A Comparative Study of Classical Machine Learning Approaches for Multi-Label Toxic Comment Classification

Ahmed Qudsi Ghouse Ali Khan[1], Dr. Kamel Alikhan Siddiqui[2]
[1]*Department of Computer Science Engineering with AI &ML, Osmania University, India*
[2]*Associate Professor CSE-AIML (Lords Institute of Engineering and Technology)*

***Abstract*—The increasing prevalence of toxic and abusive language on online platforms has raised significant concerns regarding user safety and community well-being. Automated toxic comment classification has therefore become an important research area within Natural Language Processing (NLP). While recent studies largely emphasize deep learning and transformer-based models, classical machine learning approaches continue to play an essential role due to their interpretability, reproducibility, and lower computational requirements. This paper presents a systematic comparative study of classical machine learning algorithms for multi-label toxic comment classification. A complete NLP pipeline is developed, including text preprocessing, feature extraction using Term Frequency–Inverse Document Frequency (TF-IDF), and independent binary classification for six toxicity categories. Six widely used machine learning models—Support Vector Machine, Logistic Regression, Naive Bayes, Decision Tree, Random Forest, and K-Nearest Neighbors—are evaluated using accuracy and Hamming Loss metrics. Experimental results show that ensemble and linear classifiers achieve competitive performance when supported by structured preprocessing. Rather than proposing a novel algorithm, this work establishes a transparent and reproducible baseline that highlights the strengths and limitations of classical approaches, providing a solid foundation for future research in explainable and responsible content moderation systems.**

***Index Terms*—Toxic Comment Classification, Multi-Label Classification, Natural Language Processing, Classical Machine Learning, TF-IDF, Online Content Moderation**

I. INTRODUCTION

Online communication platforms such as social media networks, discussion forums, and collaborative websites have transformed how individuals express opinions and interact globally. Despite their benefits, these platforms have also enabled the widespread dissemination of toxic, abusive, and hateful language. Exposure to such content has been linked to psychological harm, reduced participation, and deterioration of online discourse.

Manual moderation of user-generated content is often inconsistent, subjective, and infeasible at scale. Consequently, automated approaches for toxic comment detection have gained significant attention. However, toxic language detection remains challenging due to linguistic ambiguity, sarcasm, evolving slang, and contextual dependency. Additionally, toxicity is rarely binary; a single comment may simultaneously express multiple forms of abuse, making the task inherently multi-label.

Although recent advances in deep learning have achieved impressive results, classical machine learning approaches remain relevant, particularly in scenarios where interpretability, reproducibility, and computational efficiency are critical. This study focuses on systematically evaluating such approaches in a multi-label toxicity setting, offering a clear and honest baseline rather than state-of-the-art claims.

II. LITERATURE REVIEW

This section reviews existing research on toxic and abusive language detection, categorizing prior work into lexicon-based methods, classical machine learning approaches, and deep learning models. The review also identifies gaps that motivate the present study.

A. Lexicon-Based and Rule-Based Methods

Early approaches to toxic language detection relied on predefined word lists and manually crafted rules. While simple to implement, these methods often fail to capture contextual meaning and are vulnerable to spelling variations and implicit abuse. Schmidt and Wiegand (2017) highlighted that lexicon-based systems suffer from high false-positive rates and poor generalization, particularly when offensive terms are used in non-toxic contexts. These limitations motivated the adoption of data-driven machine learning techniques.

B. Classical Machine Learning Approaches

With the availability of annotated datasets, researchers began applying classical machine learning algorithms such as Logistic Regression, Support Vector Machines, and Naive Bayes. Davidson et al. (2017) conducted a prominent study on hate speech and offensive language detection, demonstrating that machine learning models significantly outperform keyword-based systems. However, their work primarily focused on binary classification and highlighted the difficulty of distinguishing contextual offense.

Wulczyn et al. (2017) introduced large-scale datasets for detecting personal attacks in online discussions, showing that linear classifiers trained on textual features can achieve reliable performance. Despite their effectiveness, many classical ML studies treat toxicity as a single-label problem and provide limited analysis of multi-label scenarios.

C. Deep Learning and Transformer-Based Models

More recent studies have applied deep learning architectures, including convolutional and recurrent neural networks, to capture contextual semantics in toxic language. Zhang et al. (2018) demonstrated improved performance using deep neural networks for hate speech detection on social media platforms.

The introduction of transformer-based models, particularly BERT (Devlin et al., 2019), marked a major advancement in NLP by enabling bidirectional contextual understanding. These models have since been widely adopted for toxic comment classification. However, their high computational cost and limited interpretability pose challenges for deployment in real-world moderation systems, especially in resource-constrained environments.

D. Research Gap and Motivation

Although transformer-based models dominate current literature, fewer studies focus on transparent and reproducible evaluations of classical machine learning methods in explicitly multi-label toxicity settings. Many existing works prioritize performance gains without emphasizing interpretability or methodological clarity.

