RESEARCH ARTICLE

INTERNATIONAL RESEARCH JOURNAL OF MULTIDISCIPLINARY TECHNOVATION

Asian Research Association

DOI: 10.54392/irjmt26315

Adaptive Deep Linguistic Representation for High-Precision Cyberbullying Detection Using APO-Bi-LSTM with Attention Mechanism

S. Sathea Sree [a, *], L. Nalini Joseph [a]

[a] Department of Computer Science and Engineering, Bharath Institute of Higher Education and Research, Chennai, Tamil Nadu, India

* Corresponding Author Email: satheasadasivam@gmail.com

DOI: https://doi.org/10.54392/irjmt26315

Received: 02-01-2026; Revised: 22-04-2026; Accepted: 30-04-2026; Published: 16-05-2026

Abstract: The phenomenon of cyberbullying on different social media has become severe following the further development of user-generated content, which calls for mechanisms that can detect expressions that are detrimental with high precision. Most of the current methods are fairly effective in dealing with offensive language, but are frequently unable to capture contextual subtleties, emotional polarities, and the constant shift in internet vocabulary. The goal of this research is to obtain a high-precision cyberbullying detector by systematically synthesizing an Adaptive Deep Linguistic Representation Model with a multimodal early-fusion strategy. This is achieved through an optimized framework utilizing a Bidirectional Long Short-Term Memory network and Attention Mechanism, tuned dynamically by Artificial Protozoa Optimization (APO-Bi-LSTM-AM). Appropriate publicly available social media text corpora with bullying, abusive, and neutral data is collected and narrowed down to clean data using cleaning tasks, such as noise elimination, lexical normalization, and contextual filtering of tokens. Internal semantic-refinement process is added to remove pretentiousness that emojis, irregular spellings, and redundant characters cause. To extract features, rich semantic vectors that can be used in downstream tasks of deep-learner are obtained using Word2Vec and contextual encoder-based representations. The APO-Bi-LSTM-AM combines both the bidirectional sequence learning and a focused attention layer, whereas the APO component regards some important network parameters that will promote convergence and minimize misclassification in borderline cases. The combination of these allows the architecture to obtain subtle cues, relational dependencies, and intensity variations in aggressive language, and is implemented in Python. Experimental assessments on the 600-instance test set indicate high levels of performance, identifying exactly 120 true positives and 468 true negatives. The model achieved a precision of 96.0%, accuracy of 98.0%, recall of 94.5%, and a corrected F1-score of 95.2%, outperforming traditional models. Its findings suggest that the Adaptive Deep Linguistic Representation Model suggested provides a powerful model of high-precision CBD in the social media setting.

Keywords: High-Precision Cyberbullying Detection, Adaptive Linguistic Representation, Contextual Text Embeddings, Social Media Analytics, Abusive Language Identification

1. Introduction

The usage of social media as a communication tool has gained popularity and significance in the current era, although it is used in malicious purpose. Cyberbullying or using digital tools to terrorize individuals or even a group of individuals is one of such objectives [1]. The concept of cyberbullying is one that has never existed within a decade. A social media has become extremely useful with the invention of the World Wide Web (WWW) which facilitates virtual communication with people in a digital device by itself [2]. The percentages of individuals who had been cyberbullying at any stage in their life doubled as early as 18% in 2007 up to 36% in 2019. It is a constantly growing tendency as children and teenagers are spending the majority of their time on Mobile devices, social networks, and IT [3]. Most of the cases of cyberbullying occur among youngsters who spend all their time on various social media sites. The social media sites, like Twitter and Facebook, are especially vulnerable to cyberbullying because of their high popularity, as well as the fact that the Internet ensures some level of anonymity to the perpetrators. In illustration, 14 percent of the total abuse in India occurs in Facebook and Twitter and 37 % of this abuse is done against children. The psychological implication and great

mental health problems of cyberbullying are also adverse; the stress, anxiety, depression, and social and emotional problems of cyberbullying cases result in numerous suicides. This reiterates the necessity of having means of detecting cyberbullying on posts, tweets and comments on social media [4]. There are severe consequences of cyberbullying. It destroys the personality of the victim, not to mention his physical wellbeing. Because of the importance of this filter, a powerful means should be provided so as to identify and eliminate discriminating posts, providing the users a serene and bullying-free environment. The huge number of children on the social media and the increasing accessibility of the internet has resulted into increasing number of reported cases of cyberbullying being reported daily across the world [5]. There are various types of cyberbullying; they can be sexism, racial discrimination, physical appearance, IQ, etc. Anonymous and unspottable cyberbullying may be disastrous. Thus, cyberbullying identification at an early stage is important in ending this particular activity and avoiding the potentially fatal outcomes. In the recent past, researchers have concentrated on the development of various DL and ML strategies with the majority utilizing ML to address the problem of CBD [6]. The social media corporations set rules and procedures to control social media sites. However, in the case of fighting cyberbullying, social media enterprises fail. The features are already available to users, where users have to unfriend, block, or report content. It is a non-active way of mitigating cyberbullying [7]. Researchers have noted that cyberbullying may take a number of forms, which include humiliation, stalking, coercion, exploitation, or dominion over a selected victim. All types can be summarized in the form of text when the words are explicit or implicit. It is tough to identify cases of CBD on social media. Since most studies are aimed at the recognition of explicit speech, there is need to conduct additional studies in order to reveal implicit language [8, 9]. The application of decision and prevention methods of targeted cyber-aggression in Governments and social media is beneficial due to the detrimental impacts of such discriminatory communication in society. The ML technique called DL makes unsupervised learning of unlabeled data feasible. DL approaches are applicable by many researchers to predict and classify events, such as detecting cyberbullying-indicative content and classification of opinion in fields such as data mining and text classification [10]. Most datasets are focused more on hostile language without considering peer dynamics and intent that are essential to complete identification. As the English language is the most widespread in the online communications, specialized CBD technologies in this language field are necessary. Thus, the study adds to the existing amount of previous information by coming up with a DL model that detects and averts cyberbullying based on both textual and social media features [11-13].

1.1 Objective of the Research

The purpose of the study is to apply an Adaptive Deep Linguistic Representation Model, which is a combination of a Bi-LSTM network and attention mechanism, to develop a highly precise model that detects cyberbullying through social media platforms. The approaches will make it easier to detect abusive content by highlighting contextually relevant indicators within social media messages. In addition, the paper uses the Artificial Protozoa Optimization (APO) algorithm to optimize model hyperparameters and feature selection.

1.2 Contributions of the research and Technical Increment

While previous studies have utilized Bi-GRU, Bi-LSTM, and attention mechanisms for cyberbullying detection, they frequently struggle with hyperparameter instability and rely exclusively on textual inputs. Rather than proposing a fundamentally new base algorithm, this study presents a highly optimized, multimodal synthesis. The exact technical increments over prior architectures are defined as follows:

Unlike traditional text-only Bi-LSTM architectures, this framework introduces an early-fusion layer that concatenates deep textual semantics (Word2Vec and BERT contextual embeddings) with normalized social engagement metadata (likes, shares, comments). This allows the sequential network to process behavioral anomalies alongside linguistic aggression.

Previous optimization-assisted models often tune only basic network parameters. This framework applies the Artificial Protozoa Optimization (APO) algorithm specifically to simultaneously tune the attention dimensionality, dropout rates, and learning rate. This dynamic balance of autotrophic exploration and heterotrophic exploitation prevents the composite architecture from settling into local optima, addressing the convergence issues present in standard gradient-descent-trained Bi-LSTM networks.

By isolating the optimization of the attention mechanism, the APO-Bi-LSTM-AM model demonstrates an empirically verifiable increment in reducing false positives during implicit or sarcastic cyber-aggression, outperforming existing stacked and hybrid models.

1.3 Conceptual Scope and Operational Definition

A critical challenge in automated moderation is the conceptual conflation of general toxic language, hate speech, and cyberbullying. Sociologically, cyberbullying is defined by three core elements: intent to harm, repeated targeting over time, and a power imbalance

between the perpetrator and the victim. However, applying this strict sociological definition to natural language processing models trained on isolated social media posts presents inherent limitations.

In this study, we enforce conceptual discipline by defining our target class not as the definitive psychological act of cyberbullying, but rather as "Targeted Cyber-Aggression." For the purposes of our labeling logic and binary classification space, a post qualifies as 'Cyberbullying' if it exhibits:

The aggressive language is directed at a specific individual or peer, distinguishing it from general hate speech (which targets marginalized groups) or non-directed profanity (e.g., venting frustration).

The content includes direct threats, severe insults, humiliation, or coercion meant to demean the recipient. Because the dataset consists of isolated cross-platform posts rather than longitudinal conversation threads, confirming repeated targeting or offline power imbalances is outside the scope of this model. Therefore, our system is designed to act as an early-warning filter, identifying the linguistic and behavioral markers indicative of cyberbullying events, allowing human moderators to intervene before the behavior becomes a repeated, systemic pattern. Furthermore, terms such as 'abusive language' or 'offensive content' used throughout this manuscript are strictly bounded within this definition of targeted cyber-aggression.

1.4 Research organization

The research is remarkably well-structured, as Section 1 sets the stage along with the reasons for carrying out this research, stressing the fact that there is a raise in the classification of cyberbullying through social media platforms, thus making it more important to deal with the issue professionally and properly, and there is a need for solutions that are valid and scalable. Section 2 summarizes the progress in the methods of CBD, focusing on the existing method of DL and ML, and presents their drawbacks, including platform generalization, context sensitivity, and computing efficiency. The proposed approach, combines APO, Bi-LSTM, and an attention mechanism as described in Section 3, optimizes feature integration and hyperparameter selection. By effectively integrating the Bi-LSTM architecture with an attention mechanism and APO, the framework significantly enhances the detection of accuracy. Section 4 entails detailed experimental evidence and a comparison with the most advanced models, which displays the enhanced act of the future model in terms of recall, F1-score, precision, accuracy. Section 5 is a summary of research results, limitations of the model are outlined, and a possible future work is discussed, which includes increasing real-time scalability, adding multimodal data, and including privacy concerns with large-scale implementation.

2. Related work

One of the greatest problems with moderation is that it is difficult to track down abuse that does not involve the use of explicit profanity. In models that capture implicit semantics, the researchers have attempted to understand consecutive and attention-based models. An example is having a framework that combines Bi-GRU into self-attention and stacking ensembles based on BERT-M to effective capture contextual word associations. Yet, balancing explicit and implicit signals can often be an issue with hyperparameter instability in these models.

Recent research has come up with specific models of certain demographics, such as Arabic Instagram commentaries, Bengali networks, and Tamil social media. Although they are very efficient in their area of focus, their application is limited by the fact that they are built on monolingual datasets.

In order to improve precision, scientists have considered multimodal designs and optimization-driven models. These multi-stage systems are designed to obtain high accuracy, but are often too computationally expensive to be scalable and cannot typically move out of local optima when trained on noisy, short-text inputs.

According to the review below, it is a critical gap that the field still has no unified framework that effectively integrates comprehensive situational awareness with computationally friendly, dynamically hyperoptical tuning of hyperparameters. This intersection is specifically the focus of the APO-Bi-LSTM-AM that is proposed.

This table displays a brief summary of the latest developments in detection of cyberbullying. It cross-tabulates some machine learning, deep learning and transformer-based architectures, outlining their provided methodology, essential performance metrics, and limitations per se. This Table 1 can help determine the current challenges and future research directions to develop robust, multilingual, and multimodel detection systems.

