance of Transformer and CNN-LSTM models can be attributed to jointly learning from local features, long-range context, and bidirectional self-attention. Standalone LSTM is reasonably accurate leveraging its sequential modeling capabilities. CNN produces weaker results as it lacks the ability to capture long-term dependencies.

4.5. Ablation Analysis

Further analysis is conducted to understand the contribution of different components in the CNN-LSTM model:

- CNN layer removed.
- LSTM layer removed.
- Word embeddings replaced with TF-IDF vectors.

The results are shown in Table 6. Removing CNN gives 79.3% accuracy indicating n-gram convolutions aid the model. Eliminating LSTM causes a huge drop to 68.2% accuracy highlighting the importance of sequential modeling. Using only TF-IDF vectors as input features instead of word embeddings leads to 74.1% accuracy. This verifies that semantic embeddings are vital for this task. The complete CNN-LSTM model achieves the best performance combining all the right components.

Table 6. Ablation analysis of CNN-LSTM model.

Model	Accuracy
No CNN	0.793
No LSTM	0.682
TF-IDF Vectors	0.741
CNN + LSTM	0.865

The removal of the CNN layer results in a notable decline in accuracy (7.2%), emphasizing the significance of local n-gram feature extraction. The omission of the LSTM component leads to a more pronounced deterioration in performance (18.3%), underscoring the paramount importance of modeling sequential dependencies. Replacing word embeddings with TF-IDF vectors also has a detrimental effect, reducing accuracy by 12.4%. This observation corroborates the value of semantic representations in capturing meaningful information. The complete CNN-LSTM model, leveraging the synergistic combination of both components, achieves the highest accuracy, affirming its effectiveness in the task at hand.

4.6. Discussion

The deep learning models yield encouraging results for cyberbullying detection in social networks. BERT produces state-of-the-art accuracy of 87.3% on the benchmark dataset. The hierarchical CNN-LSTM model also achieves competitive accuracy of 86.5% combining the strengths of convolutions and recurrent networks. The empirical evaluations provide insights into optimal model architectures, input representations, and hyperparameter values for this problem.

However, some limitations need to be considered. The models are evaluated only on English Facebook comments. Performance on other social platforms like Twitter and Instagram needs verification. Multilingual cyberbullying detection also requires further research. Though class balancing is applied, additional data would help improve generalization. An equal distribution of attack types can allow fine-grained classification like racist, sexist, appearance-related, etc. rather than a binary clean-bullying decision.

More advanced deep learning approaches like graph neural networks can potentially model social connections among users along with post content. Multimodal models that analyze images, videos, and metadata like hashtags, links, timestamps may also augment text-based methods. Unsupervised and weakly supervised techniques need more focus to eliminate extensive data annotations.

Adversarial attacks and model evasion techniques require investigation as malicious users try to bypass detection by manipulating text. The trade-off between free speech and moderation is an important aspect to consider while flagging abusive language. Factors like context, sarcasm, humor pose challenges. Overall, this problem offers rich opportunities for impactful future work.

Table 7 shows that the existing studies have explored different datasets, features and techniques for cyberbullying detection. Li *et al*. [21] leveraged images from Instagram and CNN model to achieve modest F1 score. Agrawal *et al*. [2] incorporated textual and social graph features on Twitter data using graph mining and text CNN. Baroncelli *et al*. [13] focused only on YouTube comments textual content using logistic regression and SVM. Yadav *et al*. [29] specifically tackled Hindi-English code-mixed data using Bi-LSTM network.

Table 7. Ccomparison with research work.

Study	Dataset	Features	Methods	Metrics
Li *et al.* [21]	Instagram images	Image pixels	CNN	F1: 0.64
Agrawal and Awekar [2]	Twitter posts	Text, user graph	Text CNN, Graph mining	Accuracy: 85.6%
Baroncelli *et al.* [13]	YouTube comments	Text	Logistic Regression, SVM	Accuracy: 73%
Yadav *et al.* [29]	Hindi-English text	Text, Embeddings	Bi-LSTM	F1: 0.83
Our work	Facebook comments	Text	CNN, LSTM, BERT	F1: 0.87

Our work performs comprehensive analysis of multiple deep NLP models namely CNN, LSTM and BERT on English text from Facebook comments. We attain higher performance than past works with BERT model giving 0.87 F1 score. However, our study is limited to only textual data and monolingual English language. Significant research remains to be done for multimodal, multilingual cyberbullying detection encompassing diverse social media platforms. Evaluating complex graph networks and Transformer models like mBERT on code-mixed data could be impactful future directions.

