

EOT-Net: An Emoji-Aware Obfuscated Toxic Chat Detection and Moderation Framework for Streaming Platforms

OPEN ACCESS

Received: 27/06/2026

Accepted: 07/08/2026

Published: 05/09/2026

S Sankari¹, S Silvia Priscila^{1*}

¹ Department of Computer Applications, Bharath Institute of Higher Education and Research, Chennai, Tamil Nadu, India

Citation: Sankari S, Silvia Priscila S (2026) EOT-Net: An Emoji-Aware Obfuscated Toxic Chat Detection and Moderation Framework for Streaming Platforms. Indian Journal of Science and Technology 19(32): 2264-2277. <https://doi.org/10.17485/IJST/19i32.977>

* **Corresponding author.**

silviaprisila.cbcs.cs@bharathuniv.ac.in

Funding: None

Competing Interests: None

Copyright: © 2026 Sankari & Silvia Priscila. This is an open access article distributed under the terms of the [Creative Commons Attribution License](https://creativecommons.org/licenses/by/4.0/), which permits unrestricted use, distribution, and reproduction in any medium, provided the original author and source are credited.

Published By Indian Society for Education and Environment ([iSee](https://www.indjst.org/))

ISSN

Print: 0974-6846

Electronic: 0974-5645

Abstract

Objectives: This research introduces a system that detects toxic content in a hidden manner in real time communication services by using an intelligent system that detects toxic content in a conversation. The system would try to solve the problems that the current moderation approach had, and recognize the toxic phrases in the message, substitute them with symbols and characters (emoji/symbols) and leave the automatic moderation function unchanged.

Method: This study uses an EOT-Net (Emoji-Aware Obfuscated Toxic Chat Detection and Moderation Framework), which involves ORL (Obfuscation Recovery Layer), EAE (Emoji-Aware Semantic Encoder), a module of the Contextual Feature Extraction based on the RoBERTa network, a BiLSTM network and a Multi-Head Attention mechanism. Adaptive Moderation Engine provides warnings, action suppression and action blocking due to toxicity. This framework was tested with an integrated Kaggle dataset comprising of roughly 435,000 chat messages that are toxic and non-toxic. **Findings:** The experiments have demonstrated the usefulness of the framework proposed for detection of different types of toxic communications that already exist. The accuracy, precision, recall, f1 score and AUC-ROC of EOT-Net are 97.14%, 96.52%, 96.03%, 96.27%, and 0.984 respectively, outperforming the basic models (CNN, BiLSTM, BERT, and RoBERTa). The results demonstrate the usefulness of obfuscation recovery and the knowledge of the meaning of emojis to improve the effectiveness of the hidden toxicity detection and moderation. **Novelty:** The novelty of this research is the obfuscation recovery, emoji-aware semantic encoding, contextual transformer learning, attention-based toxicity extraction and adaptive moderation. By using emojis to detect hidden abusive expressions and threats, EOT-Net outcompetes the existing methods which are only able to detect the explicit abusive expressions, so that the automated real-time moderation of next generation content moderation systems is possible.

Keywords: Toxic Chat Detection; Emoji-Aware Learning; Obfuscated Text Recovery; Deep Learning; RoBERTa; BiLSTM; Multi-Head Attention; Content Moderation; Hate Speech Detection; Streaming Platforms

1 Introduction

The amount of content produced and shared per second by users has changed drastically due to the online streaming services, multiplayer games, Social Media and real-time communication in general⁽¹⁾. While they are all social engagement and interaction and collaboration platforms, they have also been abused in the spread of hate speech, cyberbullying, harassment, identity-based abuse and personal threats^{(2), (3)}. All of these negative interactions have a negative effect on the mental health, trust and credibility of the users of the platform, which is why automated and scalable content moderation is essential.

The first systems that recognize the toxic content were primarily designed with keywords or manually crafted rules, and used traditional machine learning techniques such as Naïve Bayes, Support Vector Machines (SVM) and Random Forest^{(4), (5)}. Computationally efficient but they were not able to cover the contextual semantics and were unable to deal with abusive language in different linguistic forms. The deep contextual representation learning has enabled great advances in toxicity detection in recent years, particularly thanks to the development of Natural Language Processing (NLP). Transformer-based models, which learn the context of text using transformers, have been proven to achieve better results by capturing the context of text in online conversations^{(6), (7), (8), (9), (10), (11), (12), (13), (14)} as well as hybrid models which combine RNNs and attention mechanisms to model the semantic and sequential relationships in conversations.

Despite all these enhancements, none of the current tools to detect toxicity can catch up with new ways in which users are trying to avoid automated moderation. Offensive words are always disguised using substitutions of characters, symbols, numbers, abbreviations and misspellings (e.g., "1d10t", "stup!d", and "l0ser") making the common token-based language models much less effective⁽¹⁵⁾. Additionally, there has been a shift in the use of emojis from representing feelings into conveying sarcasm, insults, threats and offensive intent, as well as using them to represent feelings, on their own and in combination with text⁽¹⁶⁾. As a result, there are models that rely on just textual data that will miss out on the underlying toxicity from conversations ramping emoji usage.

There have been several recent studies that have been researching the detection of toxic language: Adverse Text Normalization, Character Level Representations, Emoji Embeddings and Multimodal Sentiment Analysis. However, most of these techniques are developed to extract hidden information from text or interpret emoji and they're not able to detect recent toxicity signals. Furthermore, few models exist that integrate the context-aware understanding of a language with the methods for adaptively moderating that can be implemented in such a way that it can be acted upon in real-time by platforms such as streaming and gaming platforms. There is not extensive research conducted on a single system which can solve these problems together.

This work presents a EOT-Net (Emoji-Aware Obfuscated Toxic Chat Detection and Moderation Framework), an integrated deep learning framework to address these gaps in the research, for real-time toxic chat detection and moderation. The proposed framework contains of an Obfuscation Recovery Layer (ORL) designed to recover the semantic meaning of expressions and introduced with intentional obfuscation, as well as an Emoji-Aware Semantic Encoder (EAE) for encoding a semantic representation of emojis in the context. The enriched representations are then input into a RoBERTa based contextual feature extractor, followed by a BiLSTM network and two different feature extractors: a Multi-Head Attention network for extracting contextual and sequential toxicity patterns and a semantic toxicity pattern extractor. Finally, an Adaptive Moderation Engine converts the toxicity confidence level into actionable, in-the-moment alerts that can be suppressed and/or blocked in the real world in streaming and gaming environments. This study makes a unique contribution to the literature and it is listed as follows

- Detect toxicity hidden in the form of character substitutions, symbols and any special characters by reconstructing toxic words, through the development of an Obfuscation Recovery Layer (ORL).
- Design and development of Emoji-Aware Semantic Encoder (EAE) for capturing the semantic contribution of emoji and symbolic representation that will facilitate the recognition of the implicit abusive, threatening and offensive communication.
- Incorporating the RoBERTa, BiLSTM and Multi-Head Attention models into a single context learning model to effectively model the toxicity patterns at semantic, contextual and sequential levels to achieve a successful toxic chat classification.
- Design an Adaptive Moderation Engine to turn toxicity confidence scores into flexible real-time actions – such as warning, suppression and blocking – for intelligent content moderation in streaming and social communication platforms.

EOT-Net is a scalable solution that addresses both the obfuscated abusive language and emoji-driven toxicity, offering robust solutions for the next generation of content-moderation systems, which can lead to safer and more inclusive online communication environments.