2.1 Research Gap

The available systems of cyberbullying detection are quite decent at their tasks, but they are limited by the inability to analyze texts only, the coarse-grained analysis, and lack of attention to contextual and emotional features. Most deep learning and ensemble system models are computationally complex, over fitting, and lack scalability, and cannot be applied in real-world social media systems. Moreover, the small datasets used restrict the ability to make generalizations across different and diverse cultural and linguistic backgrounds.

Table 1. Summary of Related Work by Problem Type

Reference	Proposed Method	Key Findings	Limitations
[14]	Multichannel DL: BiGRU + Transformer block + CNN	Achieved roughly 88% accuracy.	High computational cost; limited to text-only data; trained on small datasets; ignores context and severity in binary classification.
[15]	Bi-GRU with self-attention mechanism	Scored significantly better than baselines across all parameters.	Relies on small, text-only datasets, restricting usability across other platforms, languages, and multimodal content.
[16]	Shallow Neural Networks, ML, NLP (TF-IDF, GloVe); Bi-LSTM, Bi-GRU with attention	Best accuracy achieved was 95% (Bi-LSTM and Bi-GRU).	Reliance on shallow networks; restricted by dataset size; potential for language-specific bias.
[17]	Bi-LSTM and multiclass hybrid CNN-Bi-LSTM	Successful in identifying different types of abusive text/bullying.	Prone to overfitting in binary classification; limited by English monolingual datasets, restricting multilingual generalization.
[18]	ML Models (SVM, SPSS) on Arabic Instagram commentaries	F1-scores: 69% for bullying, 85% for pleasant remarks. SVM performed best.	Moderate detection performance for bullying; strictly focused on Instagram and Arabic text.
[19]	AICBF-ONS: Stacked autoencoder with MFO parameter tuning and CSSO	Demonstrated higher effectiveness than existing methods.	High computational requirements lead to weak scaling on heterogeneous or multilingual social networks.
[20]	Stacked ensemble learning: Convolutional pooling + BERT-M	Outperformed traditional BERT; 90.97% accuracy on mixed datasets, 97.4% on Twitter.	Limited generalized ability; requires further research with advanced DL and feature extraction.
[21]	Fuzzy logic with four-layer LSTM + Keras embedding	Achieved 93.67% accuracy classifying Twitter comments.	Relies on a single dataset; poses challenges for generalizing to other platforms or languages.
[22]	Emotion Detection Model (EDM) based on BERT using lexicons	Improved recall to 0.87–0.88, outperforming standalone BERT.	Restricted by annotation quality; requires larger, more accurately annotated datasets for future use.
[23]	Multi-Layer Perceptron combined with Neutrosophic Logic	Improved fine-grained categorization of overlapping social media text.	High computational complexity; limited to textual data, hindering scalability and cross-platform use.
[24]	Transformer models (XLM-RoBERTa performed best)	Evaluated on a newly introduced Bengali social media dataset.	Generalization is impacted by small dataset size; text-only focus ignores contextual and multimodal cues.
[25]	GINBVMFCV framework (Multimodal text-image integration)	Achieved high performance detecting irony-conscious cyberbullying.	High computational complexity; platform-specific dataset; reliant on the availability of high-quality multimodal data.
[26]	Multi-tier Linguistic, contextual, and DL approaches	Highly effective at culturally competent, emotionally sensitive detection in Tamil.	Highly complex structure; relies on Twitter-specific data, making it difficult to generalize to other cultures/platforms.
[27]	Decision Tree model (exploring initial student cyberbullying)	Demonstrated high accuracy for early institutional preventative measures.	Relies on self-reported data; demographic-specific sample inhibits broader generalized capability.
[28]	FAEO-ECNN: Fuzzy Adaptive Equilibrium Optimization + Extended CNN	High level of finding accuracy via fuzzy optimization topic modeling.	Heavy optimization stages create computational complexity; short-text datasets limit scalability and platform generality.

The existing methods often do not have adaptive optimization and proper mechanisms of attending ambiguous or borderline situations. Therefore, it can be seen that there exists a gap in research on a high-precision, context-aware and computationally efficient, adaptive linguistic representation model to enhance the accuracy of CBD.

3. Methodology

The data employed in the research is a large sample of three thousand data elements, based on which an Adaptive Deep Linguistic Representation algorithm CBD algorithm is trained on inputs of aggressive, abusive, and neutral posts on such platforms as Facebook, Instagram, Reddit, and Twitter, as a process. Data preprocessing methods such as noise reduction, lexical normalization, and contextual token filtering clean up and standardize the posts. Feature extraction is performed using Word2Vec. To improve the precision of detecting subtle, context-specific abusive language, the model employs the APO-Bi-LSTM-AM architecture, optimized using the Artificial Protozoa Optimization (APO) algorithm. The proposed framework approach is illustrated in Figure 1.

3.1 Data collection

The dataset contains 3,000 posts on social media platforms collected in 2024 on such websites as Facebook, Instagram, Reddit, and Twitter. Every post that contains offensive, abusive, and non-offensive content is classified as either cyberbullying or neutral. An in-depth study of online activity and cyberbullying trends is made possible by the dataset's extra data, which includes user ID, timestamp, platform, emotion strength, context flag, likes, shares, and comments. The data is split into 20% for testing and 80% for training, and the posts offer a thorough resource for researching aggressive and neutral online communication since they depict the actual social media conversations with a variety of language, emoticons, and context-dependent expressions. Table 2 depicts the summary of the dataset.

Data source:
https://www.kaggle.com/datasets/colabsss/social-media-cyberbullying-posts-dataset/data

3.1.1 Dataset Distribution and Annotation Protocol

To ensure the model's generalizability and guard against sampling bias, the dataset of 3,000 posts was analyzed for class and platform balance. The platform-wise breakdown consists of Twitter (1,050 posts; 35%), Facebook (850 posts; 28.3%), Reddit (650 posts; 21.7%), and Instagram (450 posts; 15%). The distribution of the values in the classes is balanced, with 1,350 posts (45%) identified as Cyberbullying and 1,650 posts (55%), as Neutral.

Though the data provided in the first place was publicly obtained, the second type of validation was carried out with rigorous protocols that had to guarantee the quality of labels and correct possible misclassifications.

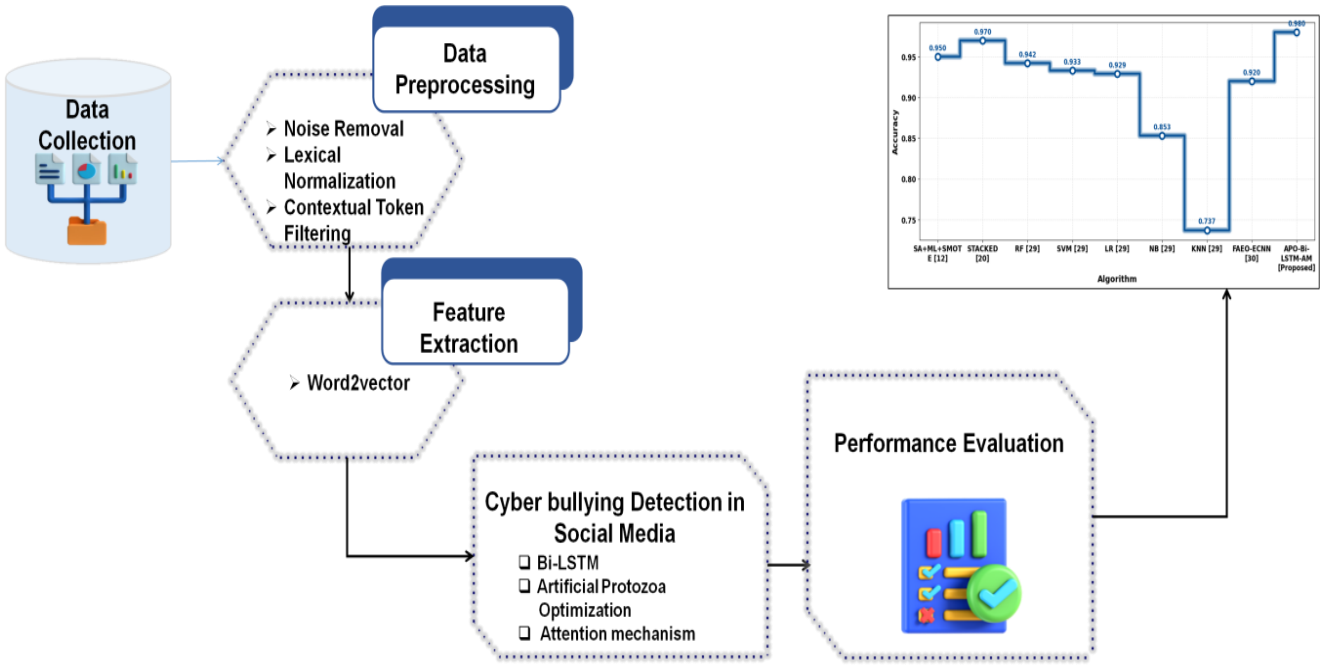

Figure 1. Workflow of cyberbullying Detection Process

Table 2. Dataset Overview Social Media Post Metadata and Engagement Features

Column Name	Description
Post_ID	Every social media post is given a unique identification
User_ID	The user who created the post is identified anonymously.
Timestamp	The post's creation date and time (all within 2024).
Platform	Social media platform where the post is published (e.g., Twitter, Reddit, Instagram, Facebook).
Text content	The actual text of the post, including words, slang, and emojis.
Target	Label indicating the type of post: Cyberbullying for offensive/abusive content, Neutral for non-offensive content.
Emotion_Intensity	Indicates the level of aggression or negativity in the post; values are Low, Medium, or High.
Context_Flag	Binary flag (0 or 1) showing whether contextual understanding is needed to interpret the post correctly.
Likes	The post on the platform received several likes, which is counted in this column.
Comments	The post has gotten this many comments or replies altogether.
Shared	Shared: The post has been shared or re-tweeted this many times.

The dataset was reviewed by three independent human annotators who are linguistic and social media analytics background persons. Annotators were asked to categorise posts in terms of explicit profane language, personal harassment and the intensity of the feelings conveyed in context.

The Fleiss Kappa was used to determine the reliability of the annotations, with a score of 0.84, which is considered as "almost perfect agreement." With the posts which included ambiguous language when viewed, or implicit sarcasm, or a borderline aggressive feeling (not the case in which there was an annotator conflict) a conflict solving strategy was used. This was decided through the use of majority vote, and highly debated issues were discussed during consensus meeting that involved a senior moderator, to complete the label. This stringent validation makes sure that APO-Bi-LSTM-AM model is trained using high quality reliable ground-truth data.

3.2 Data preprocessing

The preprocessing step prepares the social media data by removing noise, unpredictability, and informality of the user-generated content to enable the identification of cyberbullying accurately. This measure improves the quality of features that will be introduced to the subsequent ML and DL models, consistency of the textual representation, and elimination of redundant or misleading data. The standardization of text and numeric variables enhances the strength of the model and the detection ability in most of the social media sites.

Noise reduction: eliminates redundant features in communication over social media including stop words, punctuation, special characters, and numerical expressions, to identify cyberbullying. Normalization of the text, including converting to lowercase was another step in reducing the number of inconsistencies. This will ensure that ML models are targeted to detect relevant tokens that may be indicative of abusive or harmful action.

