RESEARCH PAPER

Metaheuristic Feature Selection for BERT Embeddings in Toxic Comment Classification

Divya Singhal[1], Ankit Verma[2], Amit Kumar Gupta[3], Vipin Kumar[4], Shashank Bhardwaj[5], Mahima Tayal[6]

[1]Department of Computer Science, Noida Institute of Engineering & Technology, Greater Noida, UP. Email: divyasinghal021@gmail.com

[2]Department of Computer Applications, Krishna Institute of Engineering & Technology (KIET), Ghaziabad, Delhi-NCR, Uttar Pradesh. Email: ankit.mca4u@gmail.com

[3]Department of Computer Applications, Krishna Institute of Engineering & Technology (KIET), Ghaziabad, Delhi-NCR, Uttar Pradesh. Email: amit.id29@gmail.com

[4]Department of Computer Applications, Krishna Institute of Engineering & Technology (KIET), Ghaziabad, Delhi-NCR, Uttar Pradesh. Email: vipin.kumar.mca@kiet.edu

[5]Department of Computer Applications, Krishna Institute of Engineering & Technology (KIET), Ghaziabad, Delhi-NCR, Uttar Pradesh. Email: shashank12swe@gmail.com

[6]Department of Computer Applications, Krishna Institute of Engineering & Technology (KIET), Ghaziabad, Delhi-NCR, Uttar Pradesh. Email: mahima.tayal@kiet.edu

Received: 25th May, 2026; **Revised:** 6th June, 2026; **Accepted:** 8th June, 2026; **Available Online:** 22nd June, 2026

ABSTRACT

Online platforms generate a large volume of user-generated content, which often includes toxic or hate comments. Automated detection of such instances has become an important problem in natural language processing. Recent trends in this area have shown a high dependence on the use of transformer-based models, which are known to generate high-dimensional features. However, these embeddings may contain redundant features that increase computational complexity. This study proposes an optimization-driven framework for toxic comment classification using BERT embeddings combined with the Whale Optimization Algorithm (WOA) for feature optimization. BERT is first used to generate semantic embeddings for textual comments. WOA is then employed to identify an optimal subset of embedding features that improves model efficiency while preserving classification performance. The reduced feature set is evaluated using multiple machine learning classifiers. Experimental results on the Jigsaw Toxic Comment dataset demonstrate that the proposed approach reduces the feature space by approximately 24% while maintaining comparable classification performance.

Keywords: BERT Transformer, Toxic Comments, Whale Optimization Algorithm, Feature Extraction, Performance.

How to cite this article: Singhal D, Verma A, Gupta AK, Kumar V, Bhardwaj S, Tayal M. Metaheuristic Feature Selection for BERT Embeddings in Toxic Comment Classification. Int J Drug Deliv Technol. 2026;16(62s): 1914-1922. DOI: 10.25258/ijddt.16.62s.192

Source of support: Nil.

Conflict of interest: None

Introduction

The swiftprogress of social media platforms has significantly increased the amount of user-generated textual content available online. While these platforms enable communication and information sharing, they also facilitate the spread of toxic and offensive language[1]. Toxic comments, including abusive or insulting statements, can negatively impact online communities and contribute to harmful digital environments.

The traditional approaches for feature extraction in the context of text classification follow the idea of manually designing the features that transform the input text into mathematical vectors based on the patterns of word occurrences rather than the actual meaning of the input text. The most common Vectorization techniques include TF-IDF, n-gram approaches, and Bag of Words [2]. Although these approaches are simple, intuitive, and computationally efficient, there are some notable shortcomings associated with these approaches. The approaches are word-level features that do not consider the actual meaning of the input text. For example, the word "kill" in the sentence "this song kills it" and the word "kill" in the sentence "I will

kill you" are the same features in the context of the traditional approaches, although the actual meaning of the word is quite different in both cases[3]. Also, these approaches are associated with increased memory requirements. Recent developments in the context of deep learning and the transformer approaches have enhanced the performance of the classifiers in the context of text classification problems. The BERT approach is able to capture the contextual relationships between the words in the input.

