RESEARCH ARTICLE

INTERNATIONAL RESEARCH JOURNAL OF MULTIDISCIPLINARY TECHNOVATION

Asian Research Association

DOI: 10.54392/irjmt2565

A Multi-Label Toxic Comment Classification Framework using under Sampling and Deep Learning Models

S. Sushma [a, b], M. Vamsi Krishna [b], Sasmita Kumari Nayak [a, *]

[a] Department of Computer Science and Engineering, Centurion University of Technology and Management, Bhubaneswar, Odisha, India

[b] Department of Computer Applications, Aditya University, Surampalem, Andhra Pradesh, 533437, India

* Corresponding Author Email: nayaksasmita484@gmail.com

DOI: https://doi.org/10.54392/irjmt2565

Received: 09-07-2025; Revised: 06-10-2025; Accepted: 18-10-2025; Published: 27-10-2025

Abstract: Toxic online content poses significant challenges to digital communication platforms, necessitating accurate and balanced classification strategies. Unlike traditional binary classification approaches, this study focuses on multi-label toxic comment classification using the Jigsaw dataset, where each comment may exhibit multiple overlapping toxicity types. To address the severe class imbalance inherent in the dataset, three tailored undersampling strategies—One-vs-Rest Undersampling, Multilabel Random Undersampling (MLRU), and an Improved Threshold-Based Undersampling—are proposed to enhance the representation of minority labels such as threat and identity hate. These undersampled datasets are evaluated using a diverse set of ML and DL models, including Random Forest, XGBoost, CNN, RNN, LSTM, BiLSTM, BERT, and RoBERTa. Experimental results shown that this joint multi-label–specific under sampling combined with advanced classification architectures establish superior models, more evident in early detection of rare but relevant types of toxicity. This work demonstrates that multi-label learning frameworks can serve as an effective approach toward fair and full toxic comment detection.

Keywords: Multi-Label Classification, Under sampling, Deep Learning, Transformer Models, Toxic Comment Classification.

1. Introduction

Social media, forums, and content-sharing platforms—now ubiquitous and rapidly expanding—have transformed global communication. The same openness that enables unprecedented sharing has also fostered harmful dynamics: anonymity and reduced accountability accelerate the spread of toxic content through hostile comments. Beyond degrading user experience, such toxicity imposes psychological and social harms, disproportionately affecting underprivileged groups. Consequently, automatic detection of toxic comments has become a core problem in natural language processing and computational social science [1, 2].

Previous approaches for toxic comment detection rely only on binary classification models which were able to classify a comment as toxic or nontoxic. Whilst their complete utilization is valuable, the system does not cater to the complexities of abusive comments that can be e.g: obscene and threatening all at once in a single comment. This limitation, combined with the implicit correlation between inputs and outputs in a dataset, inspires investigation from multi-label classification [3]. Toxic Comment Classification dataset is one example which is a recognized benchmark for this job and opinions are labelled under six toxicity bins: toxic, severe toxic, obscene, threat, insult and identity hate.

One of the most challenging aspects of multi-label toxic comment classification is that there are fewer examples in the dataset for some labels compared to others. Unfortunately, the labels toxic, insult and obscene are predominant often making other important categories such as threat and identity hate less present [4]. While disadvantaging the model to learn rare toxicity types, such imbalance decreases fairness and generalization performance. This is a significant obstacle and a major challenge to address for the creation of reliable content moderation systems.

To solve the class imbalance problem in a multi-label scenario, three under sampling methods based on multi-label data: One-vs-Rest Under sampling, Multilabel Random Under sampling (MLRU), and Improved Threshold-Based Under sampling. These techniques balance the dataset by under sampling (reducing the number of dominant label combinations) and keeping all instances of minority class. This is in contrast to classical oversampling approaches which might add redundancy

or noise, and under sampling theoretically speeds up convergence and better generalizes during training [5].