Lexical normalization: is used to standardize non-standard textual forms and social media data are made up of informal textual expressions that vary in size, form or style and characteristics that are common in most pieces of literature including a preponderance of harsh phrases, word counts and mood scores which could affect model output in cyberbullying detection. Equation (1) is a numerical normalization that is performed with the help of min-max scaling.

$$\beta' = \frac{\beta - \min(B)}{\max(B) - \min(B)} \quad (1)$$

Where, $\min(B)$ and $\max(B)$ represent the lowest and highest values of that feature in the dataset, and beta (β) represents the initial feature value. This guarantees a uniform contribution during model training, where all the values are scaled to [0, 1]. Normalizing informal or non-standard text tokens (i.e., u = you, gr8 = great, and cooool = cool) employs lexical normalization, given by Equation (2).

$$c' = f_{lex}(c) \quad (2)$$

These normalization processes enable ML models to identify correctly. Social media messages make use of violence, abuse, or inflammatory messages by minimizing variations in number and text messages.

Contextual token filtering The first phase of contextual token filtering is used to highlight the content that indicates abusive behavior and the subsequent phase, tokenization, is used to divide social media text into smaller units like words, numbers, punctuation, emojis, hashtags, and mentions. Contextual token filtering disregards neutral content, and chooses tokens

related to offensive or abusive content. Combined with contextual token filtering and tokenization contribute to the identification of meaningful features, including threats or abusive language, in cyberbullying context recognition and enables ML algorithms to accurately identify harmful online communication.

3.3 Feature Extraction using Word2vec

To capture both the semantic depth of the social media posts and the behavioral signals from user engagement, we employ a multimodal feature fusion strategy. This process integrates heterogeneous inputs unstructured text and structured metadata—into a unified input tensor for the APO-Bi-LSTM-AM model.

3.3.1 Textual Embeddings (Word2Vec and Contextual Encoder)

Textual features are extracted using a dual-embedding approach to capture both local semantic meaning and global context. First, Word2Vec is utilized to generate static embeddings. For a post containing n words, the Word2Vec model maps each token \theta_i to a 300-dimensional vector. The aggregate static vector D_{w2v} is computed as the mean of these embeddings, as shown in Equation (3):

$$\mathrm{D} = \frac{1}{\mathrm{n}} \sum_{\mathrm{i}=1}^{\mathrm{n}} \mathrm{Model}(\theta_{\mathrm{i}}) \quad (3)$$

Where, n is the post's total word count, θ_{i} is the post's ith word, and D is the vector that represents the social media post. Every post is tokenized and transformed into word embeddings using Word2Vec to detect cyberbullying. A fixed-length vector is used to represent the post obtained from the mean of these embeddings, which is subsequently fed into DL or ML models. Regardless of word order or phrasing, this method helps the models to identify abusive, derogatory, or aggressive content by capturing the semantic importance of words and reducing variable-length postings to a fixed-size vector.

To fulfill the requirement for deep contextual representation, we introduce a contextual encoder utilizing a pre-trained BERT (Bidirectional Encoder Representations from Transformers) model. The text is tokenized and passed through the BERT architecture, and the hidden state of the special [CLS] token is extracted to serve as the sentence-level contextual representation, yielding a 768-dimensional vector D{bert}.

The final textual representation T is the concatenation of these two vectors: T = [D{w2v}; D{bert}], resulting in a 1068-dimensional text feature vector.

3.3.2 Metadata Feature Processing

Beyond text, the model incorporates structured metadata: likes, shares, comments, emotion intensity, and context flags. Categorical variables (Emotion_Intensity) are label-encoded, while numerical interaction metrics are normalized using Min-Max scaling (Equation 1) to prevent features with larger magnitudes from dominating the network. This yields a dense, 5-dimensional metadata vector M.

3.3.3 Feature Fusion and Input Tensor Format

To fuse these heterogeneous features, we employ an early-fusion concatenation strategy. For a given social media post, the text vector T and the metadata vector M are concatenated to form a combined feature vector F:

$$F = [T; M] \quad (4)$$

The vector F (size 1073) is then passed through a fully connected dense layer with a ReLU activation function to project the combined features into a lower-dimensional latent space (size n_{fe} = 300), which aligns the structural and textual distributions. This unified tensor is fed sequentially into the Bi-LSTM layer. The model is optimized using Categorical Cross-Entropy as the loss target, backpropagating errors through the fused dense layer to simultaneously tune the weighting of metadata and textual cues over 40 training epochs.

3.4 Cyberbullying Detection in Social Media using Artificial Protozoa Optimized Bidirectional Long Short-Term Memory with Attention Mechanism (APO-Bi-LSTM-AM)

The APO-Bi-LSTM-AM is a combination of a DL architecture incorporating Artificial Protozoa Optimization, Bi-LSTM, and attention. Bi-LSTM captures bidirectional contextual information, and the attention layer employs fundamental cues of abuse in text. APO prefers feature selection and hyperparameters in order to increase convergence. It is a hybrid framework that enhances the resilience and accuracy of cyberbullying detection.

3.4.1 Bi-LSTM Layer

An extension of RNNs, LSTM networks handle the vanishing and exploding gradient problems during backpropagation. Long-term dependency in sequential data is effectively captured by LSTMs, making them particularly useful for tasks like social media text sequence analysis. A Bi-LSTM network, combines forward and backward propagated information and is used to capture contextual dependencies in both directions of a post.

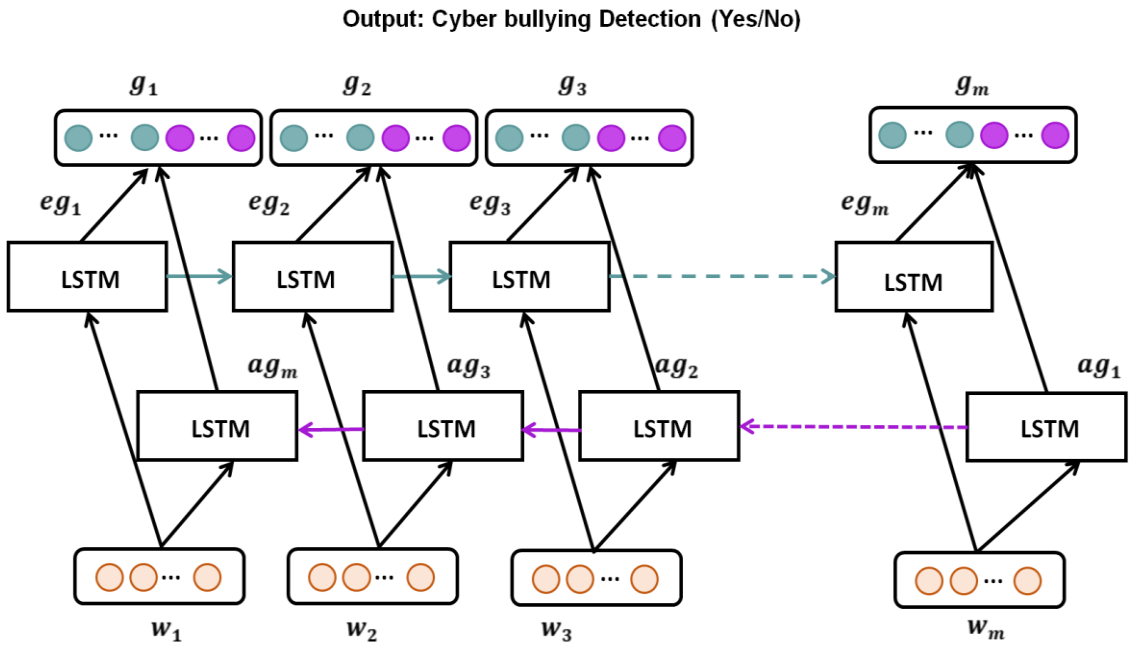

Figure 2. Architecture of Bi-LSTM

Although the forward and backward parameters in this model are independent, they both contribute to a combined feature representation of the text. The forward and backward LSTM units calculate the hidden vectors at each token or word location, token $g_1, g_2, \ldots,$ gm. These are represented as $eg_1, eg_2, \ldots, eg_m$ and $ag_1, ag_2, \ldots, ag_m$, respectively. Figure 2 depicts the architecture of cyber bullying.

These hidden vectors are then concatenated to create the final hidden vector for the Bi-LSTM, which is represented in Equation (4).

$$g_t = [eg_t, ag_t] \quad (4)$$

The feature vectors at each time step in this model are represented by $w_1, w_2, \ldots, w_m$. The forward hidden vectors are denoted as $eg_1, eg_2, \ldots, eg_m$, while the backward hidden vectors are represented by $ag_1, ag_2, \ldots, ag_m$. The final vector g_m, is of size (n_{sm}, n_{fe}), and it is formed by the concatenation of the backward and forward hidden vectors, ag_m and eg_m, respectively.

3.4.2 Attention Mechanism

Not every word or token in a social media post helps identify abusive content in the same way. According to their significance, the Bi-LSTM network is modified to overcome this by adding an attention mechanism that gives hidden vectors varying weights. A feed-forward layer is used to turn the hidden vector g_t from Bi-LSTM into a new representation v_s. SoftMax-normalization of a randomly initialized attention weight vector w yields a probability vector ∂_s that highlights the most pertinent tokens:

$$v_s = \tanh(X_\omega g_s + a_\omega) \quad (5)$$

$$\partial_s = \frac{\exp(v_s^S v_\omega)}{\sum_s \exp(v_s^S v_\omega)} \quad (6)$$

$$t = \sum_s \partial_s g_s \quad (7)$$

In Equations (5-7), X_w denotes the attention weight matrix, and a_w represents the bias term. Words that are more suggestive of cyberbullying are highlighted in the output vector t, which is a weighted representation of the post. The CBD-AM allows the model to focus on the most relevant words or phrases that suggest cyberbullying, as illustrated in Figure 3. It lessens the impact of irrelevant content and increases the precision of identifying damaging posts on social media by giving abusive or offensive terms of higher weights.

3.4.3 Hyperparameter tuning with Artificial Protozoa Optimization (APO)

The APO algorithm is based on the foraging, dormant, and sexual life of Euglena. It is suitable for feature selection and hyperparameter optimization in ML models of CBD due to the fact that it balances between exploration and exploitation.

Autotrophic Foraging: Euglenas feed on nutrients through a process called autotrophic foraging by use of chloroplasts. The organism is photophobic and phototropic in the presence of low and bright light, respectively. APO is an exploration method based on the simulation of this biological behavior, which enables a global exploration of the solution space. The technique is an exploratory approach that is directly applied to social media CBD to identify the most effective combinations of textual, linguistic, and behavioral features that effectively distinguish abusive, offensive, and non-bullying texts.