Despite their effectiveness, transformer-based embeddings typically produce high-dimensional feature vectors. Not all dimensions in these embeddings contribute equally to classification performance, and redundant features may increase computational complexity. Feature selection techniques provide valuable features and eliminate unnecessary features. It has been proven effective in solving feature selection problems by using meta-heuristic optimization algorithms. WOA is also another optimization algorithm that addresses this issue by efficiently balancing between exploration and exploitation during the search process. WOA is inspired by the bubble-net hunting behavior of humpback whales, which are capable of efficiently searching for optimal solutions in complex landscapes [19].

In this paper, we propose a feature selection framework that utilizes BERT embeddings and WOA to reduce the feature dimension of semantic representation for solving toxic comment classification problems, with the aim of determining if feature selection is capable of efficiently reducing the feature dimension of embeddings for solving classification problems. The main contributions of this work are as follows:

i. WOA-basedfeature selection framework to identify informative dimensions from BERT-based semantic embeddings.
ii. The proposed method significantly reduces the embedding dimensionality while maintaining comparable classification performance across multiple ML classifiers.
iii. A lightweight toxic comment detection pipeline is developed by combining optimized transformer embeddings with classical machine learning models.

II. Related Work

The spread of toxic information from online sources has become a focal point for many researchers in recent years. Automated recognition of toxic and offensive language is the most investigated topics in the field of NLP. Initially, most studies focused on traditional ML

techniques using lexical features. In this regard, it is noted that Logistic Regression (LR) is better than Naïve Bayes (NB), Random Forest (RF), and XGBoost [4]. Although these methods are computationally efficient, they may not be able to identify relationships in the text data. Deep learning methods, such as convolutional networks, Long Short-Term Memory and recurrent networks, have been used for the automated recognition of toxic comments [5].

TF-IDF Vectorization is a popular method of feature extraction used in text classification tasks. In this method, the text data is converted into mathematical vectors based on the frequency of the words [6]. In addition, the feature vectors obtained from the transformer embeddings are usually of a high dimension, which might contain some irrelevant information. Recently, a new class of models called transformers, which includes BERT, has achieved great success in understanding the context of the meaning of the text data [7]. The BERT model has the capacity to understand the situational meaning of the text data based on the preceding and succeeding words in a sentence [8]. In addition, Zhao et al. revealed in their study that the predictive capability of BERT model is better compared to other language models in classifying toxic comments [9]. Bonetti et al. revealed classic ML classifiers when tested with pre-trained BERTweet transformer just obtain 91.40% accuracy at high computational cost [10].

Traditional feature identification methods include filter-based and wrapper-based approaches. However, these techniques may struggle with high-dimensional data. Meta-heuristic optimization algorithms have emerged as effective tools for

solving feature selection problems. Algorithms such as Genetic Algorithms, Particle Swarm Optimization, and Ant Colony Optimization have been applied to select optimal feature subsets. WOA is a relatively recent metaheuristic algorithm due to its simple structure and efficient search capability. However, limited research has explored the application of WOA for selecting informative features from transformer-based embeddings in toxic comment classification. This work aims to address this gap by integrating WOA with BERT-based representations.

III. Research Methodology

Figure 1 shows the architecture of the proposed model, which uses the BERT model to generate semantic embeddings and the WOA algorithm to perform optimization-based feature selection. The selected features are then used by the machine learning classifier to detect toxic comments.