2 Methodology

The proposed EOT-Net aims to detect toxic content that is hidden using obfuscation or emoji substitution and symbolic representation that are often used in streaming platforms, gaming chats, and social media communication. All these elements are part of this framework that is integrated into one pipeline: normalization of the text, understanding of the meaning of the emojis, deep contextual feature extraction, classification of the toxicity, and adaptive moderation. Fig. 1 shows an overall workflow of EOT-Net. It shows the entire process of EOT-Net for real-time toxic chat detection and moderation. The framework comprises of the following functions: Text Cleaning, Obfuscation Recovery, Emoji Semantic Translation, BiLSTM and Attention-based Contextual Feature Extraction, Toxicity Classification and Adaptive Moderation. Auto warning and suppression and blocking of messages were according to their toxicity level.

2.1 Dataset Description

To present transparency and reproducibility, publicly available datasets were downloaded from open source datasets to facilitate experimentation. The Jigsaw Toxic Comment Classification Dataset⁽¹⁷⁾ is a dataset of user comments that was downloaded from Kaggle with about 160K samples that are labeled for different categories such as toxic, severe toxic, obscene, threat, insult and identity hate. Additionally, the dataset of around 25,000 manually labelled tweets for hate speech and offensive language was added⁽¹⁸⁾ to improve the linguistic diversity and to represent social media offensive language. The data set also included almost 50,000 chatroom and gaming-chat messages as part of a Toxic Chat Messages Dataset to capture toxic communication patterns in real-time conversational contexts. The Jigsaw Toxic Comment Classification Dataset can be found at this URL: <https://www.kaggle.com/c/jigsaw-toxic-comment-classification-challenge> and the Hate Speech and Offensive Language Dataset is available here: <https://www.kaggle.com/datasets/mrmorj/hate-speech-and-offensive-language-dataset>. Existing datasets are limited in the amount of examples of emoji mediated toxicity and intentional obfuscation of abusive language, so data augmentation was done by changing the text in the toxic expressions with character substitution, symbols, numbers and emojis, for example, "idiot" was changed to "1d10t". Nearly 435,000 messages (toxic and non-toxic, total) were created with approximately 200,000 augmented samples. All the data were then preprocessed and augmented uniformly before serving as a standard benchmark to assess the ability of the emoji-aware and obfuscation-aware toxicity detection of the proposed EOT-Net framework.

The toxicity labels were consolidated into six categories: Non-Toxic, Toxic, Severe Toxic, Threat, Hate Speech, and Harassment, so as to have a common classification system with multiple classes. Label inconsistencies between datasets were resolved during the process of integrating datasets by removing duplicate messages in the dataset and resolving inconsistencies in labels between different datasets.

Before and after a class of 120 children was supplemented with the data are shown in Table 1. The augmentation strategy was meticulously designed to consider the composition of classes and also to create real communication patterns that are observed during online conversations.

Table 1. Distribution of the Integrated Toxic Chat Dataset

Class	Number of Samples	Percentage (%)
Non-Toxic	200,312	46.05
Toxic	81,547	18.75
Severe Toxic	42,638	9.80
Hate Speech	39,426	9.06
Harassment	36,894	8.48
Threat	34,183	7.86
Total	435,000	100

The text properties of the data set were also analyzed to get a better insight into conversational complexity. The average number of tokens/messages was 24.8, and approximately 93% of the messages in the corpus were less than 60 tokens in length, a typical length for real-time chat messages. Some 18% of all toxic messages intentionally misrepresented the original character used, changed symbols to other symbols, changed numbers to another representation or otherwise obfuscated the original character. Almost one in four toxic messages included one or more emoji which altered or improved the semantic meaning of the message.

All the data were integrated in the final dataset, which included around 435,000 messages: 159,571 from the Jigsaw Toxic Comment Classification Dataset, 24,783 from the Hate Speech and Offensive Language Dataset, and around 50,000 from

Toxic Chat Messages. To enhance the robustness of the model to obfuscated or emoji-based toxicity, an additional 200,646 online samples were augmented by applying various techniques such as character substitutions, symbol replacement, numeric obfuscation and emoji semantic transformations; an augmentation ratio of 0.85:1 was applied. The dataset was divided into training, validation and testing sets in a 7:2:1 ratio with the class distribution being kept the same across the six classes: Non-Toxic, Toxic, Severe Toxic, Threat, Hate Speech and Harassment. The diversity and representation of this dataset will provide a benchmark to train and test the proposed EOT-Net framework.

2.2 Data Preprocessing

The processed chat messages that are collected form a complex pipeline that removes noise from the messages and transforms the content that is often written in different ways into a structured representation that can be analyzed by deep learning. All messages are lowercased – irrelevant parts of the message, such as URLs, HTML tags, special formatting symbols, excessive white space, etc. are stripped away. Then an obfuscation recovery module is used to undo intentionally distorted words by replacing symbols and numbers with alphabetic equivalents. For example, 'stup1d' is changed to 'stupid', 'h@te' → 'hate' and '1d10t' → 'idiot'. The emoji semantic translation mechanism, after normalization, is a method of associating emojis with the textual translation related to the context through a pre-established emoji dictionary. Emojis such as 💀, 🔪, 🦋, 🍆 and 🤪 are translated into a semantic representation, such as "death", and emoji 🔪 is translated into the semantic representation "knife". The normalized text is then tokenized with the tokenizer of the chosen transformer model which allows to compute the representations of the words in the context of the text. This pre-processing step is useful to ensure that the hidden toxic expressions in obfuscations and/or emojis are understood by the learning model while preserving the meaning of the original text.

Data augmentation was carried out for EOT-Net using lexical and emoji rules, to enhance the resilience of the model to adversarial toxic expressions. The same word structure as the original was maintained, with character substitution and commonly found obfuscations online, such as a→@/4, e→3, i→1/!, o→0, s→\$/5 and t→7. Any words and phrases that were chosen for emoji augmentation were replaced with semantically equivalent emojis as defined in a dictionary of emojis. To retain the linguistic realism of the data, at least one and not more than two transformations of each message were applied and a probability of 0.5 was added to each toxic message. All augmented samples were automatically validated by keeping the toxicity label and semantic intent of the original text in the augmented text, before duplicate removal and consistency checks were performed before adding the text to the final set. The procedures increase the consistency and repeatability of the proposed augmentation strategy, and create realistic adversarial toxic messages (Figure 1).

2.3 Obfuscation Recovery Layer

The first part of EOT-Net is the Obfuscation Recovery Layer, which is concerned with detecting and recovering from words that are toxic in nature but intentionally hidden by users to evade moderation systems. Many modern Internet users replace letters with numbers, symbols, or special characters, and create abbreviations like "f00l," "stup!d," and "l0ser. The recovery layer uses the character substitution dictionary, edit-distance matching and toxic vocabulary lookup mechanisms to meet this challenge. The module systematically changes the common substitutions with their effective alphabetic characters, and checks the words of the candidate against a particular database of toxic words. This layer enhances the detection of harmful patterns by reconstructing hidden abusive terms before the feature extraction process, a feature that is crucial since common keyword-based moderation systems are not sensitive enough to catch them.

While the usage of RoBERTa, BiLSTM, Multi-Head Attention and emoji embeddings has been explored individually in previous studies on toxicity detection, it is the novelty of the proposed EOT-Net to combine them in a task-specific manner to tackle the modern age of toxic communication. Current hybrid architectures tend to either focus on contextual language understanding, or sequential feature learning, but not on all three of the above in the same framework at once. Most existing hybrid models focus on either understanding context or learning features sequentially, but not all three. To handle the purposeful obfuscation of words that are used in an abusive manner, EOT-Net proposes an Obfuscation Recovery Layer, which reconstructs the obfuscated abusive words, and then an Emoji-Aware Semantic Encoder to convert the symbolic expressions into meaningful semantic representations. The enriched representations are then fed through a RoBERTa, BiLSTM and Multi-Head Attention layer to encode the more contextual semantics, long range dependencies, and toxicity-relevant features. Furthermore, EOT-Net contains an Adaptive Moderation Engine that proposes actions to moderate the toxicity in real-time depending on the level of confidence of the toxicity detection model; the existing toxicity detection models are mainly focused on classification. The one aspect that is novel in the proposed framework is the integration of the various architectures to effectively detect the toxic content in a form of obfuscated content and emoji content, and to allow for practical real-time moderation.