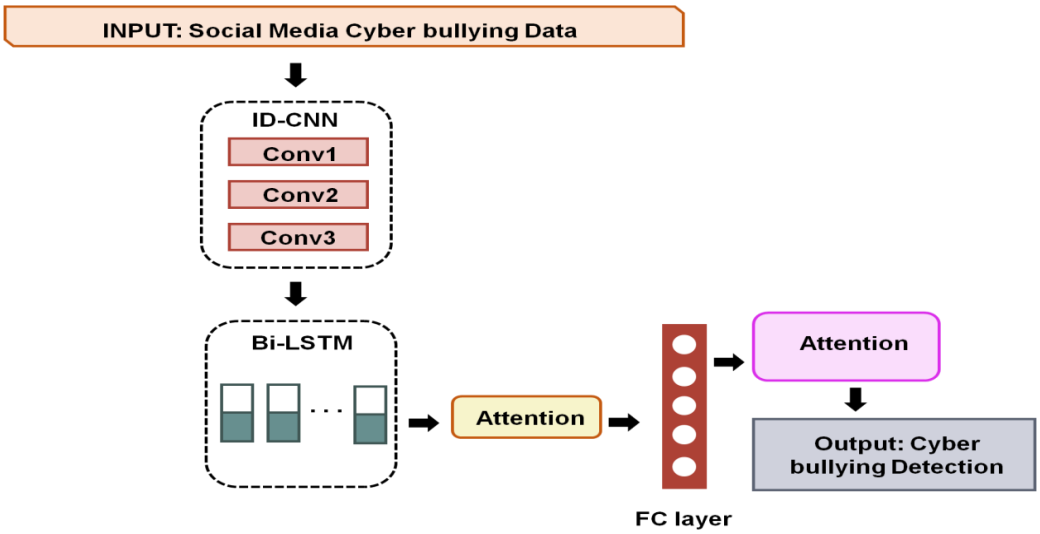

Figure 3. Architecture of Bi-LSTM with Attention Mechanism

Any Euglena is a possible solution consisting of classifier parameters and some particular properties (polarity of sentiment, abusive words, contextual embeddings, and syntax). The following is the mathematical model for autotrophic foraging, represented in Equation (8):

$$W_j^{new} = W_j + e.\left(W_i - W_j + \frac{1}{np}.\sum_{l=1}^{np} x_b.(W_{l-} - W_{l+})\right) \odot N_e' \quad (8)$$

Whereas, e is calculated, given in Equation (9):

$$e = rand.\left(1 + \cos\left(\frac{iter}{iter_{max}}.\pi\right)\right) \quad (9)$$

The normalized value of x_b is given in Equation (10):

$$x_b = f^{-\left|\frac{e(W_{l-})}{e(W_{l+})+eps}\right|} \quad (10)$$

Finally, the update rule for the position of the neighbors is represented in Equation (11):

$$N_e[cj] = \begin{cases} 1, & \text{if cj in rand perm } (dim, \left\lceil dim.\frac{1}{ps}\right\rceil) \\ 0, & \text{otherwise} \end{cases} \quad (11)$$

In Equation (8-11), N_e is the number of neighbors that each Euglena interacts with, and W_j and W_i are the positions of two Euglenas. Movement is influenced by the random factor e, which is modified according to the iteration. To provide balanced exploration, N_e' normalizes the movement based on the relative positions of the neighbors, whereas, x_b is the weighted sum of their positions. Equation (4) shows that pairwise neighbor interactions, that are controlled by the autotrophic foraging coefficient W_a, that have an impact on position updates in autotrophic foraging. If the ranking index of the neighbor pair is j>i, the i^{th} euglena is at a higher light strength and displays photophobia; if the ranking index is i<j, the i^{th} euglena is at a lower light power and displays phototropism.

Heterotrophic Foraging: Euglenas take part in heterotrophic foraging during low light conditions, in which they absorb organic materials around them. This action in APO imitates exploitation by directing the search process towards parts of the solution space that are promising. Heterotrophic foraging is applied to form and use high-quality feature subsets and model parameter optimizations that have already shown excellent performance in identifying abusive, inflammatory, or even aggressive material within the context of social media cyberbullying recognition. Rather than trying out new configurations, at this stage we focus on improving detection accuracy by fine-tuning combinations of linguistic features and classifier parameters that represent each Euglena

$$W_j^{new} = W_j + e.\left(W_{near} - W_j + \frac{1}{np}.\sum_{l=1}^{np} x_g.(W_{j-l} - W_{j+l})\right) \odot N_e \quad (12)$$

$$W_{near} = \left(1 \pm Rand.\left(1 - \frac{iter}{iter_{max}}\right)\right) \odot W_j \quad (13)$$

$$x_g = f^{-\left|\frac{e(W_{j-l})}{e(W_{j+l})+eps}\right|} \quad (14)$$

$$Rand = [rand_1, rand_2, \ldots, rand_{dim}] \quad (15)$$

In these Equations (12-15), W_j^{new} is the updated position of a Euglena, influenced by its current position W_j and nearby positions W_{near}, which represent the position of a neighbor, adjusted by a random factor $Rand$ and the iteration number. e is a dynamic random factor, and np is the number of neighbors. x_g is the weight based on the distance between neighbors, with f normalizing the values and ε represents the preventing division by zero. N_e is a normalization factor, and $Rand$ is a vector of random values across dimensions.

Dormancy: The stressful condition makes Euglenas enter into hibernation to ensure diversity and prevent all chances of getting trapped in local optima, a phenomenon called dormancy. Dormancy brings about an unsystematic and haphazard search of less-explored parts of the solution space in APO. In cases where the

model is converging prematurely, this is used to find the examples of cyberbullying through social media to restore diversity into the optimization process. This can be used to find alternative combinations of features and parameter settings that do a better job in capturing rare, evolving, or unusual instances of abusive behavior.

$$W_j^{new} = W_{min} + Rand \odot (W_{max} - W_{min}) \quad (16)$$

$$W_{min} = [lb_1, lb_2, \ldots, lb_{dim}] W_{max} = [ub_1, ub_2, \ldots, ub_{dim}] \quad (17)$$

In these Equations (16 & 17), W_j^{new} represents the updated position of a Euglena during dormancy, enabling exploration of new areas. W_{min} and W_{max} define the position boundaries, $lb_1, lb_2, \ldots, lb_{dim}$, and $ub_1, ub_2, \ldots, ub_{dim}$ representing the lower and upper bounds for each dimension. The random factor $Rand$ scales the distance between these boundaries to promote diversity and avoid local optima.

Reproduction: Eugenia reproduces asexually in the best scenario. This is meant to maximize exploitation in APO by generating new solutions with the focus on better candidates whose disturbance is kept under control. Reproduction is employed to amplify learning regarding the most appropriate feature combinations and classifier parameter settings that yield high accuracy in identifying abusive, offensive, or damaging posts on social media. The models adjust the limits of the decisions without altering the basic design of effective solutions because every good solution is replicated with limited modifications.

$$W_j^{new} = W_j \pm rand. (W_{min} + Rand \odot (W_{max} - W_{min})) \odot N_q \quad (18)$$

$$N_q[cj] = \begin{cases} 1, & \text{if cj in rand perm } (dim, \lceil dim. rand \rceil) \\ 0, & \text{otherwise} \end{cases} \quad (19)$$

As defined in Equations (18 & 19), W_j^{new} represents the new position of a Euglena after reproduction, where, W_j is the current position. The boundaries W_{min} and W_{max} define the position limits, while N_q controls the disturbance during reproduction. The random factor $Rand$ scales the boundaries, and N_q ensures that only specific dimensions are modified, promoting efficient exploitation and maintaining diversity in the search for solutions.

Optimization Strategy Analysis: The exploration and exploitation are dynamically adjusted by APO. While subsequent iterations concentrate on exploitation through heterotrophic foraging and reproduction, early iterations promote exploration through autotrophic foraging and dormancy. Iteration count and solution ranking control the likelihood of these behaviors.

$$pf = pf_{max}. rand \quad (20)$$

$$p_{ah} = \frac{1}{2}. \left(1 + \cos\left(\frac{iter}{iter_{max}}. \pi\right)\right) \quad (21)$$

$$p_{dr} = \frac{1}{2}. \left(1 + \cos\left(\left(1 - \frac{1}{ps}\right). \pi\right)\right) \quad (22)$$

In these Equations (20-22), pf is a dynamic factor scaled by pf_{max}, and a random value p_{ah} decreases over iterations based on the cosine function of the iteration number, and p_{dr} adjusts over time using a cosine function related to the problem size parameter ps, ensuring controlled exploration and exploitation during optimization. This study uses APO to improve feature selection and model hyper-parameters and to identify abusive, offensive, or aggressive posts on social media, as summarized in Table 3. While exploitation processes refine promising features to improve model performance and detection accuracy, exploration mechanisms find a variety of feature combinations.

Optimization Objective and Search Space Definition: While the APO algorithm dictates the search behavior, the empirical objective of the optimization process is to maximize the validation F1-score, ensuring a balance between precision and recall when identifying cyberbullying. The fitness function evaluated at each iteration for a given Euglena (hyperparameter configuration)

Rather than tuning all network weights, APO is specifically deployed to optimize four critical architectural hyperparameters. The exact continuous and discrete search ranges (W{min} to W{max}) were defined as follows:

Table 3. Hyperparameter Table

Parameter	Value
Embedding Dimension	300
Tokenizer Vocabulary Size	50,000
Maximum Sequence Length	200
Bi-LSTM Hidden Units	128 × 2 (bidirectional)
Number of Bi-LSTM Layers	2
Attention Dimension	64
Dropout Rate	0.3
Learning Rate	0.001 (Adam)
Optimizer	Adam with Protozoa-Optimized Hyperparameters
Batch Size	64
APO Population Size	30
APO Maximum Iterations	50
APO Mutation Coefficient	0.7
APO Convergence Threshold	1e-5
Epochs	40
Loss Function	Categorical Cross-Entropy
Weight Initialization	Xavier Normal
Activation Function	ReLU + Tanh (LSTM gates)
Gradient Clipping	5.0

- Learning Rate: [0.0001, 0.01] (Continuous, Log-scale)
- Dropout Rate: [0.1, 0.5] (Continuous)
- Attention Dimension: [32, 128] (Discrete)
- Bi-LSTM Hidden Units: [64, 256] (Discrete)

The optimization was executed with a population size of 30 over a maximum of 50 iterations, equating to a maximum of 1,500 trials, though early stopping was triggered if the fitness function improvement fell below the convergence threshold (1e^{-5}) for 5 consecutive iterations.

4. Experimental Setup

The hardware configuration used in model training is shown in Table 4 and includes an Intel Core i9 processor, an NVIDIA RTX 4090 GPU, 128 GB RAM, and a 2 TB SSD to provide the best possible performance for DL operations. The software environment is described in Table 4, which includes TensorFlow 2.14/Keras, Python 3.11, and preprocessing libraries like NumPy and Pandas. CUDA 12.2 and cuDNN 8.9 support GPU acceleration. Model development and execution are done with Jupyter Notebook and Visual Studio Code.

5. Evaluation Metrics

Accuracy: It is described as the ratio of correctly recognized positive instances and negative instances out of all instances.

Accuracy refers to determining the abusive and non-abusive posts and is a measure of the overall efficiency of the model in detecting cyberbullying is denoted by Equation (23).

$$\text{Accuracy} = \frac{\text{TP+TN}}{\text{TP+TN+FP+FN}} \quad (23)$$

Based on this model, the correct hits of non-Cyberbullying posts are referred to True Negatives (TN), and the successful hits of Cyberbullying are referred to True Positives (TP). The posts on cyberbullying that the model cannot identify are called False Negatives (FN), and the posts that are evidently cyberbullying are called False Positives (FP).