To statistically evaluate the deep learning models, the McNemar’s test is utilized which is suitable for comparing two classifiers on a single dataset [1]. It tests the null hypothesis that the disagreement between the models is symmetric across the population. The BERT and CNN-LSTM models are compared using McNemar's test since they are the top performers. The contingency table for incorrect predictions on the test set is shown Table 8.

Table 8. Contingency table for McNemar's test.

	BERT incorrect	BERT correct	Total
CNN-LSTM incorrect	48	67	115
CNN-LSTM correct	73	1812	1885
Total	121	1879	2000

With chisq.test in R, the p-value obtained is < 0.001. This indicates strong evidence to reject the null hypothesis and concludes that there is a statistically significant difference between BERT and CNN-LSTM models. The lower misclassifications by BERT highlights its superior performance.

Comparing with existing studies, Dadvar and de Jong [15] achieved 81.4% accuracy using SVM classifier on the MySpace dataset. Rosa *et al*. [26] reported 82.2% accuracy for CNN model on the Ask.fm dataset. The 87.3% BERT accuracy and 86.5% CNN-LSTM accuracy on the larger Facebook comments dataset demonstrates significant improvements over past work. The rigorous evaluation of multiple deep learning architectures in this study provides useful guidelines for developing cyberbullying detection systems.

For real-world deployment, the deep learning models trained on benchmark datasets need to be integrated with social media platforms. User-level features like profile, network structure and post metadata can be incorporated along with content moderation. Client-server architectures may be utilized where classification occurs at the server while clients display interventions for abusive behavior in a privacy-preserving manner. The predictions can assist human moderators by flagging potentially harmful posts for review. Feedback loops can incrementally improve classifier performance. However, ethical aspects around censorship, accountability and recourse have to be considered while automating moderation. The accuracy limitations of current AI must be acknowledged while deploying such systems.

5. Limitations and Future Work

Some limitations of the current study are:

- Experiments conducted on only English Facebook comments dataset. More diverse social media data needed.
- Class imbalance tackled by random under sampling. Smart oversampling techniques can help.
- Only binary clean vs cyberbullying classification is done. Fine-grained classification into subtypes needed.
- Lack of multimodal features like images, videos, and metadata.
- No analysis of model robustness against adversarial attacks.

The future work directions are:

- Evaluate models on multilingual datasets from different platforms like Twitter, Instagram, Reddit.
- Employ advanced class balancing methods like SMOTE synth.
- Explore graph neural networks to incorporate social graph information among users.
- Develop multimodal models using text, images, videos and metadata.
- Leverage semi-supervised and weakly supervised approaches to reduce labeling efforts.
- Employ adversarial training to improve model robustness against attacks.
- Analyze model fairness and biases, ensure transparency and interpretability.
- Deploy and evaluate system on real-world social media data. Measure operational metrics like moderation overhead, user satisfaction.

Investigate legal and ethical aspects around automatic moderation. Balance free speech concerns with safety.

6. Conclusions

This paper presented an empirical study of deep learning techniques for cyberbullying detection in social networks. The performance of CNN, LSTM, CNN-LSTM, and BERT models is analyzed on a dataset of Facebook comments. BERT achieves the highest accuracy of 87.3% followed by a hierarchical CNN-LSTM architecture with 86.5% accuracy. The multi-layer CNN captures informative n-gram features while LSTM models long-range sequential dependencies. BERT benefits from bidirectional context modeling using self-attention. The results demonstrate the effectiveness of deep learning for automated cyberbullying detection. However, enhancements in multilingual, multimodal support and adversarial robustness are required. Testing on diverse social media platforms and content types needs to be done. User privacy and freedom of speech considerations have to be addressed while developing real-world moderation systems. This study provides useful insights and forms a strong baseline for future research on this important problem.

Acknowledgment

"This research is funded by the Deanship of Research in Zarqa University/Jordan."

References

[1] Abuowaida S., Elsoud E., Al-momani A., Arabiat M., Owida H., Alshdaifat N., and Chan H., "Proposed Enhanced Feature Extraction for Multi-Food Detection Method," *Journal of Theoretical and Applied Information Technology*, vol. 101, no. 24, pp. 8140-8146, 2023.