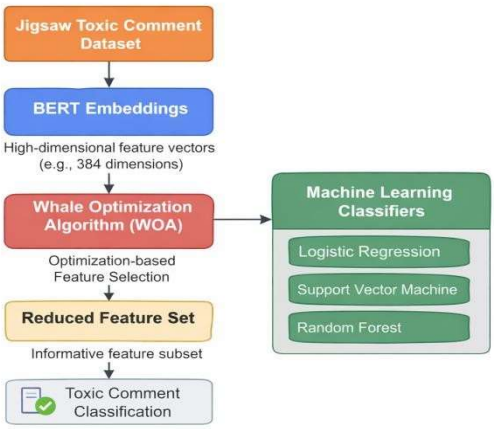

Figure 1 Proposed Workflow Methodology for Toxic Comments Detection

a) Dataset Description

There are many publicly available datasets for the detection of toxic and offensive language, as discussed in Table 1. These datasets were instrumental in the advancement of toxicity detetction research. However, the available datasets are limited in many ways, for example, they are either too small or based only on a single platform or are binary in classification and do not have more categories, or they are limited in the definition of toxicity, for example, only hate speech. Additionally, many multilingual datasets are available, but the focus of the present work is the detection of toxicity in the English language only.

Table 1 Existing Toxic Comment Classification Datasets

Ref	Dataset	Feature	Platform	Category	Records
[13]	Davidson Hate Speech dataset	5	Twitter	Hate Speech	24,783
[14]	Offensive Identification Dataset	5	Twitter	Offensive Language	14,100
[15]	YouTube	15	YouTube	Videos Comments	1000
[16]	Jigsaw Unintended Bias in Toxicity Classification	6	English-language news sites	Civil Comments	5,00,000
[17]	Jigsaw Rate Severity of Toxic Comments	7	Wikipedia	Multiple	1,59,571
[18]	social media toxic comments	8	Multiple platforms	Multiple	50,000

In this study, we selected the Jigsaw Toxic Comment dataset as the training domain because it provides multi-label annotations across different toxicity categories, including toxic, severe toxic, obscene, threat, insult, and identity hate [16]. This fine-grained labeling enables richer supervision and more detailed semantic learning compared to purely binary datasets.

b) BERT-based Feature Extraction

We adopt a transformer-based feature extraction approach rather than relying on traditional handcrafted features such as word vectorization or embedding [1]. Traditional feature engineering methods are frequency-based and highly dependent on vocabulary overlap between training and testing datasets. For obtaining the contextual semantic information contained within the comments, the Sentence BERT model 'all-MiniLM-L6-v2' is used for generating the sentence embeddings for the comments. The Sentence BERT model is a lightweight model with 6 layers in its transformer architecture, and it is used for generating 384-dimensional feature vectors that are compact and contain syntactic structure, semantic relationships, and contextual dependencies. Even though the model is termed a lightweight model, it still comprises around 22 million parameters. This helps to reduce the domain sensitivity and make the representation more robust, allowing the model to generalize better.

c) WOA for Feature Selection

In the transformer model, input texts are converted into a dense representation of meaning with hundreds of dimensions. Despite the richness of the information being provided, not all of the information being represented is of equal importance to the process of classification. This could result in more complex problem due to redundant information being represented in the model. Therefore, the optimal selection of the features becomes a critical process, and the WOA algorithm acts as a meta-heuristic optimization technique for the selection of the features based on the BERT model. It was developed as a population-based optimization technique that can efficiently balance the exploration and exploitation of the search space. The whales represent the solution to the problem of feature selection, which is a set of the selected features. This is normally represented as a binary vector, where a 1 represents the selection of the feature and a 0 represents the non-selection of the feature. The WOA initializes a set of whales with different feature subsets randomly and then iteratively updates the solution using the position of the best-performing whale.

d) Classification models

In this study, classical ML classifiers were employed to assess the effectiveness of the proposed feature selection framework. The primary objective is to investigate the effectiveness of the WOA for feature selection on BERT-based embeddings, rather than developing new classification architecture. The role of the classifier is primarily to evaluate whether the selected feature subset preserves the discriminative information contained within the original embedding space. Also, using lightweight classifiers significantly reduces computational complexity and training time, which is particularly important when conducting optimization-based feature selection experiments that require repeated evaluation of candidate feature subsets during the optimization process[11]. Employing computationally efficient classifiers ensures that the optimization process remains tractable while still providing reliable performance evaluation.

e) Model Configuration

The hyper-parameter configuration used in the experiments is summarized in Table 2. Since the primary goal of this study is to evaluate the effectiveness of the WOA-based feature selection framework rather than to optimize classifier architectures, commonly adopted default parameters were used for the classification models. The optimization process employed 20 whales and 30 iterations, while the fitness evaluation used a Logistic Regression classifier with 3-fold cross-validation.