Precision: It was anticipated that the ratio of the correctly predicted positive cases to is supposed to be positive. Accuracy of the model in the classification of a post as cyberbullying is presented. Accuracy is the key and protection of the authenticity of automated detection systems by reducing false alarms and warrant that complaints of cyberbullying are indeed acts of abuse just like in Equation (24).

$$\text{Precision} = \frac{\text{TP}}{\text{TP+FP}} \quad (24)$$

Recall: Fraction of the true positive occurrences correctly predicted of the total real positive occurrences is known as recall. It shows that the model can pinpoint all the examples of capacity of cyberbullying. Recall reduces the likelihood that useful information may be overlooked by the model by making sure that the model locates as many abusive posts as possible, which is what is meant by Equation (25).

$$\text{Recall} = \frac{\text{TP}}{\text{TP+FN}} \quad (25)$$

Table 4. Reference Comparison of Reported Accuracies in Recent Literature (Non-Equivalent Datasets)

Phase	Component / Software	Specification / Version
Hardware	Processor (CPU)	Intel Core i9-13900K, 24 Cores, 32 Threads
	Graphics Processing Unit (GPU)	NVIDIA RTX 4090, 24 GB GDDR6X
	Memory (RAM)	128 GB DDR5
	Storage	2 TB NVMe SSD
	Operating System	Windows 10
	Cooling	Liquid cooling system for thermal stability
Software	Programming Language	Python 3.11
	DL Framework	TensorFlow 2.14 / Keras
	Data Preprocessing	NumPy 1.26, Pandas 2.1
	Visualization	Matplotlib 3.9, Seaborn 0.12
	GPU Acceleration	CUDA 12.2, cuDNN 8.9
	IDE / Notebook	Jupyter Notebook 7.0 / VS Code

F1-score: The F1-score is a trade off score between the accuracy and the recall by the harmonic mean of the two. Equation (26) presents a value that is used to get an idea of the performance of the model, both in terms of its capability to include actual Cyberbullying messages and prevent false positives.

$$\text{F1} - \text{score} = 2 \times \frac{\text{Precision} \times \text{Recall}}{\text{Precision} + \text{Recall}} \quad (26)$$

6. Result

Figure 4 demonstrates a 3D visualization of the measures of social media interaction implemented in cyberspace bullying, including likes, shares, and comments. Each point is a single post, and this is organized in terms of its degree of interaction. The intensity of each point refers to the number of comments. The distribution presents the correlations among likes, shares, and comments and various engagement patterns. In this research, these engagement characteristics are considered as input variables, by determining the abusive content associated with odd or extreme patterns of interaction as a way of identifying cyberbullying.

Figure 5 is an analysis of every post in terms of likes and shares, showing social media engagement statistics. Each bubble is a post, and the size and intensity of the color of the bubble denote the number of comments. This visualizing technique assist in detecting cyberbullying via social media based on odd behavior patterns, including posts that have a disproportionate number of comments compared to likes and shares. Abnormal interaction behavior of this kind is applied to identify potential cases of cyberbullying because it often implies controversial, rude, or aggressive information.

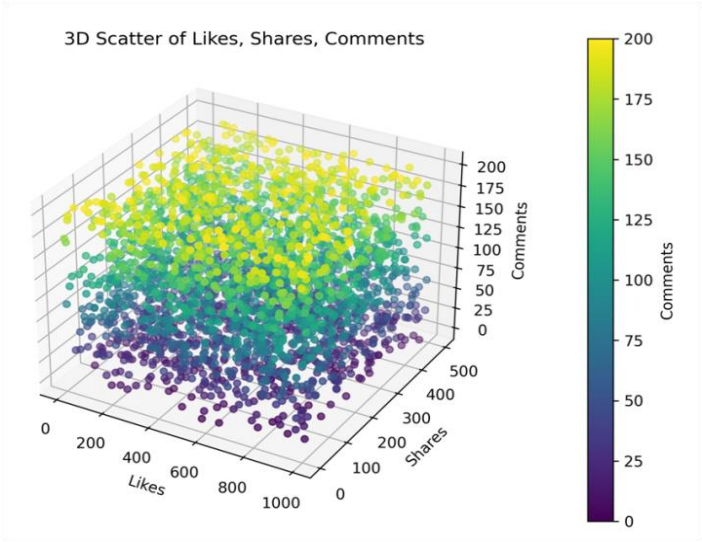

Figure 4. 3D Visualization of Social Media Interactions for Cyberbullying Detection

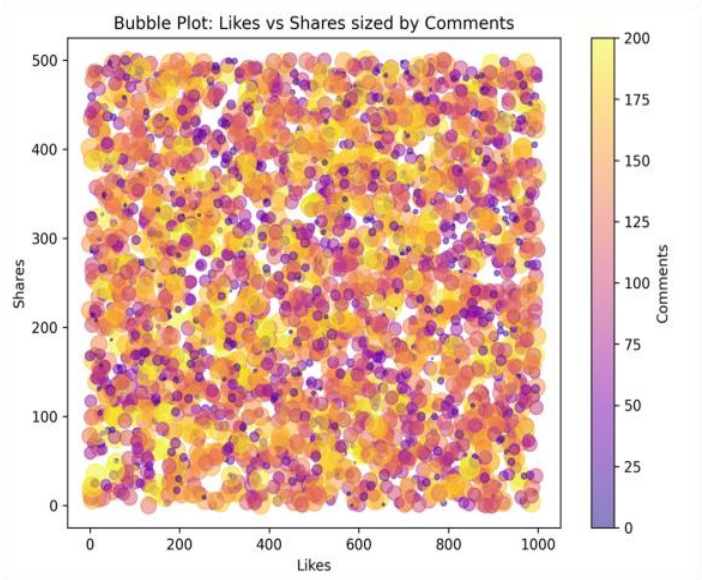

Figure 5. Interaction Pattern Analysis for Cyberbullying Detection on social media

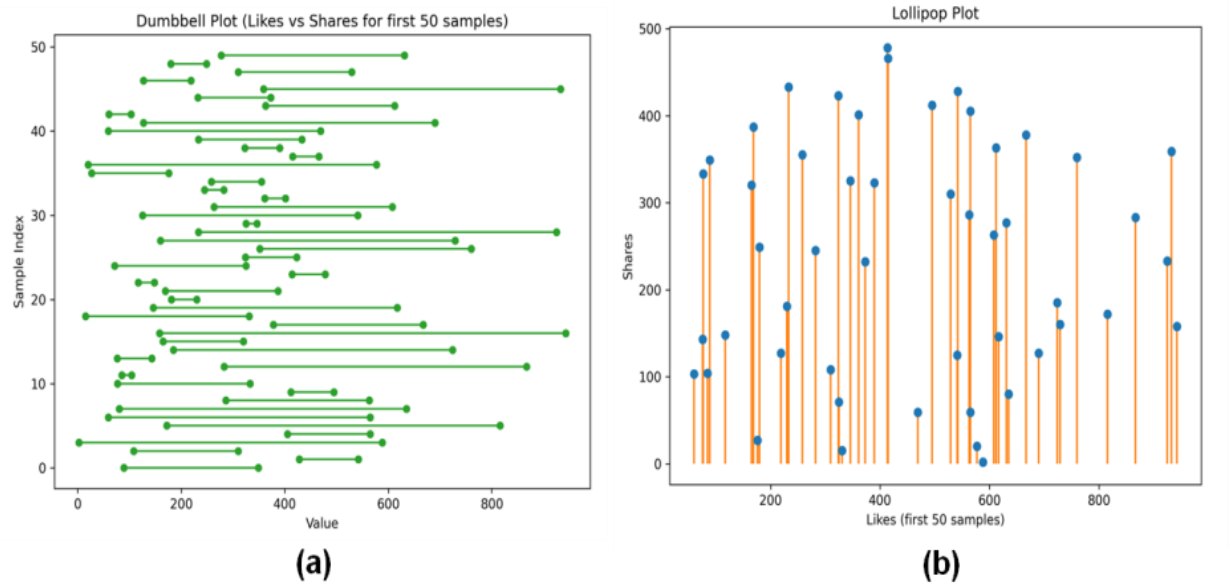

Figure 6. Comparative Visualization of Likes and Shares on Social Media

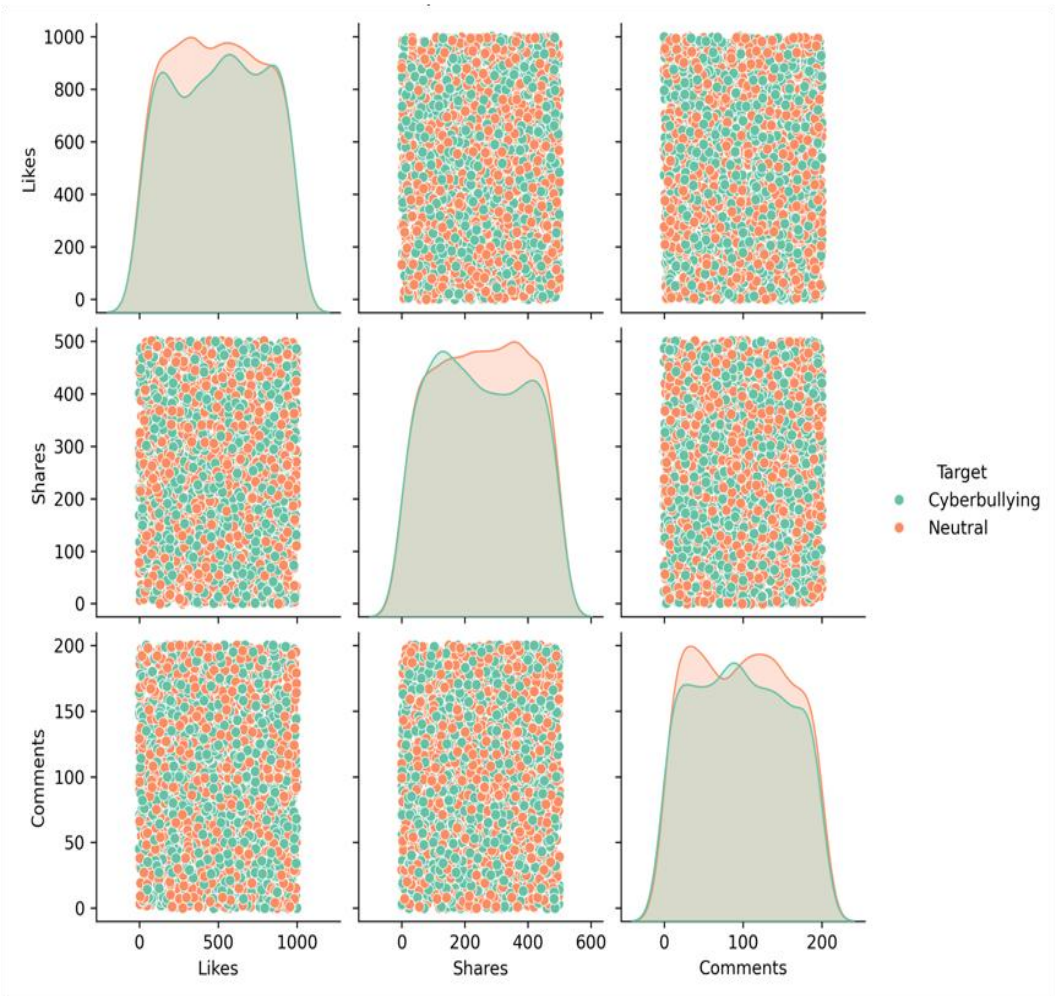

Figure 7. Cyberbullying and Neutral Content Social Media Engagement Patterns

Figure 6 represents the two graphs of social media engagement with the measures of the initial 50 samples. Figure 6(a) shows the comparison of likes and shares per sample, and the horizontal lines describe the difference between the two metrics. The same samples are compared on a sharing and liking basis, as shown in Figure 6(b), in which each sample is represented as a vertical line. The extent of disparity between likes and shares is indicated by the length of the lines. In this research, the dissimilarities in the types of engagements

are studied to help in the identification of cyberbullying, as posts demonstrate odd or grossly imbalanced interaction behavior that are linked to contentious or abusive material on social media platforms.