[2] Agrawal S. and Awekar A., "Deep learning for detecting cyberbullying across multiple social media platforms," *in Proceedings of the European Conference on Information Retrieval,* Grenoble, pp. 141-153, 2018.

[3] Alazaidah R., Ahmad F., Mohsen M., and Junoh A., "Evaluating Conditional and Unconditional Correlations Capturing Strategies in Multi Label Classification," *Journal of Telecommunication, Electronic and Computer Engineering*, vol. 10, no. 2, pp. 47-51, 2018. https://jtec.utem.edu.my/jtec/article/view/4315/3162

[4] Alazaidah R., Ahmad F., Mohsin M., and AlZoubi W., "Multi-Label Ranking Method Based on Positive Class Correlations," *Jordanian Journal of Computers and Information Technology*, vol. 6, no. 4, pp. 377-391, 2020.

[5] Alazaidah R., Almaiah M., and Al-Luwaici M., "Associative Classification in Multi-Label Classification: An Investigative Study," *Jordanian Journal of Computers and Information Technology*, vol. 7, no. 2, pp. 166-179, 2021. DOI:10.5455/jjcit.71-1615297634

[6] Alazaidah R., Samara G., Almatarneh S., Hassan M., Aljaidi M., and Mansur H., "Multi-Label Classification Based on Associations," *Applied Sciences*, vol. 13, no. 8, pp. 5081, 2023. https://doi.org/10.3390/app13085081

[7] Al-Garadi A., Varathan K., and Ravana S., "Cybercrime Detection in Online Communications: The Experimental Case of Cyberbullying Detection in the Twitter Network," *Computers and Security*, vol. 63, pp. 433-443, 2017. https://doi.org/10.1016/j.chb.2016.05.051

[8] Alhusenat A., Owida H., Rababah H., Al-Nabulsi J., and Abuowaida S., "A Secured Multi-Stages Authentication Protocol for IoT Devices," *Mathematical Modelling of Engineering Problems*, vol. 10, no. 4, pp. 1352-1358, 2023. https://doi.org/10.18280/mmep.100429

[9] Al-luwaici M., Junoh A., AlZoubi W., Alazaidah R., and Al-luwaici W., "New Features Selection Method for Multi-Label Classification Based on the Positive Dependencies among Labels," *Solid State Technology*, vol. 63, no. 2, pp. 9896, 2020.

[10] Alshraiedeh F., Hanna S., and Alazaidah R., "An Approach to Extend WSDL-Based Data Types Specification to Enhance Web Services Understandability," *International Journal of Advanced Computer Science and Applications*, vol. 6, no. 3, pp. 88-98, 2015.

[11] Alzyoud M., Alazaidah R., Aljaidi M., Samara G., Qasem M., Khalid M., and Al-Shanableh N., "Diagnosing Diabetes Mellitus Using Machine Learning Techniques," *International Journal of Data and Network Science*, vol. 8, no. 1, pp. 179-188, 2024. DOI:10.5267/j.ijdns.2023.10.006

[12] Badjatiya P., Gupta S., Gupta M. and Varma V., "Deep Learning for Hate Speech Detection in Tweets," *in Proceedings of the 26th International Conference on World Wide Web Companion*, Perth, pp. 759-760, 2017. https://doi.org/10.1145/3041021.3054223

[13] Baroncelli A., Perkins ER., Ciucci E., Frick P., Patrick C., Sica C., "Triarchic Model Traits as Predictors of Bullying and Cyberbullying in Adolescence," *Journal of Interpersonal Violence*, vol. 37, no. 5, pp. 3242-3268, 2022. https://doi.org/10.1177/0886260520934448

[14] Cheng L., Guo R., and Li J., "Hierarchical Attention Networks for Cyberbullying Detection with Textual and Social Context," *in Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics*, Dublin, pp. 1234-1245, 2022. DOI:10.1137/1.9781611975673.27

[15] Dadvar M. and De Jong F., "Cyberbullying Detection: A Step Toward a Safer Internet Yard," *in Proceedings of the 21st International Conference on World Wide Web*, New York, pp. 121-122, 2012. https://doi.org/10.1145/2187980.2187995