Table 2 Hyper-parameter Configuration used in Experiments

Component	Parameter	Value
Sentence-BERT	Model	All-MiniLM-L6-v2
	Embedding Dimension	384
	Transformer Layer	6
Whale Optimization Algorithm	Population Size	20
	Iterations	30
	Fitness Function	F1-Score+ Feature Sparsity
	Encoding	Binary Feature Selection
Logistic Regression (Fitness Evaluator)	Solver	L-BFGS
	Cross Validation	3-fold
Logistic Regression (Classifier)	Regularization	L2
Linear SVM	Kernel	Linear

IV. Experimental Results and Discussion

a) Feature Reduction Analysis

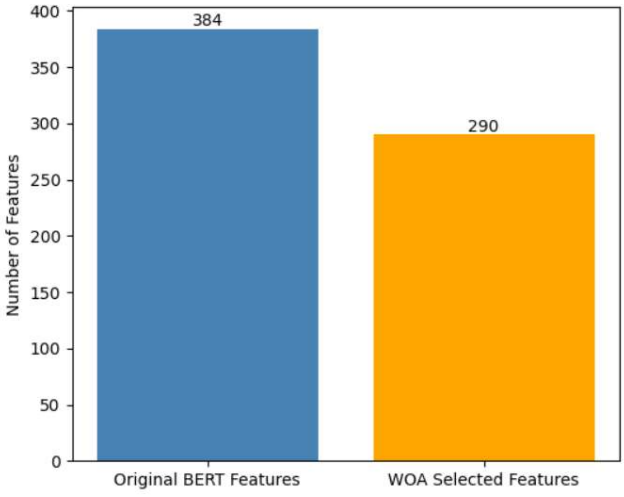

Figure 2 Feature Reduction using WOA

As shown in Figure 2, the original embedding representation of the BERT model has 384 features, which are the semantic features extracted during the embedding process. After applying the proposed WOA-based feature selection approach, the number of features is reduced to 290

features. This represents a reduction of 24.48% in the number of features of the embedding representation of the BERT model, which implies that nearly a quarter of the embedding dimensions are redundant and can be eliminated in the feature selection process without the need to manually extract the features of the embedding representation of the BERT model and without the need to use the optimization algorithm to search the high-dimensional feature space and select a subset of the most informative features.

b) Classification Performance

The classification performance of the original BERT embeddings and the WOA-selected feature subset was analyzed using three ML classifiers: LR, Linear Support Vector Machine (SVM), and RF. The performance of each classifier was assessed using evaluation metrics. These metrics provide a comprehensive evaluation of the classification capability, particularly in datasets that may contain class imbalance.

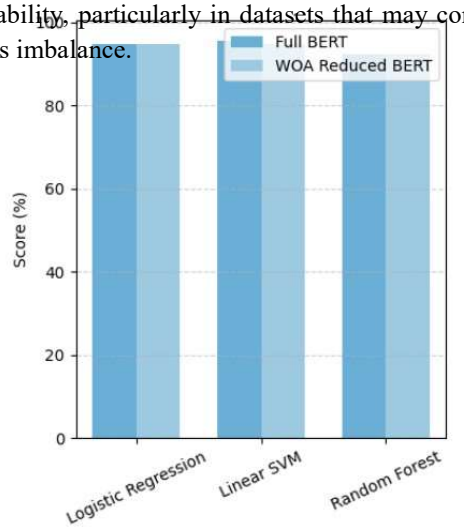

Figure 3 Accuracy across different ML Classifiers

Accuracy assesses the overall performance of the model by calculating correctly identified cases among all predictions. The outcomes shown in Figure 3 demonstrate that, with values surpassing 92%, both the WOA-selected feature subset and the complete BERT representation achieve good accuracy across all classifiers. The WOA-reduced feature representation attains 94.84%, whereas the full BERT representation yields the greatest

accuracy of 95.60%. The comparatively minute difference suggests that the overall classification performance is little impacted by the feature dimensionality reduction. These results show that most of the discriminative information found in the original BERT embeddings is retained in the chosen feature subset.