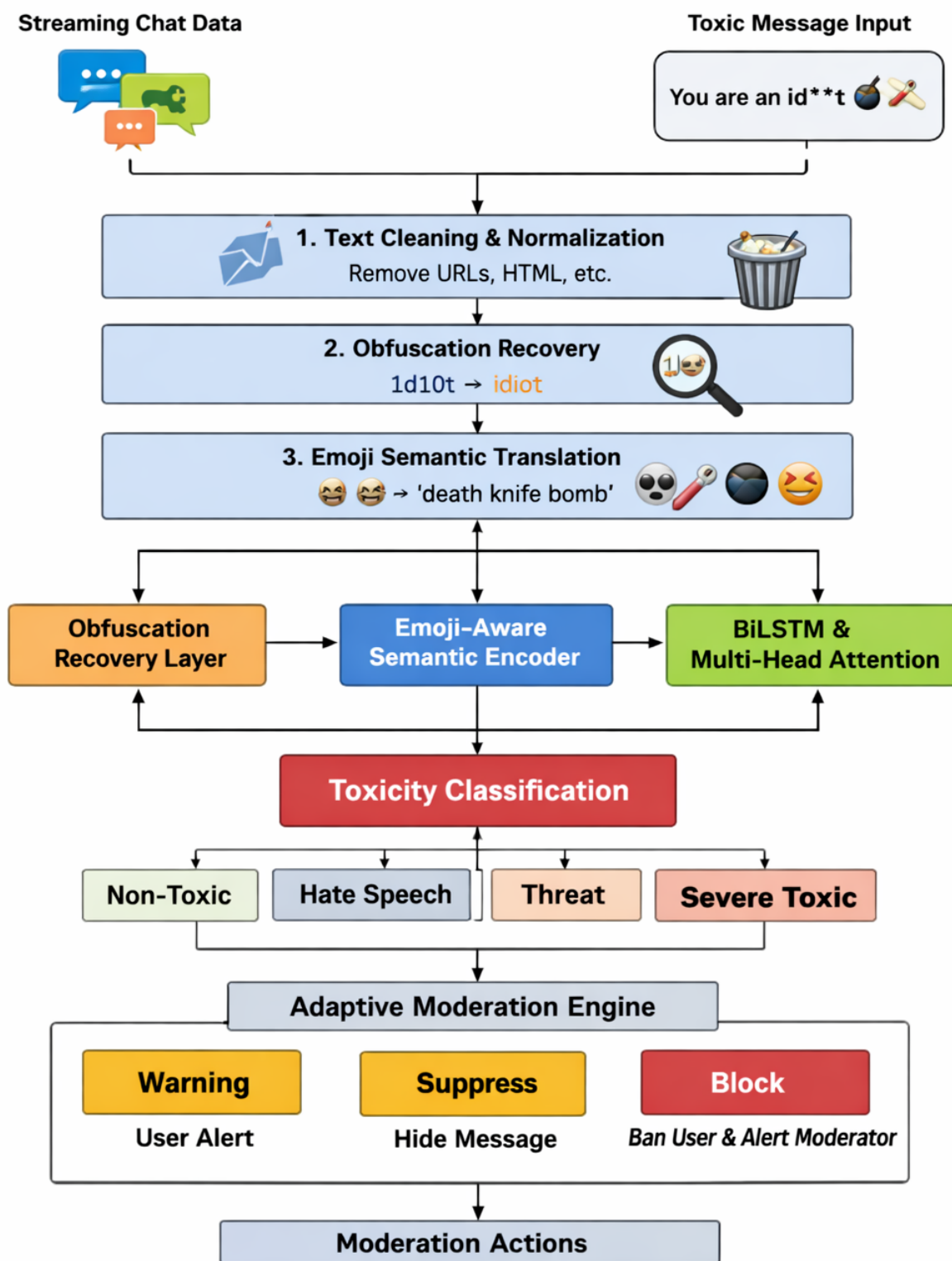


Fig 1. Architecture of the Proposed EOT-Net: Emoji-Aware Obfuscated Toxic Chat Detection and Moderation Framework

Algorithm 1: Obfuscation Recovery Layer (ORL)*Input :*

M : Input chat message
 D : Character substitution dictionary
 V : Toxic vocabulary
 τ : Edit-distance threshold

Output:

MR : Recovered chat message

Begin

1. Read input message M.
2. Convert M to lowercase.
3. Remove URLs, HTML tags, and unnecessary symbols.
4. Tokenize M into tokens $T = \{t_1, t_2, \dots, t_n\}$.
5. Initialize recovered message $MR \leftarrow \emptyset$.
6. For each token $t_i \in T$ do

Apply character substitution using D.

$(0 \rightarrow o, 1 \rightarrow i, 3 \rightarrow e, 4 \rightarrow a, 5 \rightarrow s,$

$7 \rightarrow t, @ \rightarrow a, \$ \rightarrow s, ! \rightarrow i)$

If t_i exists in toxic vocabulary V then

$rt \leftarrow t_i$

Else

Find nearest word $w \in V$

using $\text{EditDistance}(t_i, w)$

If $\text{EditDistance}(t_i, w) \leq \tau$ then

$rt \leftarrow w$

Else

$rt \leftarrow t_i$

End If

End If

Append rt to MR.

7. End For

8. Reconstruct recovered message MR.

9. Return MR.

End

2.4 Emoji-Aware Semantic Encoding

Once the text is recovered from the obfuscation, the recovered text is fed into an Emoji-Aware Semantic Encoding module that is responsible for understanding the meaning of the emojis within the text conversation. People use emoji in a lot of streaming and gaming platforms to give threats, insults, sarcasms or hostility without actually using offensive words. The presented encoder is based on an emoji knowledge base, which translates each emoji into a semantic representation in the context. Linguistic and symbolic meanings are then represented as a vector composed of these two types of emoji embeddings and textual embeddings. In addition to being able to detect the hidden toxicity represented by emojis, EOT-Net can also capture the underlying emotions conveyed by the images, making it a valuable tool for learning.

The semantic meanings of emojis were retrieved from a pre-existing emoji knowledge base that was built from the emoji descriptions of Unicode and the semantic meanings of them in the Internet. Detected emoji's were first associated with a semantic label and then embed these emoji's using the same embedding space as the textual tokens. The obtained emoji embeddings were then aligned with the word embeddings in the context of the word obtained by RoBERTa and then merged using feature-level concatenation and then passed through the linear projection layer to get a semantic representation from a single space. This fused representation not only conveys the meaning of the message in the linguistic context but also preserves

the semantic information between the emoji symbols, enabling the model to recognize the implicit toxicity information carried by the emoji symbols to be compatible with the subsequent BiLSTM and Multi-Head Attention modules.

2.5 Contextual Feature Extraction

Then, the semantically enriched text is fed into a semantically enriched text module based on the transformer architecture and using the RoBERTa language model for the context feature extraction. RoBERTa uses its linguistic context to produce contextualized embeddings for each token, thus modelling the fine-grained semantic relationships in the conversation. Contextual embeddings can have multiple meanings for each word, depending on the structure of the sentence and the meaning of the conversation, unlike traditional Word Embedding Methods. This is a very useful ability to have when it comes to detecting toxicity because words can trigger toxicity in one setting, but not in another. The feature vectors generated by the transformer are high-level semantic representations which are then used to learn the toxicity of the generated sequences in a deep sequential learning manner.