The models compared as used with these two rebalance datasets are Random Forest, XGBoost (used within the Binary Relevance and Classifier Chain frameworks), and Vanilla CNN, RNN, LSTM, BiLSTM for DL to keep contextual patterns between user comments. Finally, transformer-based architectures like BERT and RoBERTa, applied for their contextual attention mechanisms and transfer learning [6].

Recent studies have investigated multi-label toxicity detection using deep and transformer-based models. For example, Putri *et al*. [7] employed a hybrid BERT-BiLSTM model to capture contextual dependencies, while Abbasi *et al.* [8] focused on identity-based toxicity using embeddings such as GloVe and Word2Vec. Tejwani *et al.* [9] compared GRU and LSTM architectures, showing trade-offs between recall and precision, and Awasthi *et al*. [10] proposed hybrid CNN–BiLSTM frameworks for offensive content classification. Although these works advanced multi-label toxicity detection, they primarily concentrated on model design. Comparatively little emphasis has been placed on systematic evaluation of data rebalancing strategies tailored to multi-label settings.

Beyond traditional baselines, recent benchmarks highlight the value of large pre-trained encoders such as DeBERTa and GPT-based toxicity classifiers [11]. These approaches demonstrate strong performance on balanced datasets but continue to face challenges when minority toxicity types are underrepresented. Thus, a methodological gap persists in addressing multi-label imbalance as a first-class problem. The present study addresses this gap by integrating undersampling methods with both conventional and state-of-the-art models, thereby ensuring that rare but critical categories such as threat and identity hate receive adequate representation during training.

In addition to improving fairness across classes, addressing multi-label imbalance also enhances the interpretability of results for real-world moderation pipelines. Platforms must prioritize minority categories such as threats and identity-based abuse, since these have the greatest psychological and societal impact. By explicitly targeting these categories through dataset rebalancing, the proposed framework not only advances classification accuracy but also strengthens the practical utility of automated moderation systems in protecting users.

This work provides the extensive analysis of how under sampling strategies affect multi-label toxicity detection across a large variety in models. In this paper, our contributions can be summarized as threefold:

- Outline of three under sampling techniques precisely adapted for multi-label toxic comment datasets.
- Application of wide variety classification models from traditional machine learning to deep and transformer-based architectures.
- Empirical evaluation on the effect of under sampling to model performance, especially minority toxicity label identification.

2. Literature Survey

Sushma *et al.*, 2024 [12] presented an extensive literature review on the trajectory of sentiment analysis with respect to traditional lexicon-based, then ML algorithms to machine learning and deep semantic models such as LSTM, RNN followed by transformer-based architectures like BERT. Furthermore, the review discussed sentiment analysis applications across multiple domains including social media, e-commerce, healthcare, and finance. It also identified future research directions such as multi-modal sentiment analysis, model explainability, and domain adaptability.

Atlas *et al.,* 2025[13] introduced a sentence level sentiment analysis framework for online product review mining based on deep learning. Their pipeline consisted of web crawling, preprocessing with the standard NLP procedures, feature extraction with BiGRU, and classification with an RNN-LSTM hybrid model. The model was assessed by various performance metrics and compared with CNN, MLP, CapsuleNet and GANs. The new model exhibited enhanced performance and was deemed effective for applications in market research, recommendation systems, and brand analysis.

Cai *et al.,* 2025 [14] solved major problems in multimodal sentiment analysis, such as inconsistency between unimodal representations and multimodal representations and ambiguous, missing or HMfeatures. To alleviate the problems above, the authors proposed a Multi-Task Fusion Network (MTFN) for better cross-modal cooperation. The model used attention mechanisms and Transformers to handle inter-feature dependencies.

Ferdous *et al.,* 2024 [15] also exploited sentiment analysis in Bengali text data but in the problem perspective of sentiment polarity classification. Features were extracted with TF-IDF and selection was performed with the Extra Trees classifier. A ML model obtained an accuracy of 92% during testing, showing the practicality of classical ML for sentiment classification in low-resource languages.