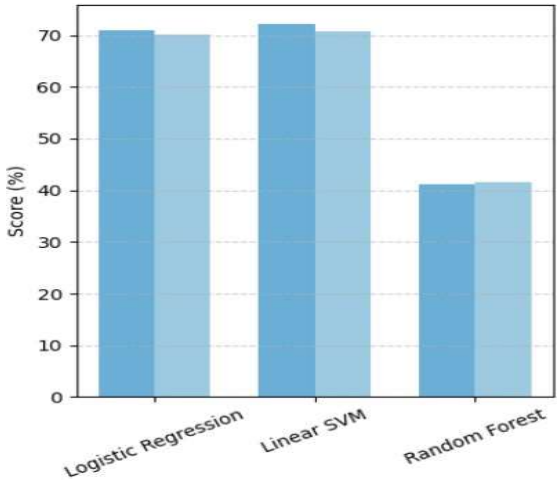

Figure 4 F1-Score across different ML Classifiers

The F1-score is a crucial measure of model efficacy since toxic comment detection necessitates both precise identification and low misclassification. The outcomes shown in Figure 4. demonstrate that the WOA-selected feature subset and the entire BERT representation have similar F1-scores for the linear classifiers. The highest F1-score obtained in the experiments is 72.19%, while the reduced feature representation achieves 70.68%. The slight variation suggests that the suggested feature selection method effectively eliminates unnecessary features while maintaining the most informative embedding dimensions.

Table 3 Comparative Results of WOA reduced feature set across multiple classifiers

Model	Feature Set	Accuracy	Precision	Recall	F1 Score
Logistic Regression	**Full Bert**	94.82%	82.50%	62.31	71%
	WOA Reduced Bert	94.67%	81.51%	61.54 %	70.13 %
Linear SVM	**Full Bert**	95.60%	84.44%	63.05 %	72.19 %
	WOA Reduced Bert	94.84%	83.62%	61.20 %	70.68 %
Random Forest	**Full Bert**	92.32%	93.16%	26.44 %	41.19 %
	WOA Reduced Bert	92.34%	92.66%	26.84 %	41.62 %

The comparative outcomes presented in **Table 3** demonstrate that the differences between the two feature representations are relatively small across all evaluated metrics. The feature selection approach has similar performance when the number of feature dimensions is reduced from 384 to

290. The variation in performance when using a feature subset selected by WOA in comparison with complete BERT representation characteristics is low for all classifiers, which is a key observation from this data set. In past studies on this data set using an LSTM classifier, a classification accuracy of around 94% has been reported [12]. This implies that the characteristics removed by feature selection have a negligible impact on the classification task, which in turn implies that WOA is successful in selecting embedding characteristics that are less relevant.