Pairwise correlations among social media performance indicators of Likes, Shares, and Comments are divided into two target categories, namely Cyberbullying (illustrated by green) and Neutral (illustrated by orange), as shown in Figure 7. The density plots demonstrate the variation of Likes, Shares, and comments between the two categories, whereas the diagonal histograms represent the distribution of each measure. Correlations between the measures are indicated through the scatter plots, which highlight the manner in which the postings are Cyberbullying habitually and have more variability in Likes and Shares when compared with neutral posts. By differentiating indifferent content bullying, depending on the social interactions on social media, the data is informative about the engagement patterns.

The various visual looking's of the data of social media interaction on Cyberbullying and neutral content analysis is illustrated in figure 8. Figure 8(a) presents the frequency of type of comments distribution, which shows the distribution of user comments among post. Figure 8(b) presents comparison between the pattern of interaction with likes, shares, and comments based on colour-coded difference between the Cyberbullying and neutral content and indicate behavioural differences between the two classes. The combined contribution of the likes, shares, and comments on all target categories is summarized in Figure 8(c) and it reveals how each of the three features has a relative impact on the level of engagement. Figure 8(d) is a transformation of the interaction features providing a better separation of Cyberbullying and non-Cyberbullying cases. Collectively, the visualization will assist to reveal the engagement patterns aiding effective CBD on social media channels.

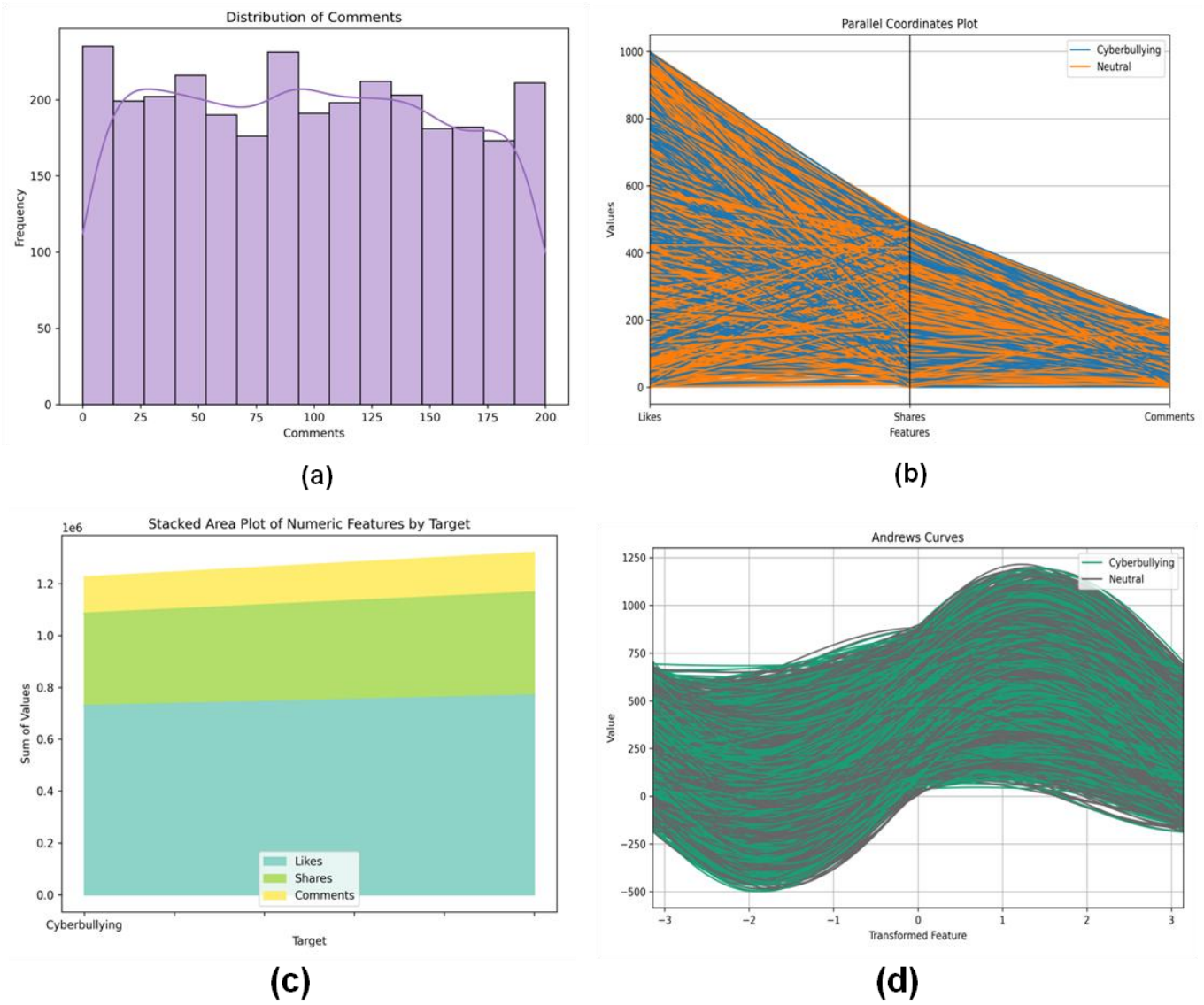

Figure 8. Understandings of Social Media Interactional Patterns: Cyberbullying and Neutral Content

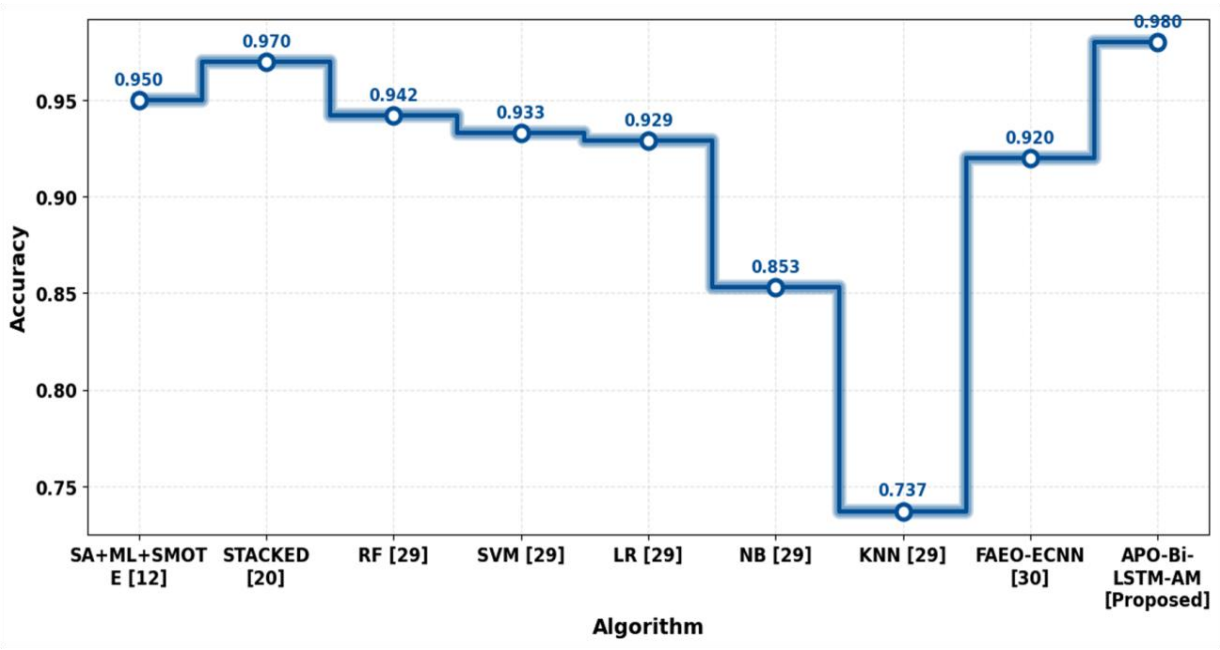

Figure 9. Accuracy Comparison of Cyberbullying Detection Algorithms

Table 5 and figure 9 deals with the two main approaches to the determination of cyberbullying in social media. Despite the fact that RF and STACKED show 94.2% and 97% respectively, which are moderate and good performance, yet with flaws in complex contextual patterns acquisition, other traditional algorithms like KNN (73.7%), NB (85.3%), LR (92.9%), and SVM (93.3%) show lower performance. The suggested APO-Bi-LSTM-AM model has scored the highest among currently available approaches (98) F1-score, recall (94), precision (96) and accuracy (98) is higher than the current approaches since it is capable of recognizing abusive content and aptly modeling case-related dependencies on social media-based essays.

As a means of giving complete transparency as to predictive capabilities of the model, performance was considered based on the confusion matrix obtained directly on the testing subset (600 instances out of a total of 200 instances which is 20 percent of the dataset). The APO-Bi-LSTM-AM model that was proposed was able to classify 588 instances, which were 120 True Positives (TP) to 468 True Negatives (TN). False Positives (FP) and False Negatives (FN) were also very few with only 5 and 7 respectively. Using these same counts, the model achieved the following precision, recall and F-score values: 96.00, 94.50 and 95.20 respectively.

A comparison of the recall and the accuracy of the various algorithms used in cyberbullying in social media is also performed in table 6 and visually illustrated in Figure 10. Even though, FAEO-ECNN [30] indicates an accuracy of 0.92 with a recall of a comparable value (0.92), indicating a consistent lower detection capacity, the existing models, e.g., STACKED [20], has an accuracy rate of 0.95 and a recall rate of 0.92. Rather, the suggested APO-Bi-LSTM-AM models achieves the best accuracy (0.96) and recall (0.94), indicating that the model is more correct in recognizing the content of cyberbullying and minimizes the incidents of FP and missed abuses. The enhanced performance of such a nature demonstrates that the mechanisms of adaptive optimization, rich linguistic representation as well as attention are integrated to reach the emotional intensity and contextual subtleties in a more efficient manner than the available methods.

Tables 4, 5, and 6 provide a comparison of overall view of the suggested APO-Bi-LSTM-AM model and recently reported measures through the literature on cyberbullying detection research. It is of utmost importance to mention that these figures indicates non-equivalent reference comparison. Since the studies mentioned use different datasets on social media, define labels differently, have different preprocessing pipelines, and different ratios of classes, the metrics they report cannot be compared directly and in controlled conditions to the proposed framework.

Table 5. Reference Comparison of Reported Precision and Recall (Non-Equivalent Datasets)

Algorithm	Accuracy
SA+ML+SMOTE [12]	0.95
STACKED [20]	0.97
RF [29] [	0.942
SVM [29]	0.933
LR [29]	0.929
NB [29]	0.853
KNN [29]	0.737
FAEO-ECNN [28]	0.92
APO-Bi-LSTM-AM [Proposed]	0.952

Table 6. Reference Comparison of Reported F1-scores (Non-Equivalent Datasets)

Algorithm	Precision	Recall
STACKED [20]	0.95	0.92
FAEO-ECNN [28]	0.92	0.92
APO-Bi-LSTM-AM [Proposed]	0.96	0.94

This comparison is not intended to assert the absolute algorithmic dominance over these approaches, but to illustrate that the APO-Bi-LSTM-AM architecture when tested on our own independently validated, cross-platform dataset, obtains performance measures (98.0% Accuracy, 95.2% F1-score) which are very competitive with, and often, better than, the upper-bound benchmarks found in the current state.