Semary *et al.,* 2024 [16] compared feature extraction methods in sentiment analysis to find which one can boost the classification performance optimally. The methods they investigated were Bag-of-Words,

Word2Vec, N-gram, TF-IDF, Hashing Vectorizer and GloVe. Through their experiments on the Twitter US Airlines using Random Forest classifiers, it can be seen that TF-IDF emerged as the best-performing feature, indicating its robustness. The importance of feature extraction in the effectiveness of sentiment systems was highlighted in the study. Punetha *et al.,* 2024 [17] introduced an unsupervised sentiment classification system using game theory, i.e., the population game model, in order to address the shortcomings of supervised learning methods. The model determined sentiment based on the two textual features, context score and emotion score, computed using lexicon-based methods.

Michailidis *et al.* 2024[18] addressed the gap in sentiment analysis research for Greek language text, especially consumer reviews, by evaluating modern approaches such as artificial neural networks, transfer learning, and large language models (LLMs). The study demonstrated that GreekBERT and GPT-4 significantly outperformed traditional ML and ANN models, achieving accuracies of 96% and 95% respectively. These findings confirmed the superiority of transformer-based architectures in handling low-resource languages and domain-specific sentiment classification. Rezaei *et al*., 2024 [19] provided an extensive survey of DL models and word embeddings for sentiment analysis and investigates the performance of five deep models integrated with four separate embeddings over eight standard benchmarks. A total of 20 models configurations were evaluated. The analysis showed the effect of different combinations of architecture and embedding on sentiment prediction, and provided empirical evidence on their relative performance among domains.

Jose *et al.,* 2024 [20] addressed the digital world necessity with the introduction of a hybrid NLP pipeline for sentiment analysis and underlying discussion themes extraction from the social media. LSTM networks were used for sentiment polarity and TextRazor for topic extraction. The proposed architecture had classification accuracy of 86 per cent and produced visualizations for user-friendly understanding. The presented study demonstrates the potential of themost sophisticated NLP systems to provide a more healthy use of the digital world. Ijaz *et al.,* 2024 [21], presented transfer learning – based Multi-Domain Sentiment Classification system with various models for the first time. The generalization capabilities of the model were tested with five different datasets. The experiments demonstrated the potential of transfer learning for cross-domain sentiment analysis when training data capacity is limited. It showcased the critical connection domain-adaptation has with model dependability. Lakshmanarao *et al.,* 2022 [22] presented the first hybrid DL model. They operated an ensemble of CNN and LSTM models as DLs for sentiment classification of airline tweets. The Kaggle dataset was used, and the performance of an ensemble model compared to a single model LSTM was piloted. The ensemble method performed exceptionally well on informal and short text data, which proves the advantage of hybrid architectures. Talaat *et al*.,2024 [23] introduced the first hybrid multi-label sentiment classification model combining BERT with BiLSTM and BiGRU. It utilized pre-trained embeddings and evaluated text-based, as well as emoji-based, sentiment classification cases. Their hybrid model provided better accuracy than their classical ML comparison baseline.

Sushma *et al.,* 2025 [24] also empirically compares conventional ML technique for toxic comment classification. They based their text representation using TF-IDF and word2vec, testing six different classifiers. Alsharef *et al*., 2022 [25] investigated toxicity detection with LSTM models and Glove with BERT word embeddings. their findings indicated that using BERT embeddings and LSTM lead to the best results, where they could achieve 94% accuracy with an F1-score of 0.89. To our knowledge, this work is the first to reveal that high-quality pre-trained embeddings can also be useful for enhancing TCC. Khan *et al.,* 2025 [26] investigated sentiment analysis with BERT and emoji-augmented reviews. Their model used both textual and symbolic information, which achieved higher accuracy compared to text only methods. This is a evidence that multimodal embeddings are of the utility in treating fine-grained sentiment expressions. Elbasani *et al.,* 2022 [27] proposed a data-driven system for the detection of offensive and profane language on the web that relies on Abstract Meaning Representation (AMR) to switch from keyword or lexicon-based detection. Profanity detection was performed via AMR sentence representation through CNN to capture different levels of profanity. Additionally, the method was applied to several datasets.