c) Optimization Convergence Analysis

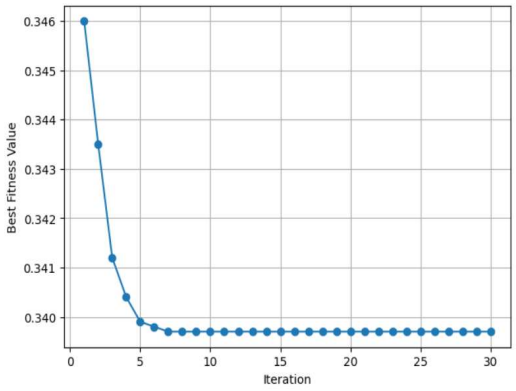

Figure 5 WOA Convergence Curve

During the optimization process, the convergence characteristics of the WOA algorithm have been investigated to check the effectiveness of the proposed feature selection approach. In this approach, each candidate solution or the whales represent the potential sets of the embedding features. A fitness function that balances the compactness of the feature set and the classification accuracy is used to measure the quality of the candidate solutions. Specifically, the fitness function considers two important factors: feature sparsity, which encourages the selection of the compact and smaller feature set, and the F1-score, which measures the classification accuracy of the selected feature set. The LR classifier is used as the fitness function to measure the quality of the feature subset. Using the lightweight classifier during the optimization process is effective because the optimization process can efficiently explore the solution space while maintaining the reliability of the performance estimate. Duringeach evaluation step, the candidate feature subset is used to train an LR model, and its performanceis assessed using three-fold cross-validation. The F1-score obtained from cross-validation is thenused to guide the optimization process. As a result, the algorithm evaluates a large number ofcandidate solutions throughout the optimization procedure..

Considering the optimization configuration used in this study, the total number of model trainings performed during the optimization process can be estimated as given in Eqn (1):

□□□□□ □□□□□□□□□□ = □□. □□ □*h*□□□□ X □□. □□ □□□□□□□□□□□ X □□_□□□□□ (1)

For the experiment design used in this paper, the WOA algorithm has 20 whales and 30 iterations, and the fitness value is computed using the 3-fold cross-validation method. As such, the optimization process entails the evaluation of the model over 1,800 training evaluations in the pursuit of an optimal set of features. This is an extensive process that ensures the algorithm is able to test many possible combinations of features in the embedding space. The convergence behavior observed in the optimization curve shows that the algorithm improves the fitness value significantly in the initial iterations and then stabilizes as it converges towards the optimal value. The stabilization of the fitness value suggests that the algorithm successfully identifies a near-optimal subset of embedding features.

d) Embedding Space Visualization

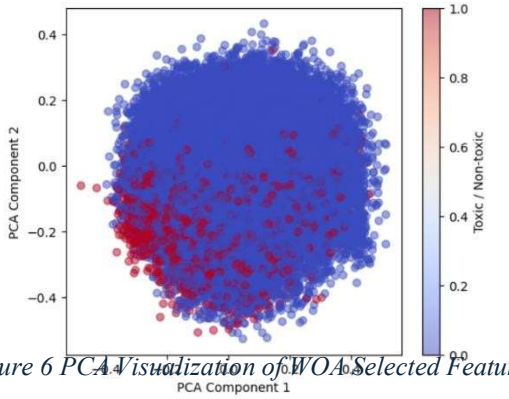

Figure 6 PCA Visualization of WOA Selected Features

The primary purpose of this visualization is to determine if the reduced feature subset is capable of retaining the semantic relationships that are originally present in the embedding space. The WOA is used to reduce the feature subset by eliminating some of the dimensions present in the embedding space. Therefore, it is necessary to determine if the primary structure of the feature representation is retained. If the selected feature subset retains the semantic information present in the embeddings, then the overall structure of the data points should remain coherent and have distinguishable patterns according to their respective classes. Fig. 6 is used to represent the PCA visualization of the feature embeddings obtained after applying the feature selection technique based on the WOA algorithm. The PCA visualization clearly represents that the reduced feature representation retains a similar structure in the distribution of the feature embeddings. Therefore, this represents that the feature selection technique based on the WOA algorithm is effective enough to retain the primary features present in the embedding space while eliminating the unnecessary features. The structure retained in the feature representation clearly represents that

the semantic information required to distinguish between toxic and non-toxic comments is retained.

V. Conclusion

This study offered a framework based on optimization for the optimization of the efficiency of transformer-based text representations for the classification of toxic comments. By leveraging the optimization of the combination of contextual sentence embeddings and the population-based optimization method, the study explored the potential of the optimization method in the selection of the relevant sets of semantic features. This is because the optimization method is capable of efficiently exploring the embedding space and selecting the compact feature representations that maintain the key characteristics of the original semantic features. This reveals the fact that the transformer-based embeddings have some level of redundancy and can be optimized through the optimization method for the selection of the relevant features. Therefore, the optimization method can be used as a tool for the optimization of the selected features of the embeddings.