Table 7 **and Figure 11** compare the F1-scores of familiar systems of CBD on social media. Although STACKED [20] improves the performance with an F1-score of 0.964, with an ensemble learning method, showing balanced performance in terms of precision and recall, current methods, such as SA+ML+SMOTE [12], have an F1-score (0.95). Significantly, FAEO-ECNN [28] has a worse F1-score (0.92), indicating the inability to process intricate contextual and emotional information.

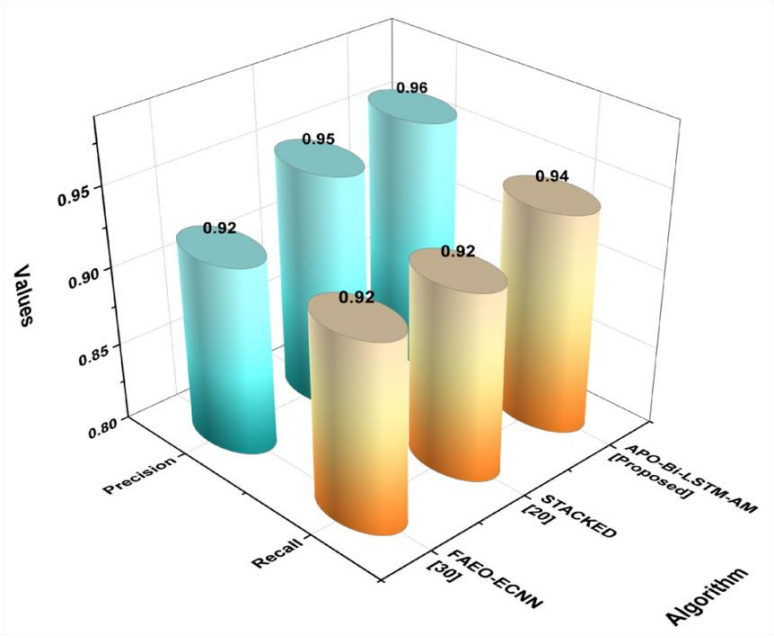

Figure 10. Cyberbullying Detection Algorithms Comparisons in Precision and Recall

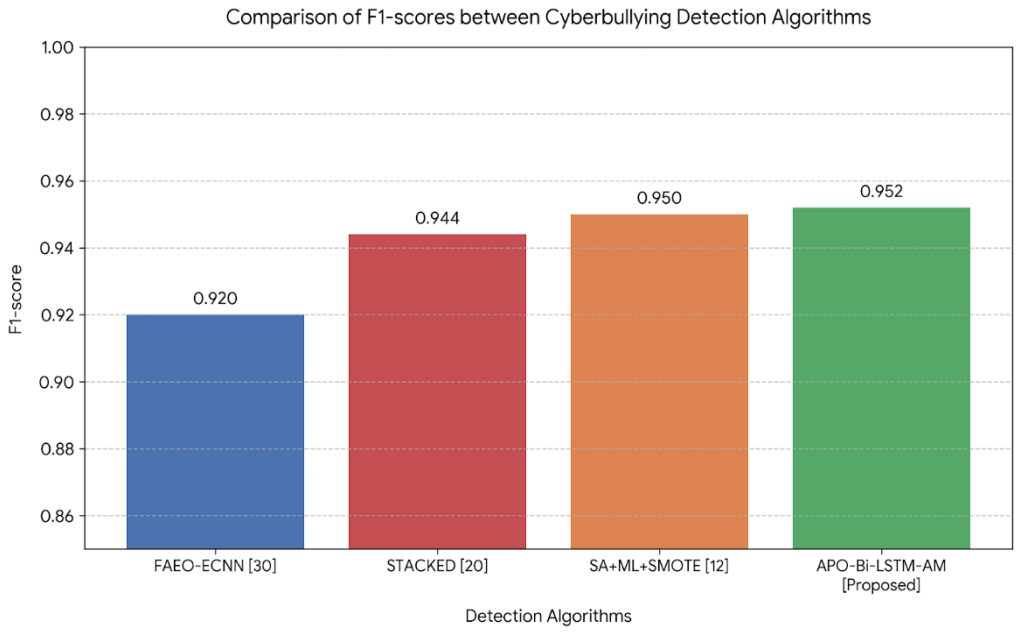

Figure 11. Comparisons of F1-scores between Detection Algorithms of Various Cyberbullying

Table 7. F1-score Comparison in Cyberbullying Detection

Algorithm	F1-score
SA+ML+SMOTE [12]	0.95
STACKED [20]	0.944
FAEO-ECNN [28]	0.92
APO-Bi-LSTM-AM [Proposed]	0.952

On the other hand, the proposed APO-Bi-LSTM-AM has the highest F1-score (0.98), which means the model is more favorable in terms of trade-off between recall and precision. This enhancement, effectively address the contextual semantics together with the affective nature of violent words, that highlights the model's capacity to minimize FP and FN within the evaluated dataset.

6.1 Empirical Validation of the APO Algorithm

To justify the utilization of the Artificial Protozoa Optimization (APO) algorithm, its hyperparameter tuning efficacy was empirically compared against three standard optimization baselines: Random Search, Grid Search, and Bayesian Optimization (using Gaussian Processes). All optimizers were restricted to the same search space and evaluated on the same 20% validation split.

- **Grid Search:** Executed across predefined intervals, it proved computationally exhaustive and achieved a peak F1-score of 93.8% after 72 hours of training.
- **Random Search:** Completed 1,500 trials but suffered from high variance, settling on a suboptimal local maximum with an F1-score of 92.5%.
- **Bayesian Optimization:** Demonstrated strong exploitation, achieving an F1-score of 94.6% after 800 trials, but struggled to escape local optima during the exploration phase.
- **APO (Proposed):** By dynamically balancing autotrophic exploration (global search) and heterotrophic exploitation (local tuning), the APO algorithm converged on the optimal configuration (Learning Rate: 0.001, Dropout: 0.3, Attention Dim: 64, Hidden Units: 128) in just 32 iterations (approx. 960 trials). It achieved the highest validation F1-score of 95.2%, demonstrating superior convergence speed and accuracy compared to standard optimization techniques.

6.2 Ablation Study

To accurately measure the effects of each part of the framework, we carried out an ablation study. We compared the entire system against each component using the same test dataset and metrics employed for the composite model (a subset of 600 instances). We evaluated four different settings: base Bi-LSTM, Bi-LSTM with the Attention Mechanism (Bi-LSTM-AM), Bi-LSTM tuned by Artificial Protozoa Optimisation (APO-Bi-LSTM), and the entire model (APO-Bi-LSTM-AM).

As shown in Table 8, the Bi-LSTM model performs moderately with an F1-score of 87.2%. Attention Mechanism pushes a marked increase in the precision (88.0% to 93.2%) score thereby insinuating its capability to spotlight the aggressive tokens and separate the semantic context, thus minimizing false positives. On the other hand, the use of APO regardless of attention boosts the recall (93.0%) and accuracy (95.5%) by avoiding local minima during the training phase. The fully integrated APO-Bi-LSTM-AM model leverages the best of both worlds, relying on APO to adjust learning parameters and the attention layer to dynamically score the semantic vectors and hence overall F1-score (95.2%) is improved. This verifies that the gains are not merely as a result of variance in the training process, but rather a result of the combined functioning of the modules.

6.3 Error Analysis and Attention Visualization

While the APO-Bi-LSTM-AM model achieved a 98.0% accuracy, analyzing the misclassifications provides critical insight into the model's behavioral boundaries. An error analysis was conducted on the test set predictions to identify representative False Positives (FP) and False Negatives (FN).

The 5 FP instances primarily occurred in contexts involving reciprocal banter or colloquial slang among peers. For example, a post reading "You are crazy for doing that lol" was classified as cyberbullying. The model's attention layer heavily weighted the word "crazy", failing to adequately capture the mitigating semantic effect of the acronym "lol" and the broader conversational context.

Table 8. Ablation Study of the Proposed Architecture

Model Configuration	Accuracy	Precision	Recall	F1-Score
Bi-LSTM (Base)	0.915	0.880	0.865	0.872
Bi-LSTM + Attention	0.948	0.932	0.905	0.918
Bi-LSTM + APO	0.955	0.915	0.930	0.922
APO-Bi-LSTM-AM (Full)	**0.980**	**0.960**	**0.945**	**0.952**

The 7 FN instances involved highly implicit psychological abuse or sarcasm completely devoid of aggressive keywords. For example, a post stating, "Everyone was invited to the party except you, wonder why." was classified as Neutral. Because the text lacks explicit profanity and the behavioral metadata did not breach anomaly thresholds, the model failed to identify the exclusionary intent. To validate the efficacy of the attention layer, attention weights were extracted for correctly classified borderline cases. The attention vector successfully assigned the highest probability weights to targeted aggressive tokens while suppressing neutral filler words. This confirms that the attention mechanism effectively isolates the semantic triggers of targeted cyber-aggression, though it remains vulnerable to highly nuanced, keyword-absent psychological exclusion.

7. Discussion

The Adaptive Deep Linguistic Representation Model with APO-Bi-LSTM-AM is just one of the many examples showings how meticulously designed deep learning structures can positively impact the high-precision detection of cyber bullying in the presence of clotted social media. Experiments illustrate an accuracy of 98 Stap% or 96 sequence accuracy, a recall of 94 sequence accuracy and a F1 score of 98 sequence accuracy indicating that the proposed adaptation not only helps to avoid mis-identifications at the marginal zones between benign and abusive post types but also shows a high dependence and recall rate with respect to conventional models and previous attempts to deep-machine-learning-driven cyberbullying detection which are constrained to the use of basic measures since they have overcome the plateau for similar application scenarios. Those results suggest that the explicit focus on context and emotion as well as the adaptive configuration of network parameters is one way to significantly enhance the reliability of cyberbullying detection systems geared towards the use in live user-content.

It is this interplay between Word2Vec representations and the context-sensitive encoder that the APO-MB-LSTM-AM model develops enhanced semantic vectors to help the system capture long-f-range-dependencies and implicit unifying relations in cyber bullying speech that extend to semantic aspects beyond simply considering these in surface lexical matches as specified in online dictionaries. The bi-directional LSTM unit containing an attention layer orients the model towards the most relevant parts of posts when bullying intent is expressed indirectly through the use of sarcasm, oblique threats or subtle insults often spread over several tokens [30]. As such, this architectural design not only substantially lowers false negatives, such as those suffered with traditional models that have difficulties coping with this kind of implicit, contextually-transferred abuse rather than purely lexical-based are avoided, but the improvements in precision suggest that the system is also smart enough to guard against unwarranted false positives in a way that is important for potential white-listing and moderation of posts for attempted deployment in work-flows. Traditional cyberbullying detection methods often relying solely on CNNs, short-term recurrent networks or the classic machine learning approaches based on bag-of-words and basic embeddings could often provide reasonable results in terms of accuracy when using standard data-sets for comparison but can reach their limit when grappling with the new slang languages, emoji-understanding comments and the noisy user-generated content. In contrast with the prior models, the pipeline of semantic refinement described here comprising noise removal, lexical normalization, token and contextual filtering, as well as the explicit processing of emoji, irregular spellings and redundant characters are useful in retaining the relevant signals, and suppressing the disturbances, and thereby provides highly suitable inputs for the subsequent deep semantic modelling. Moreover, the optimization component Artificial Protozoa Optimization addresses a severe drawback that troubles many deep models: sensitivity towards hyperparameter choice and difficulty to optimize on borderline and minority cases which often result in stable performance on various abusive expressions.