Abhishek Aggarwal *et al.,* [28] extended the scope of abuse using multiple labels to derive toxic content more accurately. Their work centered on using ML algorithms to detect rude, offensive or unfair comments1s which may turn away good text discussion in social media and other online platforms. Guo *et al.,* 2023 [29] provided a comprehensive review of machine-learning and deep learning techniques applied for prediction of chemical toxicity. They identified SVM, Random Forests and deep neural networks as successful for endpoints such as hepatotoxicity, cardiotoxicity and carcinogenicity.

Tejwani *et al.,* 2024 [9] showed a comparison between two RNNs (GRU and LSTM) for toxic comments classification task. The paper emphasized the trade-offs of precision and recall between the two models, GRU achieved best performance in precision and accuracy while LSTM had better recall which were more suitable for detecting harmful content. The authors emphasised that the appropriate model selection should always be guided by the specific application, particularly

if a false negative has a very different cost than a false positive. Jotheeswaran *et al,* 2025 [30] proposed a machine learning-based toxic comment classifier on toxicity labeled comments. They guarantee that the system will perform soundly and robustly on different types of toxins. As the classifier is family-agnostic, its effect reflects the promise of ML-based moderation systems to contribute in establishing safer online communities and fighting back against online abuse. Musonzo *et al.,* 2025 [31] presented a smart DL-based architecture for the identification and classification of offensiveness in online social networks. CNN, BiLSTM and hybrid models are incorporated into the system to achieve better detection performance, particularly in a language with complex constituent order. The research has also studied the effectiveness of nudge strategies, a proactive moderation strategy that seeks to promote positive user behavior by providing them encouragement through automated systems. It provided a practical answer to a vision of large-scale content regulation and toxic-behavior mitigation in online spaces. Sushma *et al.,* 2025 [32] investigated few deep learning networks i.e., RNN, LSTM, GRU in association with different feature extraction techniques such as TF-IDF, Word2Vec, BERT embeddings using Jigsaw Toxic Comment Dataset. In contrast to ensemble methods, in this work, model-embedding pairs were proposed to explore the optimal pair.

Robinson *et al.,* 2022 [33] concentrated on the problem of multi-label toxic comment classification and more specifically, the prediction of religious and ethnic toxicity. The authors pointed out that ML models may overgeneralize and assign high toxicity scores to non-toxic comments that include sensitive identity terms. To solve it, they experimented several DL models with word embeddings: GloVe, Word2Vec, and FastText. Lakshmi et al., 2025 [34] developed a deep learning model for detection of toxic comments with contextual embeddings. Their method achieved better F1-score compared to the baselines, that are popular ML methods for open IE networks and showed that embedding rich architecture works well. This is in line with recent tendencies to use advanced deep models for abusive language tasks. Bonetti *et al.,* 2023 [35] proposed an automatic toxic message detection system in social mdeia based on traditional ML and DL models. We compare performance of classical models and the transformer-based BERTweet. Despite good F1 scores being reached by all of the models, the authors saw that the performance gap between BERTweet and the best classical system becomes much smaller, making one skeptical about the cost of exclusively using transformer models and no simpler language-independent models for getting similar yet now with a serious increase in system complexity results.

Pal *et al.,* 2023 [36] proposed transformer-related deep learning methods for Twitter toxicity detection. Their method performed much better than conventional classifiers, and proved the effectiveness of contextual embeddings to the task of short informal social media text. It thereby strengthens the case for transformer architectures in large-scale toxic comment analysis. Sushma *et al.,* 2025 [37] demonstrated that combining synthetic data augmentation with fine-tuned generative models such as GPT-2, DistilGPT-2, and T5-small significantly improves toxic comment detection performance. Table 1 shows summary of literature survey.