References

[1] B. R. Naidu, N. Tangudu, K. Kavitha, P. V. Reddy, J. Sahukaru, and R. G. Lopinti, "Toxic Comment Classification using Deep Learning 1," no. April, pp. 93–104, 2023.

[2] J. Sarker, A. K. Turzo, M. Dong, A. Bosu, and W. State, "Automated Identification of Toxic Code Reviews Using ToxiCR," *ACM Trans. Softw. Eng. Methodol.*, vol. 32, no. 1, pp. 1–33, 2023, doi: 10.1145/3583562.

[3] T. Garg, S. Masud, T. Suresh, and T. Chakraborty, "Handling Bias in Toxic Speech Detection : A Survey," vol. 1, no. 1.

[4] N. Reddy, N. Ram, K. K. Kumar, and K. S. R. Murthy, "Toxic Comments Classification," no. June, 2022.

[5] A. Abbasi, A. R. Javed, F. Iqbal, and N. Kryvinska, "Deep learning for religious and continent - based toxic content detection and classification," *Sci. Rep.*, pp. 1–12, 2022, doi: 10.1038/s41598-022-22523-3.

[6] V. R. Pawar, S. D. Garud, A. A. Kadam, and A. G. Khairnar, "Purging the Poison : A Machine Learning Approach to Filtering Toxic Comments," vol. 02, no. July, pp. 2065–2073, 2024.

[7] J. Sarker, S. Sultana, S. R. Wilson, and A. Bosu, "ToxiSpanSE : An Explainable Toxicity Detection in Code Review Comments," 2021.

[8] A. Akshaya, K. Sindhuja, N. Rohan, and Y. Sahas, "OPEN ACCESS MULTILINGUAL TOXIC COMMENTS CLASSIFICATION USING BERT," vol. 15, pp. 67737–67742, 2025.

[9] Z. Zhao, "comparative study of using pre-trained language models for toxic comment classification" 2021.

[10] A. Bonetti, M. Martínez-sober, J. C. Torres, J. M. Vega, and S. Pellerin, "applied sciences Comparison between Machine Learning and Deep Learning Approaches for the Detection of Toxic Comments on Social Networks," 2023.

[11] K. Sharma and V. Pratap, "Classifying Toxic Comments with Machine Learning and Deep Learning Approaches," pp. 1074–1082, 2025.

[12] K. S. Ashok, "A Neuro-NLP Induced Deep Learning Model Developed Towards Comment Based Toxicity Prediction," pp. 94–99, 2022.

[13] T. Davidson, D. Warmsley, M. Macy, and I. Weber, "Automated hate speech detection and the problem of offensive language," in *Proceedings of the 11th International AAAI Conference on Web and Social Media (ICWSM '17)*, Montreal, Canada, 2017, pp. 512–515.

[14] M. Zampieri, S. Malmasi, P. Nakov, S. Rosenthal, N. Farra, and R. Kumar, "Predicting the Type and Target of Offensive Posts in Social Media".

[15] https://www.kaggle.com/datasets/reihanenamdari/youtube-toxicity-data.

[16] C. J. Adams, D. Borkan, I. Incer, J. Sorensen, L. Dixon, L. Vasserman, and N. Thain, *Jigsaw Unintended Bias in Toxicity Classification*, Kaggle, 2019.

[17] Ian Kivlichan, Jeffrey Sorensen, Lucas Dixon, Lucy Vasserman, Meghan Graham, Tin Acosta, and Walter Reade. Jigsaw Rate Severity of Toxic Comments.

https://kaggle.com/competitions/jigsaw-toxic-severity-rating, 2021. Kaggle.

[18] https://www.kaggle.com/datasets/miadul/toxic-comments-detection-dataset/data.

[19] Singhal, D., Verma, A., Radhakrishnan, G. V., Parashar, J., Date, S. S., & Upreti, K. (2025). Whale Optimization and AutoML for Precise Phishing Detection. Journal of Mobile Multimedia, 21(5), 855-880.