[16] Devlin J., Chang M., Lee K., and Toutanova K., "BERT: Pre-Training of Deep Bidirectional Transformers for Language Understanding," *in Proceedings of the Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies,* Minneapolis, pp. 4171-4186, 2019. https://doi.org/10.18653/v1/n19-1423

[17] Elbasi E. and Zreikat A., "Heart Disease Classification for Early Diagnosis Based on Adaptive Hoeffding Tree Algorithm in IoMT Data," *The International Arab Journal of Information Technology*, vol. 20, no. 1, pp. 38-48, 2023. https://doi.org/10.34028/iajit/20/1/5

[18] Hinduja S. and Patchin J., "Bullying, Cyberbullying, and Suicide," *Archives of Suicide Research*, vol. 14, no. 3, pp. 206-221, 2010. DOI:10.1080/13811118.2010.494133

[19] Hochreiter S. and Schmidhuber J., "Long Short-Term Memory," *Neural Computation*, vol. 9, no. 8, pp. 1735-1780, 1997. DOI:10.1162/neco.1997.9.8.1735

[20] Kim Y., "Convolutional Neural Networks for Sentence Classification," *arXiv Preprint*, vol. arXiv:1408.5882, pp. 1-6, 2014. https://doi.org/10.48550/arXiv.1408.5882

[21] Li Y., Zhang Z., and Wang H., "Multi-task Learning for Cyberbullying Detection and Emotion Recognition in Social Media Texts," *IEEE Transactions on Affective Computing*, vol. 14, no. 2, pp. 456-468, 2023.

[22] Mazari A. and Kheddar H., "Deep Learning-Based Analysis of Algerian Dialect Dataset Targeted Hate Speech, Offensive Language and Cyberbullying," *International Journal of Computing and Digital Systems*, vol. 13, no. 1, pp. 965-972, 2023. DOI:10.12785/ijcds/130177

[23] Paul S., Saha S., and Singh J., "COVID-19 and Cyberbullying: Deep Ensemble Model to Identify Cyberbullying from Code-Switched Languages During the Pandemic," *Multimedia Tools and Applications*, vol. 82, no. 6, pp. 8773-8789, 2022. DOI:10.1007/s11042-021-11601-9

[24] Pericherla S. and Ilavarasan E., "Transformer Network-Based Word Embeddings Approach for Autonomous Cyberbullying Detection," *International Journal of Intelligent Unmanned Systems*, vol. 12, no. 1, pp. 154-166, 2021. https://doi.org/10.1108/IJIUS-02-2021-0011

[25] Ptáček V., Habernal I., and Hong J., Cyberbullying Classification Dataset. Kaggle. https://www.kaggle.com/datasets/vpacheco/cyberbullying-classification , Last Visited, 2024.

[26] Rosa H., Matos D., Ribeiro R., Coheur L., and Carvalho J., "A "Deeper" Look at Detecting Cyberbullying in Social Networks," *in Proceedings of the International Joint Conference*

on Neural Networks, Rio de Janeiro, pp. 1-8, 2018. DOI:10.1109/IJCNN.2018.8489211

[27] Salmivalli G., "Bullying and the Peer Group: A Review," *Aggression and Violent Behavior*, vol. 15, no. 2, pp. 112-120, 2010. https://doi.org/10.1016/j.avb.2009.08.007

[28] UNESCO, Behind the Numbers: Ending School Violence and bullying, 2021. https://en.unesco.org/news/behind-numbers-ending-school-violence-and-bullying, Last Visited, 2024.

[29] Yadav K., Lamba A., Gupta D., Gupta A., Karmakar P., and Saini S., "Bi-LSTM and Ensemble based Bilingual Sentiment Analysis for a Code-mixed Hindi-English Social Media Text," *in Proceedings of the IEEE 17th India Council International Conference*, New Delhi, pp. 1-6, 2020.

Hayel Khafajeh obtained his Ph.D. in Computer Information Systems in 2008 in Jordan. He joined Zarqa University, Jordan in 2009. In 2010, he served as Head of the CIS Department for two years. Since academic year 2014/2015, he has served and he still as the Vice Dean of the IT College at Zarqa. Hayel Khafajeh has worked for 23 years in the educational field as Programmer, Teachers Supervisor, head of IT Division, and manager of ICDL Center. He has published many educational computer books for the Ministry of Education in Jordan. His research interests include Information Retrieval, AI and E-learning. He is the author of several publications on these topics.