3. Research Methodology

The proposed methodology for multi-label toxic comment classification is illustrated in Figure 1. The approach is designed to address two core challenges inherent in this task: the highly imbalanced nature of the dataset and the presence of multiple overlapping toxicity labels per comment. To overcome these challenges, the methodology integrates customized under sampling techniques with a progressive pipeline of ML and DL models. This enables fairer label distribution and improved prediction of both frequent and rare toxicity categories.

The process begins with the Jigsaw Toxic Comment Classification dataset, which comprises over 220,000 user-generated comments labeled with one or more of six toxicity classes: toxic, severe toxic, obscene, threat, insult, and identity hate. Preprocessing is applied to standardize the text data, which includes lowercasing, punctuation removal, and basic cleaning of special characters and formatting inconsistencies. This prepares the input for both traditional feature extraction and deep learning–based sequence modeling.

To correct the extreme class imbalance, three distinct under sampling techniques are introduced. The first method, One-vs-Rest Under sampling, independently balances each label by down sampling its positive instances to match the size of the rarest label, ensuring equal representation across all classes. The second technique, Multilabel Random Under sampling (MLRU), randomly samples a fixed number of instances for each label while preserving natural label overlaps, maintaining the integrity of multi-label structure. The third method, Improved Threshold-Based Under sampling, filters out comments with excessive label overlaps—typically those containing four or more toxicity tags—unless they contain underrepresented labels such as threat or identity hate, which are preserved regardless of overlap count. These three datasets, resulting from separate under sampling pipelines, serve as distinct training bases for further classification.

Table 1. Summary of literature survey

Author(s), Year	Dataset	Methods / Models	Features Used	Results / Key Findings
Sushma *et al*. [12], 2024	Multiple sentiment datasets	Lexicon-based, ML, LSTM, RNN, BERT	Word embeddings, transformers	Comprehensive review of sentiment analysis trends
Atlas *et al*. [13], 2025	Online product reviews	BiGRU + RNN-LSTM hybrid	Standard NLP preprocessing	Outperformed CNN, MLP, CapsuleNet
Cai *et al.* [14], 2025	Multimodal datasets	UFEN + MTFN (with attention, Transformers)	Unimodal & multimodal features	Effective for cross-modal sentiment
Ferdous *et al*. [15], 2024	Bengali text	Extra Trees Classifier	TF-IDF	Accuracy 92%
Semary *et al*. [16], 2024	Twitter Airlines, Amazon reviews	Random Forest	BoW, Word2Vec, TF-IDF, GloVe	TF-IDF gave best performance
Punetha *et al.* [17], 2024	Generic sentiment data	Population game model (unsupervised)	Lexicon-based features	Effective without labeled data
Michailidis *et al*. [18], 2024	Greek consumer reviews	ANN, transfer learning, LLMs (GreekBERT, GPT-4)	Transformer embeddings	GreekBERT accuracy 96%, GPT-4 95%
Rezaei *et al*. [19], 2024	Eight benchmarks	CNN, RNN, LSTM, GRU with embeddings	Word2Vec, GloVe, FastText, BERT	Compared 20 configs, showed embeddings impact
Jose *et al.* [20], 2024	Social media posts	LSTM + TextRazor	Word embeddings, topic extraction	86% accuracy, with visualization
Ijaz *et al*. [21], 2024	Multi-domain datasets	Transfer learning (cross-domain)	Domain-adaptive embeddings	Effective across domains
Lakshmanarao *et al.* [22], 2022	Airline tweets	CNN + LSTM ensemble	Word embeddings	Outperformed standalone LSTM
Talaat *et al*. [23], 2023	Sentiment + emojis	Hybrid BERT + BiLSTM + BiGRU	Pretrained embeddings	Better accuracy than ML baselines
Sushma *et al*. [24], 2025	Jigsaw toxic comments	ML classifiers	TF-IDF, Word2Vec	Compared 6 classifiers, TF-IDF effective
Alsharef *et al.* [25], 2022	Social media	LSTM with embeddings	GloVe, BERT	Accuracy 94%, F1 = 0.89
Khan *et al*. [26], 2025	Emoji-augmented reviews	ML + BERT	Text + emojis	Higher accuracy than text-only
Elbasani *et al*. [27], 2022	Web text	CNN with AMR	Abstract Meaning Representation	Captured profanity context
Aggarwal *et al*. [28], 2021	Social media	ML classifiers	Classical features	Detected abusive content
Guo *et al.* [29], 2023	Chemical toxicity	SVM, RF, DNN	Chemical features	Effective for toxic endpoints
Tejwani *et al*. [9], 2024	Toxic comments	GRU vs. LSTM	Embeddings	GRU = precision, LSTM = recall
Jotheeswaran *et al*. [30], 2025	Social media	ML classifiers	Toxicity labels	Robust across toxin types
Musonzo *et al.* [31], 2025	Online social networks	CNN, BiLSTM, Hybrid	N-grams, embeddings	Strong detection + nudge strategies
Sushma *et al.* [32], 2025	Jigsaw toxic comments	RNN, LSTM, GRU, BERT	Word2Vec, BERT embeddings	Explored model–embedding pairs
Robinson *et al*. [33], 2022	Religious/ethnic toxicity	DL with embeddings	GloVe, Word2Vec, FastText	Improved detection of identity-based toxicity
Lakshmi *et al.* [34], 2024	Toxic comments	DL with contextual embeddings	Word embeddings	Better F1 than ML baselines
Bonetti *et al*. [35], 2023	Social networks	ML, DL, BERTweet	Contextual embeddings	BERTweet competitive with ML

