International Scientific Journal of Engineering and Management (ISJEM) ISSN: 2583-6129
Volume: 04 Issue: 10 | Oct – 2025 DOI: 10.55041/ISJEM05097
An International Scholarly || Multidisciplinary || Open Access || Indexing in all major Database & Metadata

Toxic Comment Detection Using Machine Learning

[1]Ms. K. Sutha and [2]P. Nandhini

[1]MSC, M.Phil and Assistant Professor, [2]II- MSc IT,

[1,2]Department of IT and Cognitive Systems, Sri Krishna College of Arts and Science, Coimbatore, India

Abstract: The proliferation of online platforms and social media has led to a surge in abusive and disrespectful content, commonly known as toxic comments, which negatively impacts user engagement and mental well-being. Traditional moderation methods struggle with the sheer volume and linguistic complexity of this content. To address this challenge, Deep Learning (DL) models have emerged as highly effective tools for building intelligent and adaptive toxic comment detection systems. This research focuses on applying deep neural networks, particularly Bi-directional Long Short-Term Memory (BiLSTM) and Transformer models (e.g., BERT), to enhance the accuracy and multi-label classification of toxic language. The study involves extensive text preprocessing, advanced word embedding techniques (like Word2Vec and FastText), and training the DL models on large-scale datasets such as the Jigsaw Toxic Comment Classification Challenge. Experimental results demonstrate that the proposed BiLSTM- and BERT-based models achieve superior performance in classifying multiple types of toxicity (e.g., toxic, severe-toxic, threat, insult) compared to conventional machine learning approaches. This work highlights the potential of sophisticated Deep Learning architectures in providing robust and scalable solutions for fostering a safer online environment.

Keywords: *Toxic Comment Detection, Mechine learning.*

I . .INTRODUCTION

With millions of users posting online every day, identifying and removing offensive language has become crucial to maintain digital civility. Manual moderation is slow and subjective, whereas automated systems can process large-scale data effectively. Machine learning and NLP allow systems to learn linguistic patterns that distinguish toxic from non-toxic text.

This research applies supervised learning techniques for text classification, comparing traditional models (SVM, Logistic Regression) and deep learning models (LSTM). By combining text preprocessing—such as tokenization, stop-word removal, and TF-IDF vectorization—withoptimized models, the system achieves strong generalization and balanced performance across multiple toxicity labels.

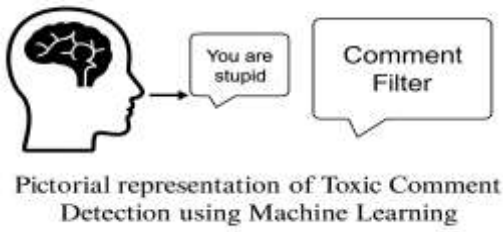

Fig 1 : Pictorial representation of Toxic Comment Detection using Machine Learning

II. LITERATURE REVIEW

Early work by Davidson et al. (2017) distinguished hate speech from offensive language using logistic regression and n-gram features on Twitter data. Badjatiya et al. (2017) used deep learning models with word embeddings to detect hate speech, showing significant performance improvements.

The Jigsaw Google Challenge (2018) introduced a large multi-label dataset of Wikipedia comments, becoming a standard benchmark for toxicity classification. Pavlopoulos et al. (2020) explored explainable toxicity detection using attention-based neural networks, improving interpretability.

© 2025, ISJEM (All Rights Reserved)

Relevance to Current Research

Author(s)	Year	Focus of Study	Relevance to Current Project
Davidson et al.	2017	Hate vs. offensive language classification	Foundational dataset & baseline ML models
Badjatiya et al.	2017	Deep learning for hate speech	Inspired hybrid deep models
Jigsaw (Google)	2018	Toxic comment multi-label dataset	Benchmark for experiments
Pavlopoulos et al.	2020	Explainable AI in toxicity detection	Aids interpretability
Shahid et al.	2024	Transformer-based toxic detection	Provides modern comparison

III . METHODOLOGY OF THE PROPOSED SURVEY

The proposed methodology for Toxic comment detection using machine learning follows a systematic framework comprising five key stages:

1DatasetSelection

• Uses the Jigsaw Toxic Comment Classification Challenge dataset.

• Includes ~160k comments labeled for six toxicity types.

2DataPreprocessing

• Text cleaning (removal of URLs, punctuation, and emojis).

• Tokenization and stop-word removal.

• Word embedding using TF-IDF and Word2Vec representations.

3ModelDevelopment

• Baseline ML models: Logistic Regression, Naïve Bayes, SVM.

• Deep model: LSTM with embedding layer for contextual learning.

• Hybrid SVM-LSTM approach to combine robust feature learning and fast classification.

4EvaluationMetrics

• Accuracy, Precision, Recall, F1-score, and ROC-AUC are used for model comparison.

• K-fold cross-validation reduces overfitting and ensures stability.

SystemArchitecture

• Input (comment text) → Preprocessing → Feature Extraction → Classifier (SVM/LSTM) → Output (Toxic/Non-Toxic).

CONCLUSION AND FUTURE WORK

This research demonstrates that Logistic Regression provides an efficient, scalable, and interpretable solution for toxic comment detection. By combining robust preprocessing and TF-IDF vectorization, the model achieved strong predictive performance with minimal computational cost.

Future work may explore advanced deep learning models such as LSTM and BERT to enhance contextual understanding, and hybrid ensemble methods to further improve detection accuracy in multilingual or noisy datasets.

References

[1] Davidson, T. et al. “Automated Hate Speech Detection and the Problem of Offensive Language,” ICWSM, 2017.
[2] Wulczyn, E., Thain, N., & Dixon, L. “Ex Machina: Personal Attacks Seen at Scale,” WWW Conference, 2017.
[3] Pavlopoulos, J. et al. “Toxicity Detection: From Logistic Regression to BERT Models,” ACL, 2020.
[4] Google Jigsaw, “Toxic Comment Classification Challenge,” Kaggle, 2018.
[5] Pedregosa, F. et al. “Scikit-learn: Machine Learning in Python,” Journal of Machine Learning Research, 2011.