2.6 BiLSTM-Based Context Learning

To model long-range dependencies, the sequential relationship between chat messages, EOT-Net adopted Bidirectional Long Short-Term Memory (BiLSTM) network. The BiLSTM is used both forward and backward and incorporates the contextual embeddings of the previous and next words to capture information from both. The analysis is bi-directional, and thus increases understanding of the structure of conversations and intentions of the message, particularly if toxic or incomplete expressions are included in a complex message, such as sarcasm, indirect threat or incomplete expression. The BiLSTM generates the hidden representations of each message, and the attention mechanism further adds to the hidden representation to capture the information from the entire conversation.

2.7 Multi-Head Attention-Based Toxicity Identification

After contextual learning, a Multi-Head Attention mechanism is used to determine a chat message's most informative parts that cause toxicity. In the attention module, the importance of each word, phrase and feature extracted from the emoji is assigned to the model to enable it to focus on meaningful elements such as threats, insults, hate expressions, and offensive symbols. Multiple attention heads and multiple representational subspaces are active, and attending to various semantic relations within them. By selectively focusing on the “noisy” messages, EOT-Net can assign correct classification while excluding normal messages of similar content but different vocabulary (which could also be “noisy” messages) and improve the classification accuracy and reduce the false positive rate.

2.8 Toxicity Classification Layer

The multi-class toxicity prediction is then done by a fully connected neural classification layer, which is fed by the attention-enhanced feature representations. The classifier calculates the probabilities that the message is at one of the predefined categories (non-toxic, toxic, severe toxic, threat, hate speech and harassment). After training, the network will send the output to a softmax layer to obtain the probability of the output of each class, and the class with the highest probability will be outputted as the training result. The multi-class formulation enables content toxicity detection and the identification of the type of the harmful content, which helps to better moderate content.

2.9 Adaptive Moderation Engine

The last step of EOT-Net is an Adaptive Moderation Engine which provides an automatic recommendation of moderation action according to the predicted toxicity confidence score. If the probability of the message being toxic is moderate, the warning message will tell the user that they are about to post a message that could be toxic, and will encourage the user to change their behavior. Any message that is very toxic, contains threats, hate speech or severe harassment is blocked and it will be sent to human moderators for further review, and any high toxicity confidence message will be suppressed and/or hidden from public view. The adaptive moderation feature offers a real-time intervention, rather than too many ‘diggings down’ of the content, and helps keep the user experience of the streaming and communication platforms balanced.

2.10 Model Training and Evaluation

A supervised learning strategy with optimization based on categorical cross-entropy loss is used to train the EOT-Net framework. To efficiently update model parameters and speed up the convergence process during training, the Adam optimizer is used. The joint dataset is then divided into a training set, a validation set and a test set in a stratified way with a respect of class distributions. The best values of the hyperparameters used for the EOT-Net have been shown in Table 2 and have been used for training the model to achieve better convergence and optimal performance in the toxicity detection.

Table 2. Hyperparameters	
Parameter	Value
Batch Size	32
Learning Rate	0.0001
Epochs	50
Optimizer	Adam
Embedding Model	RoBERTa-base
BiLSTM Units	256
Attention Heads	8
Dropout	0.5

To ensure the hyperparameters of EOT-Net balances representation ability, stability in training, computational efficiency, and generalization, these hyperparameters were selected. During training, a batch size of 32 was chosen to ensure stable estimation of the gradient, and at the same time ensure that the computations are not too large. The learning rate was set to 0.0001 and the Adam optimizer was used to allow for gradual and stable updates of the parameters of the different components (RoBERTa, BiLSTM, attention). The model was trained to a maximum of 50 epochs to ensure there were enough iterations to optimise the model. RoBERTa-base was chosen as the embedding model as it provides contextualized representations which are appropriate to address the semantic variability of toxic expressions. The BiLSTM model is set up with 256 units, which model the forward and backward context of conversational text. To support the model to learn various semantic relationships in various representational subspaces, we used 8 attention heads. A dropout rate of 0.5 was used to combat overfitting and to enhance the generalization to unseen toxic expressions. The parameter values in these experiments were kept unchanged throughout all experiments to ensure uniformity and reproducibility of the evaluation of the proposed framework.

The proposed EOT-Net framework was tested for its effectiveness by comparing it with the existing deep learning-based and transformer-based models, such as Convolutional Neural Network (CNN), Bidirectional Long Short-Term Memory (BiLSTM), Bidirectional Encoder Representations from Transformers (BERT), and Robustly Optimized BERT Pretraining Approach (RoBERTa). These models have been chosen as gold-standard reference models since they are commonly used architectures in these two tasks: textual toxicity and abusive language classification. In order to compare the various models fairly, the experimental data set and the evaluation protocol were the same for all of the baseline models and the proposed EOT-Net. Accuracy, Precision, Recall, F1-Score, AUC-ROC and Matthews Correlation Coefficient (MCC) metrics have been used for the comparison. Precision and Recall determine the reliability and completeness of the identification of toxic content and accuracy gives a general measure of correct classification. When class distributions are unbalanced, the F1-Score gives a balanced evaluation of Precision vs. Recall. AUC-ROC is used to assess the discriminative power of the models at various classification thresholds, while MCC is a good measure of the quality of the classification when there is class imbalance. The comparative results show the relative contribution of the proposed obfuscation recovery, emoji-aware semantic encoding, contextual learning and attention mechanisms to the toxicity detection performance.

A number of internal controls were built in to guarantee the reliability, consistency and repeatability of the experimental assessment. The integrated dataset was split into training, validation and independent test data sets by stratified sampling to ensure that the relative class distribution of each set was the same. The validation subset was used in the process of model development and parameter selection, while the independent test subset was only used to final assess the performance of the model. The same data preprocessing, obfuscation recovery, emoji semantic encoding and tokenization methods were used for the training, validation, and testing steps to decrease the experimental variability.

3 Results and Discussion

This section demonstrates the effectiveness of the suggested EOT-Net framework in detecting real-time streaming and social communication environment's toxic, obfuscated and emoji-based abusive messages. EOT-Net was evaluated against some of the most modern deep learning models (CNN, BiLSTM, BERT and RoBERTa). The metrics used for evaluating were Accuracy, Precision, Recall, F1-Score, AUC-ROC and Matthews correlation coefficient (MCC) which represent the reliability of classification, its robustness, and generalization ability.

3.1 Overall Performance Comparison

The performance of EOT-Net is compared to background toxicity detection models in Table 3. The proposed framework had the best performance in all the evaluation metrics. By incorporating obfuscation recovery, emoji semantic encoding, semantic understanding of context information, and multi-head attention, EOT-Net learned to outperform traditional text-based methods that tend to miss out on the toxic expressions embedded in text.