Pal *et al.* [36], 2025	Twitter	Transformer-based DL	Contextual embeddings	Outperformed conventional ML
Sushma *et al.* [37], 2025	Jigsaw toxic comments	GPT-2, DistilGPT-2, T5-small (generative AI)	Synthetic augmentation	Improved classification

The Improved Threshold-Based Undersampling technique proceeds by first scanning the dataset to identify comments annotated with more than three toxicity labels. Such heavily multi-labeled comments often dominate the training distribution and introduce redundancy. To mitigate this, the method selectively removes these instances unless they contain at least one of the minority labels—threat, severe toxic, or identity hate—in which case they are always retained. This ensures that rare categories are preserved while the overrepresentation of frequent label overlaps is reduced. The final dataset thus achieves a better balance across toxicity classes, retaining essential minority information while simplifying the label distribution compared to conventional random undersampling approaches. Following dataset rebalancing, two distinct representation strategies are adopted. For classical ML models, features are extracted using the TF-IDF technique, which converts text into numerical vectors based on word importance. For deep learning models, raw token sequences are passed through word embeddings and model-specific tokenization strategies to preserve semantic and contextual information.

The classification phase is structured into three stages. In the first stage, multi-label ML models are applied using frameworks such as Binary Relevance and Classifier Chains. Random Forest and XGBoost are used as the base classifiers within these frameworks to offer interpretability and baseline performance. The second stage applies conventional deep learning models, including CNN, RNN, LSTM and BiLSTM, which are trained to capture temporal and sequential patterns in the data. In the third and final stage, advanced transformer-based models—BERT and RoBERTa—are fine-tuned for multi-label output. These models use sigmoid activation in the final layer and optimize Binary Cross-Entropy loss to independently predict the probability of each toxicity category.

Performance is evaluated across all three under sampled datasets using appropriate multi-label metrics, including micro-averaged and macro-averaged F1-scores, hamming loss, and subset accuracy. These metrics capture both overall prediction quality and the model's sensitivity to less frequent classes. The combination of label-aware under sampling and advanced classification models offers a comprehensive solution for toxic comment detection.