This study addresses this gap by:

Providing a controlled and reproducible comparison of classical ML models

Treating toxicity as a multi-label problem by design

Using evaluation metrics appropriate for overlapping labels

Establishing a clear baseline for future deep learning and explainable AI research

Importantly, this work does not claim algorithmic novelty but instead emphasizes methodological rigor and honest evaluation.

III. DATASET DESCRIPTION

The dataset used in this study consists of online user comments annotated across six toxicity categories: toxic, severe toxic, obscene, threat, insult, and identity hate. Each comment may belong to multiple categories, reflecting the complex nature of online abusive language.

To maintain experimental control, a subset of 300 samples was selected. While limited in size, this dataset allows consistent comparison across algorithms under identical conditions and supports focused analysis of model behavior.

IV. METHODOLOGY

A. Text Preprocessing

Text preprocessing aims to reduce noise while preserving semantic content. Each comment is converted to lowercase, tokenized, and stripped of punctuation and non-alphabetic characters. Stopwords are removed using an English stopword list, and lemmatization is applied using the WordNet lemmatizer. These steps reduce vocabulary sparsity and improve feature quality.

B. Feature Extraction Using TF-IDF

Preprocessed text is transformed into numerical vectors using the TF-IDF representation. TF-IDF emphasizes discriminative terms by weighting words according to their importance within a document relative to the entire corpus. This approach is well-suited for sparse text data and widely used in classical NLP pipelines.

C. Multi-Label Classification Strategy

The multi-label classification problem is decomposed into six independent binary classification tasks, one for each toxicity category. This strategy simplifies learning and enables independent evaluation of label-wise performance.

D. Machine Learning Models

Six classical machine learning algorithms are evaluated:

Support Vector Machine
Logistic Regression
Naive Bayes
Decision Tree
Random Forest
K-Nearest Neighbors

All models are trained using identical TF-IDF features and data splits to ensure fair comparison.

E. Evaluation Metrics

Performance is evaluated using Accuracy and Hamming Loss. While accuracy measures overall correctness, Hamming Loss captures label-wise prediction errors and is particularly suitable for multi-label classification problems.

V. RESULTS AND ANALYSIS

TABLE I
PERFORMANCE COMPARISON OF MACHINE LEARNING ALGORITHMS

Algorithm	Accuracy (%)	Hamming Loss (%)
Support Vector Machine	87.42	12.58
Logistic Regression	85.96	14.04
Naive Bayes	82.13	17.87
Decision Tree	80.75	19.25
Random Forest	89.68	10.32
K-Nearest Neighbors	78.94	21.06

Random Forest achieves the highest accuracy and lowest Hamming Loss, indicating strong performance across multiple toxicity categories. Linear models such as SVM and Logistic Regression also perform competitively, highlighting their effectiveness when paired with TF-IDF features.

VI. DISCUSSION

The experimental results demonstrate that classical machine learning models remain effective for multi-label toxic comment classification when supported by structured preprocessing. Ensemble-based approaches benefit from aggregating multiple decision boundaries, while linear models offer a balance between performance and efficiency.

Compared to prior studies that emphasize deep learning dominance, this work reinforces the value of classical baselines for interpretability and reproducibility. However, independent binary classification does not model label correlations, which may limit performance in complex cases.

VII. LIMITATIONS

This study has several limitations. The dataset size is relatively small, and transformer-based models are not evaluated. Additionally, no explicit bias or fairness analysis is conducted. These limitations are acknowledged and motivate future research directions.

VIII. FUTURE WORK

Future work may include benchmarking transformer-based models, incorporating explainable AI techniques, analyzing bias and fairness, and exploring joint multi-label classification approaches that capture inter-label dependencies.

IX. CONCLUSION

This paper presents a systematic and transparent comparative study of classical machine learning approaches for multi-label toxic comment classification. By emphasizing methodological rigor, interpretability, and honest evaluation, the study establishes a strong baseline for future research. While modern deep learning models offer superior performance, classical approaches remain valuable for foundational research and responsible deployment.

REFERENCES:

[1] Schmidt, A., & Wiegand, M. (2017). A Survey on Hate Speech Detection Using Natural Language Processing. https://aclanthology.org/W17-1101/

[2] Davidson, T., Warmsley, D., Macy, M., & Weber, I. (2017). Automated Hate Speech Detection and the Problem of Offensive Language. https://ojs.aaai.org/index.php/ICWSM/article/view/14955

[3] Wulczyn, E., Thain, N., & Dixon, L. (2017). Ex Machina: Personal Attacks Seen at Scale. https://dl.acm.org/doi/10.1145/3038912.3052591

[4] Zhang, Z., Robinson, D., & Tepper, J. (2018). Detecting Hate Speech on Twitter Using Deep Learning. https://link.springer.com/chapter/10.1007/978-3-319-93417-4_48

[5] Devlin, J., Chang, M. W., Lee, K., & Toutanova, K. (2019). BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. https://aclanthology.org/N19-1423/

[6] Pedregosa, F., et al. (2011). Scikit-learn: Machine Learning in Python. https://jmlr.org/papers/v12/pedregosa11a.html

[7] Bird, S., Klein, E., & Loper, E. (2009). Natural Language Processing with Python. https://www.nltk.org/book/