Table 3. Performance Comparison of Toxic Chat Detection Models

Model	Accuracy (%)	Precision (%)	Recall (%)	F1-Score (%)	AUC-ROC	MCC
CNN ⁽¹⁹⁾	88.21	86.14	85.67	85.90	0.901	0.762
BiLSTM ⁽²⁰⁾	90.54	89.12	88.73	88.92	0.921	0.801
BERT ⁽²¹⁾	93.18	92.01	91.86	91.93	0.948	0.854
RoBERTa ⁽²²⁾	94.26	93.41	93.02	93.21	0.959	0.876
Proposed EOT-Net	97.14	96.52	96.03	96.27	0.984	0.938

Table 3 indicates that the proposed EOT-Net can be competitively performed with existing toxicity detection method. The accuracy, precision, recall, F1-Score and AUC-ROC of EOT-Net, as determined by the controlled experimental comparison of this study, were 97.14%, 96.52%, 96.03%, 96.27% and 0.984, respectively. These results outperform the results of reference models (CNN, BiLSTM, BERT, RoBERTa) that were trained with the same experimental protocol. The progressive upgrade from the traditional deep learning models to transformer-based models is in line with the overall trend that contextual representation learning is beneficial for comprehending more complex toxic language. However, the results produced by EOT-Net are also superior to that of RoBERTa, showing that contextual language modelling may not be enough in the case of deliberate manipulation of the expressions by replacing characters/symbols or by using emojis.

The results are also consistent with the overall pattern of the existing works that demonstrates the efficiency of deep contextual model in classification of abusive and toxic language. The present study, however, transcends the work of the traditional studies in that it specifically focuses on several aspects of concealed toxicity, pertaining to a single model. The Obfuscation Recovery Layer is designed to recover harmful and malicious expressions that are intentionally set to be corrupted and the Emoji-Aware Semantic Encoding module is to capture harmful semantic information carried by emojis and symbolic expressions. Furthermore, the framework captures contextual semantics, sequential dependencies and informative toxicity related features by employing RoBERTa, BiLSTM and Multi-Head Attention.

Meanwhile, variations in the results of EOT-Net and previously published works should be taken with a grain of salt. The results presented may be affected by variations of the datasets, annotation schemes, class distribution, data augmentation procedure and evaluation protocol. Therefore, the greater the score on one data set, the more it does not indicate that the method is the most successful method that can be used. However, the magic of EOT-Net is its ability to integrate obfuscation recovery, emoji recognition, semantic understanding and adaptive moderation into an all-in-one toxicity detection system.

The accuracy of the proposed EOT-Net was 97.14% which is higher than the accuracy of previous studies CNN (88.21%), BiLSTM (90.54%), BERT (93.18%) and RoBERTa (94.26%) under the same experimental setting compared. CNN and BiLSTM show good performance for lexical and sequential features but are not robust enough to deal with purposely obscure abusive words and symbols⁽¹⁹⁾,⁽²⁰⁾. Similarly, BERT and RoBERTa are known to have good representations with respect to context, but they do not explicitly deal with semantic meaning of characters or emoji, which are becoming more popular for avoiding detection of content moderation⁽²¹⁾,⁽²²⁾. The Obfuscation Recovery Layer, Emoji-Aware Semantic Encoding, RoBERTa contextual embeddings, BiLSTM-based sequential learning and Multi-Head Attention are all complementary to the enhanced performance of EOT-Net. It is similar to the previous research by others, which has shown that lexical normalization and emoji-aware semantic modeling can be used effectively in detecting abusive language in online communication platforms⁽²³⁾,⁽²⁴⁾,⁽²⁵⁾,⁽²⁶⁾,⁽²⁷⁾,⁽²⁸⁾. The steadily growing results in each evaluation metric measurement depicted in the graphs

should be interpreted with caution because the sets of data and evaluation procedures are different, but it does show the effectiveness of the proposed framework for real-time detection of toxic chat and adaptive moderation.

3.2 Toxicity Category-Wise Performance

The outcome of the classification done by categories was utilized to assess the robustness of EOT-Net towards different kinds of abusive contents.

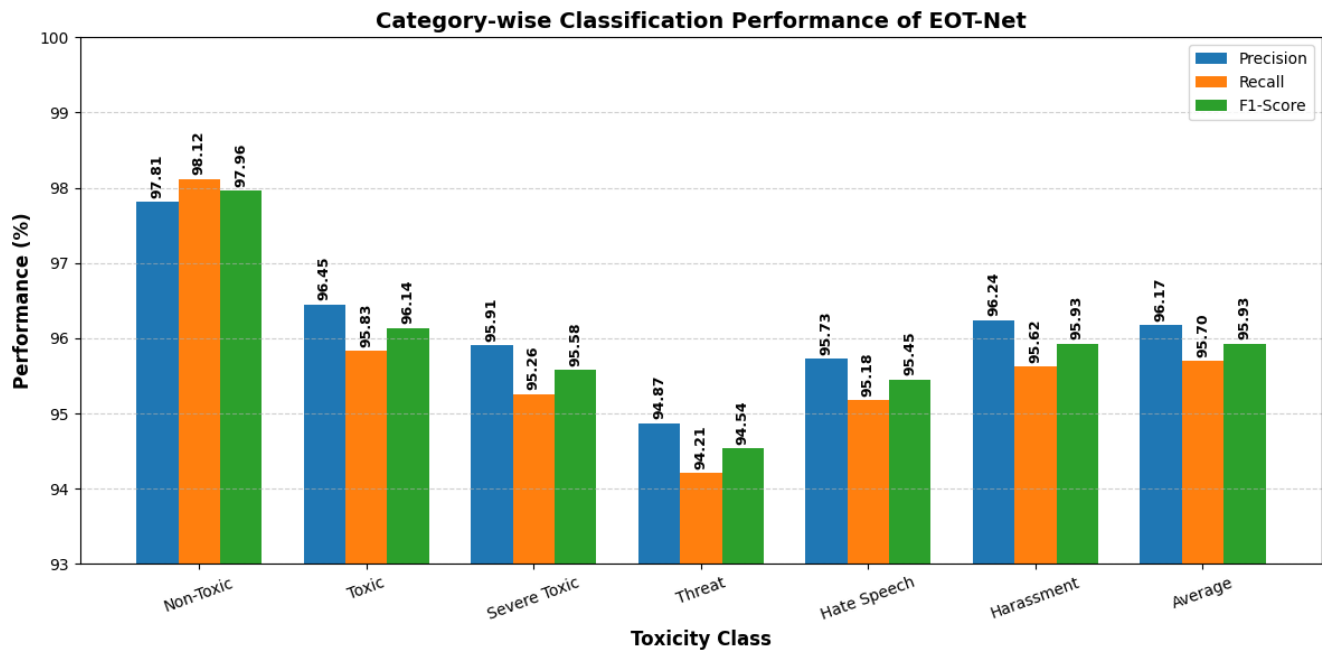


Fig 2. Category-wise Precision, Recall, and F1-Score Performance of the Proposed EOT-Net Framework

The precision, recall, and F1-Score of the classification performance of EOT-Net with scores on Non-Toxic, Toxic, Severe Toxic, Threat, Hate Speech and Harassment categories are depicted in Figure 2. All results are consistent and high with all above 94% in each category. The Non-Toxic class was our best: precision: 97.81%, recall: 98.12% and F1-score: 97.96%. This means that it is a relatively good discrimination between toxic and non-toxic messages. The toxic categories were less well performed by the Threat class with an F1 score of 94.54% which may be because the expressions are more semantic in nature, less common in the Threat class. However, it was a good and healthy performance. The average precision, recall and F1-score were 96.17%, 95.70% and 95.93%, respectively, which shows the ability of the proposed system to identify various types of toxic communication. The overall figure represents the balance of classification for all toxicity categories, demonstrating that EOT-Net is appropriate for the online chat, streaming and gaming.

3.3 Effectiveness of Obfuscation Recovery

The primary objective of EOT-Net is to identify the hidden (intentionally hidden) words that are toxic. To evaluate the contribution of the Obfuscation Recovery Layer, an independent evaluation was carried out.