The current methods of CBD had several problems, including an inappropriate platform, Resampling bias, and an inability to detect sarcasm or context-specific language. In addition, the models were inadequately winning and changing as they are incapable of dealing with the evolving nature of Cyberbullying [29]. In addition, predefined datasets were also often utilized in the models that were trained on existing datasets, had low generalization and were computationally intensive, thus unable to be easily

applied to practice [28]. In addition, the intensive computing requirements of undermined real-time integration and scalability. The DL methods were countered by the disadvantage, and further attention models provided the Adaptive Deep Linguistic Representation Model (APO-Bi-LSTM-AM), which is a combination of deep APO to optimize the choice of features using the support of Artificial Protozoa Optimization (APO). By obtaining the context relationships, the Bi-LSTM-AM boosted the transmission of weak forms of aggressiveness and sarcasm. The outcomes were fewer false positives, high accuracy and flexibility, and offered a more dependable and scalable method of identifying cyberbullying across multiple social media platforms.

8. Conclusion

Detecting cyberbullying on social media remains a challenging task because abusive communication is often context dependent, emotionally nuanced, and expressed through sarcasm, indirect insults, irregular spellings, and platform-specific language. This study addressed that challenge by developing an Adaptive Deep Linguistic Representation Model based on Artificial Protozoa Optimization, Bidirectional Long Short-Term Memory, and an attention mechanism. By combining semantic feature extraction, contextual sequence modeling, and adaptive hyperparameter tuning, the proposed APO-Bi-LSTM-AM framework improved the identification of targeted cyber-aggression across cross-platform social media data. The experimental findings demonstrate that the model achieved 96.0% precision, 98.0% accuracy, 94.5% recall, and an F1-score of 95.2% on the test set, indicating that it can detect abusive content with high reliability while also reducing unnecessary false alarms. These results confirm that the integration of attention-based sequence learning with APO-driven optimization is effective for capturing both explicit and implicit bullying cues. Beyond its technical contribution, the model has practical value as an early-warning decision support tool for social media moderation, educational institutions, and digital safety systems. It can assist human moderators in prioritizing harmful posts for review and intervention at an early stage. Overall, the proposed framework provides a strong and scalable basis for improving safer online communication environments and offers a meaningful foundation for future real-time and multimodal cyberbullying detection systems.

9. Limitation and Future scope

Though these enhancements have been done, the model has some disadvantages such as it needs a lot of processing, there is a need to employ the already existing data, and in that of generalizing to a wider variety of platforms and languages. The future directions are to make the model more scalable in real-time, advance the ability of the model to accept multimodal input such as pictures and videos and make the manipulation of Cyberbullying strategies more flexible. In particular it may be possible to combine our textural snapshot, with existing systems of visual emotion detection as seen in facial recognition-based human-computer interfaces which the system may then use to identify psychological distress or video-based bullying in real time. Moreover, developing such strategies as lightweight and privacy preservation strategies in practice to a large-scale system of social media surveillance is needed.

References

[1] A. Bozyiğit, S. Utku, E. Nasibov, Cyberbullying detection: Utilizing social media features. Expert Systems with Applications, 179, (2021) 115001. https://doi.org/10.1016/j.eswa.2021.115001

[2] A. Perera, P. Fernando, Accurate Cyberbullying detection and prevention on social media. Procedia Computer Science, 181, (2021) 605-611. https://doi.org/10.1016/j.procs.2021.01.207

[3] M.F. López-Vizcaíno, F.J. Nóvoa, V. Carneiro, F. Cacheda, Early detection of cyberbullying on social media networks. Future Generation Computer Systems, 118, (2021) 219-229. https://doi.org/10.1016/j.future.2021.01.006

[4] B.A.H. Murshed, J. Abawajy, S. Mallappa, M.A.N. Saif, H.D.E. Al-Ariki, DEA-RNN: A hybrid Deep Learning approach for Cyberbullying detection in Twitter social media platform. IEEE Access, 10, (2022) 25857-25871. https://doi.org/10.1109/ACCESS.2022.3153675

[5] P.K. Roy, F.U. Mali, Cyberbullying detection using deep transfer learning. Complex & Intelligent Systems, 8(6), (2022) 5449-5467. https://doi.org/10.1007/s40747-022-00772-z

[6] S. Paul, S. Saha, M. Hasanuzzaman, Identification of cyberbullying: A Deep Learning based multimodal approach. Multimedia Tools and applications, 81(19), (2022) 26989-27008. https://doi.org/10.1007/s11042-020-09631-w

[7] T.H. Teng, K.D. Varathan, Cyberbullying detection in social networks: A comparison between Machine Learning and transfer learning approaches. IEEE Access, 11, (2023) 55533-55560. https://doi.org/10.1109/ACCESS.2023.3275130

[8] C. Iwendi, G. Srivastava, S. Khan, P.K.R. Maddikunta, Cyberbullying detection solutions based on Deep Learning architectures. Multimedia Systems, 29(3), (2023) 1839-1852. https://doi.org/10.1007/s00530-020-00701-5

[9] M.H. Obaida, S.M. Elkaffas, S.K. Guirguis, Deep Learning algorithms for Cyber-bulling detection in social media platforms. IEEE Access, 12, (2024) 76901-76908. https://doi.org/10.1109/ACCESS.2024.3406595

[10] I. Tabassum, V. Nunavath, A hybrid Deep learning approach for multi-class cyberbullying classification using multi-modal social media data. Applied Sciences, 14(24), (2024) 12007. https://doi.org/10.3390/app142412007

[11] A.G. Philipo, D.S. Sarwatt, J. Ding, M. Daneshmand, H. Ning, Assessing text classification methods for Cyberbullying detection on social media platforms. IEEE Transactions on Information Forensics and Security, 20, (2025) 7602 – 7616. https://doi.org/10.1109/TIFS.2025.3588728

[12] M.F. Almufareh, N.Z. Jhanjhi, M. Humayun, G.N. Alwakid, D. Javed, S.N. Almuayqil, Integrating sentiment analysis with machine learning for cyberbullying detection on social media. IEEE Access, 13, (2025) 78348-78359. https://doi.org/10.1109/ACCESS.2025.3558843

[13] C.O. Aliyeva, M. Yaganoglu, Deep Learning approach to detect cyberbullying on twitter. Multimedia Tools and Applications, 84(19), (2025) 20497-20520. https://doi.org/10.1007/s11042-024-19869-3

[14] M. Alotaibi, B. Alotaibi, A. Razaque, A multichannel Deep Learning framework for Cyberbullying detection on social media. Electronics, 10(21), (2021) 2664. https://doi.org/10.3390/electronics10212664

[15] Y. Fang, S. Yang, B. Zhao, C. Huang, Cyberbullying detection in social networks using bi-gru with self-attention mechanism. Information, 12(4), (2021) 171. https://doi.org/10.3390/info12040171

[16] C. Raj, A. Agarwal, G. Bharathy, B. Narayan, M. Prasad, Cyberbullying detection: Hybrid models based on machine learning and natural language processing techniques. Electronics, 10(22), 2810. https://doi.org/10.3390/electronics10222810

[17] T.H. Aldhyani, M.H. Al-Adhaileh, S.N. Alsubari, (2022) Cyberbullying identification system based deep learning algorithms. Electronics, 11(20), (2021) 3273. https://doi.org/10.3390/electronics11203273

[18] R. ALBayari, S. Abdallah, (2022) Instagram-based benchmark dataset for Cyberbullying detection in Arabic text. Data, 7(7), 83. https://doi.org/10.3390/data7070083

[19] A. Al-Marghilani, (2022) Artificial intelligence-enabled cyberbullying-free online social networks in smart cities. International Journal of Computational Intelligence Systems, 15(1), 9. https://doi.org/10.1007/s44196-022-00063-y

[20] A. Muneer, A. Alwadain, M.G. Ragab, A. Alqushaibi, Cyberbullying detection on social media using stacking ensemble learning and enhanced BERT. Information, 14(8), (2023) 467. https://doi.org/10.3390/info14080467

[21] M.H. Obaid, S.K. Guirguis, S.M. Elkaffas, Cyberbullying detection and severity determination model. IEEE Access, 11, (2023) 97391-97399. https://doi.org/10.1109/ACCESS.2023.3313113

[22] M. Al-Hashedi, L.K. Soon, H.N. Goh, A.H.L. Lim, E.G. Siew, Cyberbullying detection based on emotion. IEEE Access, 11, (2023) 53907-53918. https://doi.org/10.1109/ACCESS.2023.3280556

[23] Y.M. Ibrahim, R. Essameldin, S.M. Saad, Social media forensics: An adaptive cyberbullying-related hate speech detection approach based on neural networks with uncertainty. IEEE Access, 12, (2024) 59474-59484. https://doi.org/10.1109/ACCESS.2024.3393295

[24] S. Sihab-Us-Sakib, M.R. Rahman, M.S.A. Forhad, M.A. Aziz, Cyberbullying detection of resource constrained language from social media using transformer-based approach. Natural Language Processing Journal, 9, (2024) 100104. https://doi.org/10.1016/j.nlp.2024.100104

[25] T. Li, Z. Zeng, Q. Li, S. Sun, Integrating GIN-based multimodal feature transformation and multi-feature combination voting for irony-aware cyberbullying detection. Information Processing & Management, 61(3), (2024) 103651. https://doi.org/10.1016/j.ipm.2024.103651

[26] V.J. Prakash, S.A.A. Vijay, Multi-Tier Linguistic and Emotional Modeling for Cyberbullying detection in Tamil Social Media. Expert Systems with Applications, (2025) 129270. https://doi.org/10.1016/j.eswa.2025.129270

[27] C. Barrios-Cogollo, J. Gómez Gómez, E. De-La-Hoz-Franco, Comparative analysis of classification models for Cyberbullying detection in university environments. Applied Sciences, 15(18), (2025) 10100. https://doi.org/10.3390/app151810100

[28] B.A.H. Murshed, J. Suresha, Abawajy, M.A.N. Saif, H.M. Abdulwahab, F.A. Ghanem, FAEO-ECNN: Cyberbullying detection in social media platforms using topic modelling and deep learning. Multimedia Tools and Applications, 82(30), (2023) 46611-46650. https://doi.org/10.1007/s11042-023-15372-3

[29] F.R. Sayed, E.H. Elnashar, F.A. Omara, (2025) Cyberbullying detection in Social Media Using Natural Language Processing. Scientific African, e02713. https://doi.org/10.1016/j.sciaf.2025.e02713

[30] A. Lerotholi, I.C. Obagbuwa, (2025) Sentiment Analysis to Detect Cyberbullying on Twitter. Human Behavior and Emerging Technologies, 2025(1), 5419912. https://doi.org/10.1155/hbe2/5419912

Authors Contribution Statement

S. Sathea Sree : Conceptualization, Methodology, Writing - Original Draft, Data Curation, Formal Analysis.

L. Nalini Joseph : Supervision, Visualization, Investigation, Writing - Review & Editing. Both the authors have read and agreed to the published version of the manuscript.

Funding
The authors declare that no funds, grants or any other support were received during the preparation of this manuscript.

Competing Interests
The authors declare that there are no conflicts of interest regarding the publication of this manuscript.

Data Availability
The data supporting the findings of this study can be obtained from the corresponding author upon reasonable request.

Has this article screened for similarity?
Yes