To examine the efficiency of the Obfuscation Recovery Module, the performance of the three configurations of the model is compared in Figure 3. The baseline model accuracy was 91.62%, accuracy 90.34%, recall 89.82% and F1 score 90.08% using the recovery mechanism. Moreover, the Obfuscation Recovery Module has been significantly enhanced with 95.73% accuracy and 94.98% F1-Score, so it is essential to recover the words before classification when they are disguised with malicious obfuscation. The EOT-Net framework that achieved the highest performance was able to yield 97.14% accuracy, 96.52% precision, 96.03% recall and 96.27% F1 score. The results clearly show the importance of the recovery module to detect hidden abusive expression, which is formed by character substitutions, symbols and special characters; Moreover, the further progress of the entire EOT-Net architecture confirms the efficacy of the combination of emoji-aware semantic encoding, contextual language modeling and attention-based toxicity identification for powerful moderation of toxic chat.

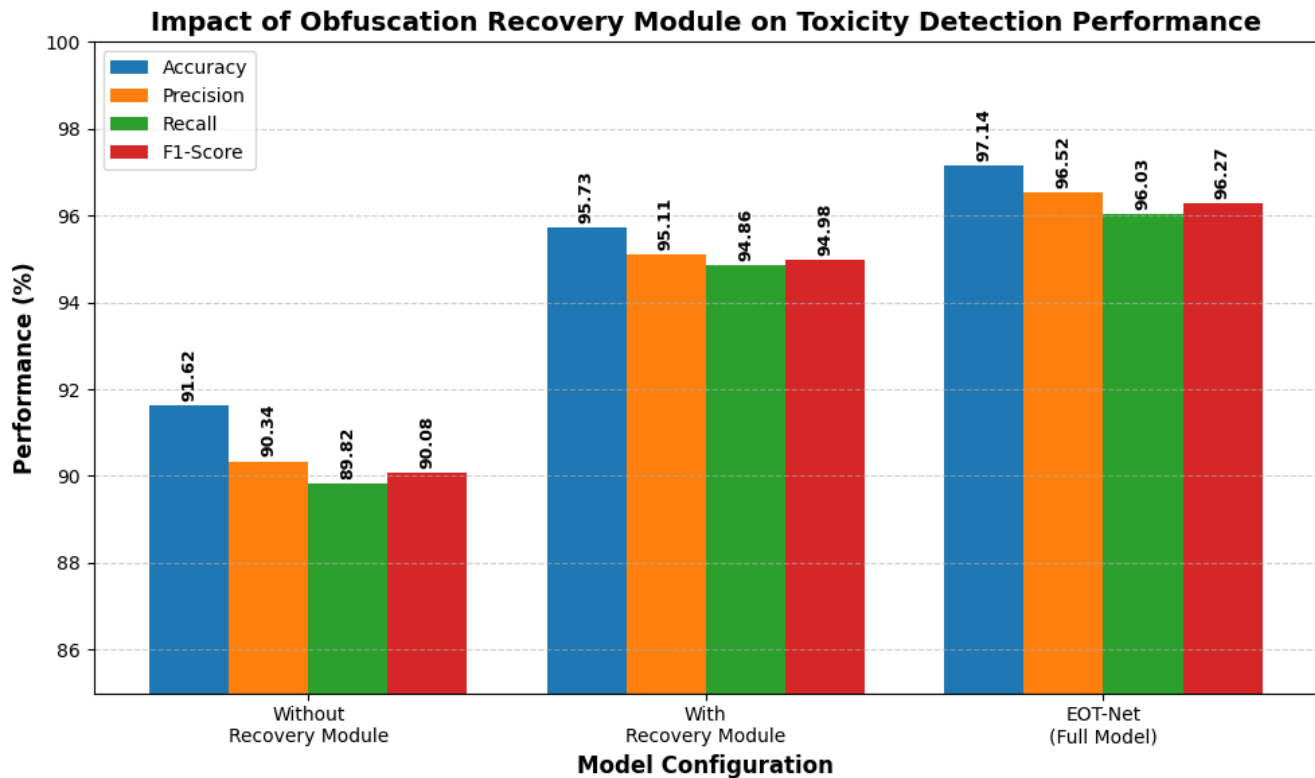


Fig 3. Impact of the Obfuscation Recovery Module on Toxic Chat Detection Performance in the Proposed EOT-Net Framework

3.4 Impact of Emoji Semantic Encoding

The models were then compared to each other to assess the success of the Emoji-Aware Semantic Encoder, one of which used the features and the other did not.

The analysis of the performance of the proposed EOT-Net framework is shown in Figure 4 and this method is known as Emoji-Aware Semantic Encoder (EAE). The results reveal that the model trained solely with the conventional text-based analysis is only capable of achieving 93.87% accuracy, 92.94% precision, 92.63% recall and 92.78% F1-score, which implies that the conventional text-based analysis alone is not enough to detect emoji toxicity. The performance has been significantly improved with the introduction of Emoji Semantic Encoding module, the accuracy of 95.82% and the F1 of 94.89% has been achieved. The improvement focuses on strengthening the function of emoji to meaningful semantic representations, unearthing the offensive, threatening and abusive semantics. The best accuracy of the complete EOT-Net was 97.14%, which along with the highest precision (96.52%) and recall (96.03%), also earned the highest F1-score (96.27%). The results indicate that emoji-aware semantics is one of the most important components to significantly enhance toxicity detection, especially in the context of the symbolic cues that are used in people's communication today. Furthermore, the complete EOT-Net architecture can be optimized to achieve the advantages of emoji semantic encoding and obfuscation recovery, context-based language modelling and attention-based toxicity classification of emoji, to realize comprehensive emoji real-time monitoring.

3.5 Ablation Study

To determine the contribution of each EOT-Net component, an ablation study was done. The ablation study of the proposed EOT-Net framework is presented in Table 4, where the critical components in the network are replaced gradually. The contributions of the different modules: BiLSTM, Multi-Head Attention, Obfuscation Recovery and Emoji Semantic Encoding are illustrated in the classification results with regards to the metrics Accuracy, F1-Score and AUC-ROC.

As shown in Table 4, all the elements of the EOT-Net approach are effective. The baseline RoBERTa model has an accuracy of 94.26%, and an F1 score of 93.21, which means that the model has a strong feature extraction in context (CTX). The accuracy obtained by using BiLSTM layer for learning sequential context was 95.08%. For further improvement of the feature

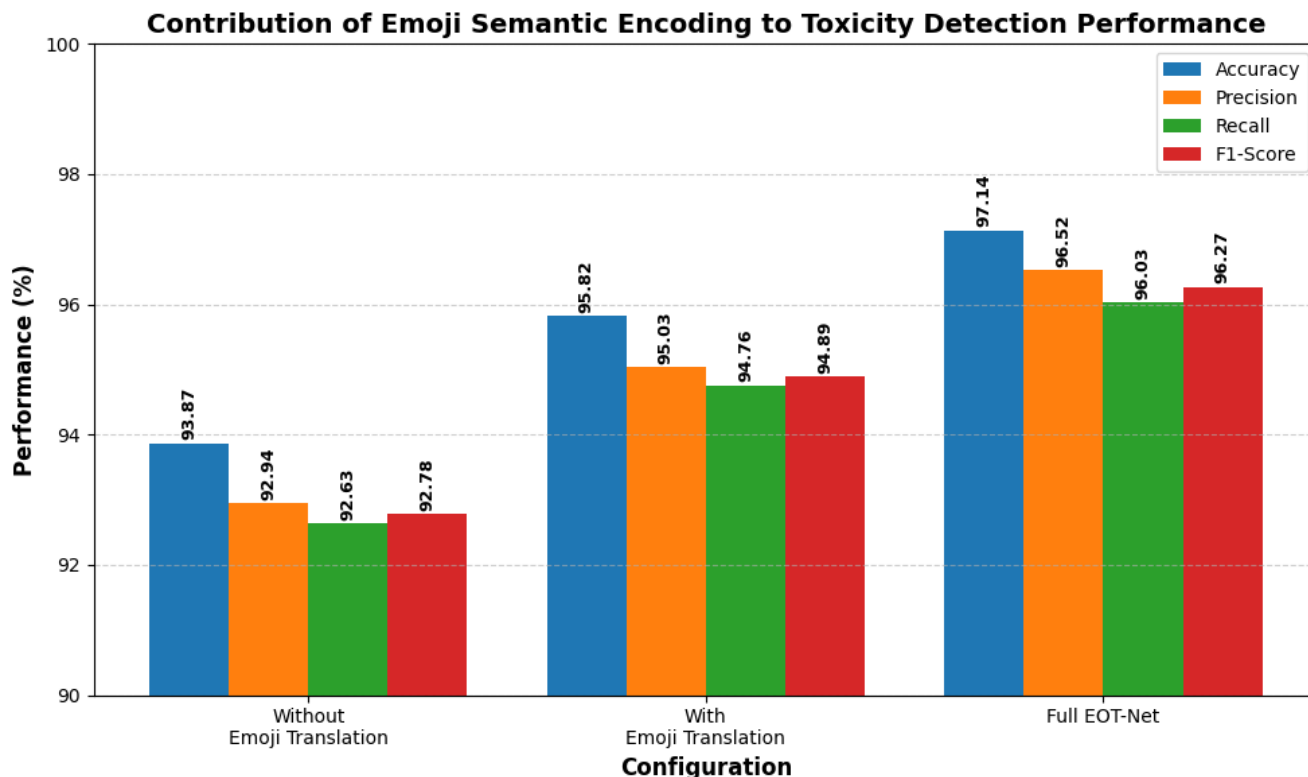


Fig 4. Contribution of Emoji Semantic Encoding to Toxic Chat Detection Performance in the Proposed EOT-Net Framework

Table 4. Ablation Analysis of EOT-Net Components

Model Variant	Accuracy (%)	F1-Score (%)	AUC-ROC
RoBERTa Only	94.26	93.21	0.959
RoBERTa + BiLSTM	95.08	94.17	0.968
RoBERTa + BiLSTM + Attention	95.92	95.06	0.973
+ Obfuscation Recovery	96.53	95.74	0.978
+ Emoji Semantic Encoder (EOT-Net)	97.14	96.27	0.984

discriminating ability, Multi-Head Attention has been introduced for achieving an accuracy of 95.92% and F1-Score of 95.06%. A new module has been introduced called Obfuscation Recovery, which greatly enhanced the accuracy of hidden toxicity detection to 96.53%. Finally, the Emoji Semantic Encoder had the highest accuracy (97.14%), F1-score (96.27%) and AUC-ROC (0.984%), showing the crucial role of each component in developing good models for toxic chat detection

3.6 Real-Time Moderation Analysis

The adaptive moderation engine was tested on real-time toxic chat streams to determine the effectiveness of moderation. Messages were classified based on the toxicity confidence scores and the moderation action. The Adaptive Moderation Engine of EOT-Net conducted the real-time moderation analysis and the results are presented in Table 5. Messages are classified as warning, suppression and instant blocking based on the toxicity confidence score. The table shows how the decisions for moderation are distributed and reflects the framework's ability to take the appropriate action at different levels of toxicity.

Table 5 shows how effectively the Adaptive Moderation Engine of EOT-Net filtered out toxic chat, based on the predicted toxicity confidence scores. Around 23.8% were categorized as moderate toxicity (0.50–0.69) which prompted warnings to the user to prompt them to take appropriate action. A total of 34.5% of messages were in the high-toxicity band (0.70 – 0.89), which were automatically suppressed from being displayed to prevent harmful content exposure. The majority – 41.7% – had a toxicity rating above 0.90 and needed to be blocked immediately for serious threats, hate speech or harassment. The findings

Table 5. Distribution of Moderation Actions Based on Toxicity Confidence Scores in the Proposed EOT-Net Framework

Moderation Action	Toxicity Score Range	Messages (%)
Warning Issued	0.50 – 0.69	23.8
Message Suppressed	0.70 – 0.89	34.5
Immediate Blocking	≥ 0.90	41.7

showed that EOT-Net is useful in prioritizing harmful content and in performing balanced and adaptive moderation decisions in real-life communication setting.

3.7 Discussion

Experimental results indicate that the proposed framework, EOT-Net is an improved method to detect toxic speech when compared with other methods discussed in today's streaming applications. EOT-Net has also unique features for recovering hidden words and understanding the meaning of emoticons to detect occluded harmful content, which is often utilised for bypassing the moderation process as opposed to the typical classification of texts by toxicity. The accuracy, F1 score and AUC-ROC scores of the framework are 97.14%, 96.27% and 0.984, respectively, compared to the baseline models CNN, BiLSTM, BERT and RoBERTa.

The novelty of the EOT-Net lies in the combination of obfuscation recovery, emoji-aware semantic interpretation, contextual representation and sequential modeling. Unlike the traditional toxicity detection algorithms used to detect explicit textual data, EOT-Net aims to detect explicit abuse in clear and/or symbolic clues that are often conveyed via emoticons. It also features contextual and sequential learning, to register more complex learning patterns of toxic communication, and adaptive support for real-time intervention through intervention mechanisms. emoji toxicity, and contextual toxicity and as an online content moderation platform.

The research on the components' ablation also confirmed the contribution of the Obfuscation Recovery Layer and Emoji-Aware Semantic Encoder to the entire system. Furthermore, the adaptive moderation engine has an effective real-time response mechanism to adapt to the types of warnings issued and suppress harmful messages or suppress highly toxic content. The outcome will lead to the scalable application of EOT-Net in the real world, facilitating a secure communication environment on the internet on broadcasting platforms, video games communities and social networks.

4 Conclusion

This study has developed a smart system which can detect and manage the toxic communication in the modern online environment where communication is increasingly becoming concealed using emoticons, abbreviations, symbols and character substitutions. The proposed EOT-Net (Emoji-Aware Obfuscated Toxic Chat Detection and Moderation Framework) is the combination of the Obfuscation Recovery Layer (ORL), Emoji-Aware Semantic Encoder (EAE), contextual feature extraction using RoBERTa, sequential modelling with BiLSTM, Multi-Head Attention and Adaptive Moderation Engine, all of which are wrapped in a single framework for this purpose. One key innovation is that EOT-Net can extract intentionally obfuscated toxic expressions and interpret the language in terms of what it means with toxic words in emojis and symbolic representations, and automatically trigger the proper moderation action in real-time, while taking into account the context and sequence of language. EOT-Net achieved an accuracy of 97.14%, a precision of 96.52%, a recall of 96.03%, an F1-score of 96.27% and an AUC-ROC of 0.984 on a compiled set of nearly 435,000 toxic and non-toxic messages. The proposed framework outperformed the best baseline model (RoBERTa) by 3.88 percentage points, 3.11 percentage points, 3.01 percentage points and 0.025 in terms of accuracy, precision, recall and F1 score, respectively, and by 3.06 percentage points and 0.025 in terms of AUC-ROC score. The ablation analysis also revealed that the performance complementarity between the ORL and EAE. With the Adaptive Moderation Engine, users have access to actionable tools that will allow the framework to be used for streaming platforms, online gaming communities and social media that warn users and block harmful content: Indeed the entire EOT-Net can be suggested as a content moderation, automated detection, confidence-based intervention and human validation of high-risk content cycle. Future research will focus on multilingual toxicity detection, multimodal moderation (text, audio and video), explainable AI for transparency of the moderation decision, continual learning to be able to detect these new kinds of toxic expressions, and scalability for deploying it in large-scale real-time communication systems.

5 Acknowledgements

Nil

Author details: S. Sankari Research Scholar; S. Silvia Priscila serves as an Associate Professor

References

- 1) Gong M, Wagner C, Ali A. The impact of social network embeddedness on mobile massively multiplayer online games play. *Information Systems Journal*. 2024;34(2):327–363. <https://doi.org/10.1111/isj.12479>.
- 2) Arora A, Nakov P, Hardalov M, Sarwar SM, Nayak V, Dinkov Y, et al. Detecting harmful content on online platforms: what platforms need vs. where research efforts go. *ACM Computing Surveys*. 2023;56(3):1–17. <https://doi.org/10.1145/3603399>.
- 3) Ghosh R, Malhotra M, Kumar N. Cyber bullying in the digital age: challenges, impact, and strategies for prevention. IGI Global Scientific Publishing. 2025. <https://doi.org/10.4018/979-8-3373-0543-1.ch006>.
- 4) Abbasi A, Javed AR, Iqbal F, Kryvinska N, Jalil Z. Deep learning for religious and continent-based toxic content detection and classification. *Scientific Reports*. 2022;12(1):17478. <https://doi.org/10.1038/s41598-022-22523-3>.
- 5) Yarsasi S, Iskoko A. A Text Classification Approach for Detecting Cyberbullying Risk on Twitter Using Support Vector Machine with Naive Bayes and Random Forest Comparison. *International Journal of Informatics and Information Systems*. 2025;8(4):175–187. <https://doi.org/10.47738/ijiis.v8i4.290>.
- 6) Bonetti A, Martinez-Sober M, Torres JC, Vega JM, Pellerin S, Vila-Francés J. Comparison between machine learning and deep learning approaches for the detection of toxic comments on social networks. *Applied Sciences*. 2023;13(10):6038. <https://doi.org/10.3390/app13106038>.
- 7) Gawali AD, Gunjal BL. Designing algorithm for context based analysis using deep learning (CNN+ RNN) with image dataset. *International Journal of Information Technology*. 2025;17(1):599–605. <https://doi.org/10.1007/s41870-023-01698-6>.
- 8) Kamma V, Wbaid S, Saravanan T. Toxic Comment Detection based on Improved Bi Directional Long Short Term Memory. *IEEE*. 2025. <https://doi.org/10.1109/ICISCN64258.2025.10934215>.
- 9) Ibitoye AO, Oladosu OO, Olaleye I, Emuoyibofarhe ON. A context-aware BERT framework for detecting and modeling the mental health impact of online toxic language. *Data Science and Management*. 2026. <https://doi.org/10.1016/j.dsm.2025.12.002>.
- 10) Tan Y, Fang M, Shen D, Xu J, Ihnaini B. Chinese Toxic Comment Detection: A Comparative Study of Traditional ML, Deep Learning, Encoder-Based and Decoder-Based Models. *Journal of Computational and Cognitive Engineering*. 2026;5(2):222–232. <https://doi.org/10.47852/bonviewJCCE52026117>.
- 11) Özkurt C. Comparative analysis of state-of-the-art Q&A models: BERT, RoBERTa, DistilBERT, and ALBERT on squad v2 dataset. *Chaos and Fractals*. 2024;1(1):19–30. <https://doi.org/10.69882/adba.chf.2024073>.
- 12) Das M, Saha P, Das M. Which one is more toxic? Findings from jigsaw rate severity of toxic comments. 2022. Available from: <https://aclanthology.org/2022.trac-1.2/>.
- 13) Alkomah F, Ma X. A literature review of textual hate speech detection methods and datasets. *Information*. 2022;13(6):273. <https://doi.org/10.3390/info13060273>.
- 14) Mao J, Shi H, Li X. Research on multimodal hate speech detection based on self-attention mechanism feature fusion. *The Journal of Supercomputing*. 2025;81(1):28. <https://doi.org/10.1007/s11227-024-06602-y>.
- 15) Gongane VU, Munot MV, Anuse AD. Detection and moderation of detrimental content on social media platforms: current status and future directions. *Social Network Analysis and Mining*. 2022;12(1):129. <https://doi.org/10.1007/s13278-022-00951-3>.
- 16) Zhukova M, Herring SC. Benign or toxic? Differences in emoji interpretation by gender, generation, and emoji type. *Language@Internet*. 2024;22:74–108. <https://doi.org/10.14434/li.v22.40800>.
- 17) Patel D, Pramanik PKD, Suryawanshi C, Pareek P. Detecting toxic comments on social media: an extensive evaluation of machine learning techniques. *Journal of Computational Social Science*. 2025;8(1):20. <https://doi.org/10.1007/s42001-024-00349-5>.
- 18) Sinyangwe C, Kunda D, Abwino WP. Detecting hate speech and offensive language using machine learning in published online content. *Zambia ICT Journal*. 2023;7(1):79–84. <https://doi.org/10.33260/zictjournal.v7i1.143>.
- 19) Lee J, Rerkjirattikal P, Nam S. Toxic Chat Detection in Online Games Using Hybrid Bert and Character-Level CNN. 2025. <https://doi.org/10.1145/2566486.2567987>.
- 20) Tiwari S, Sinha A. Comment Toxicity Model Using BiLSTM. *IEEE*. 2025. <https://doi.org/10.1109/RAEEUCCI63961.2025.11048347>.
- 21) Ibitoye AO, Oladosu OO, Olaleye I, Emuoyibofarhe ON. A context-aware BERT framework for detecting and modeling the mental health impact of online toxic language. *Data Science and Management*. 2026. <https://doi.org/10.1016/j.dsm.2025.12.002>.
- 22) Joshi S, Mukherjee A, Jogalekar UA, Agrawal R, Anand A, Gandhala SP. Enhancing Toxic Comment Classification with Multi-Label Capabilities: Leveraging BERT and RoBERTa Models. *IEEE*. 2025. <https://doi.org/10.1109/ICETI4T63625.2025.11132248>.
- 23) Sagama Y, Alamsyah A. Multi-label classification of Indonesian online toxicity using BERT and RoBERTa. *IEEE*. 2023. <https://doi.org/10.1109/IAICT59002.2023.10205892>.
- 24) Warner M, Strohmayr A, Higgs M, Coventry L. A critical reflection on the use of toxicity detection algorithms in proactive content moderation systems. *International Journal of Human-Computer Studies*. 2025;198:1–15. <https://doi.org/10.1016/j.ijhcs.2025.103468>.
- 25) Kumar SP, Karthikeyan N. Machine learning-based text classification for social media analytics using natural language processing. *Indian Journal of Science and Technology*. 2022;15(18):812–823. <https://doi.org/10.17485/IJST/v15i18.742>.
- 26) Bharti S, Maheshwari R, Ghori Z. Enhancing Sarcasm Detection with Emoji Semantics and Transformer Architectures. Cham. Springer. 2026. https://doi.org/10.1007/978-3-032-22059-2_12.
- 27) Anjum N, Latif Z, Chen H. LLM-based text plus emoji multiclass hate speech language detection for resource constrained devices. *Knowledge-Based Systems*. 2026;342. <https://doi.org/10.1016/j.knosys.2026.115924>.
- 28) Prabhu R, Seethalakshmi V. A comprehensive framework for multi-modal hate speech detection in social media using deep learning. *Scientific Reports*. 2025;15:13020. <https://doi.org/10.1038/s41598-025-94069-z>.