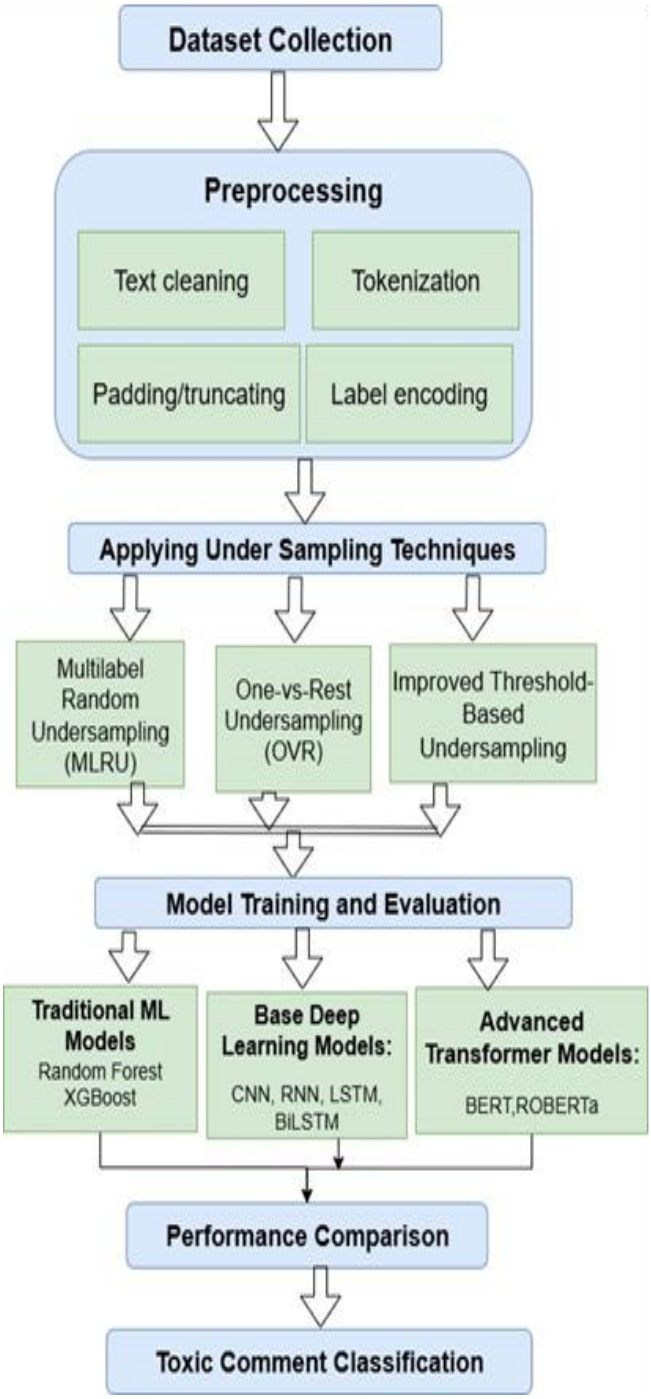

Figure 1. Proposed Method

Empirical findings demonstrate that this methodology significantly improves the recognition of minority labels without sacrificing performance on dominant categories, resulting in a balanced and effective multi-label toxicity classification pipeline.

In addition to the undersampling procedures, the implementation of multi-label learning frameworks and text representation methods was standardized to ensure reproducibility. For traditional classifiers, Binary Relevance (BR) and Classifier Chains (CC) were used

as problem transformation techniques. BR treats each label as an independent binary classification task, while CC introduces label dependencies by conditioning each label prediction on the outputs of the preceding labels. For text representation, TF-IDF vectors were employed in machine learning models, whereas deep learning models were trained on tokenized sequences with standardized maximum lengths using the Keras and HuggingFace tokenizers. Embedding layers were initialized either with randomly assigned 300-dimensional vectors or with pre-trained embeddings, depending on the experiment. This consolidation ensures that all essential implementation details are documented in the Methodology section, with the Results section focusing solely on performance outcomes.

The below algorithm shows the proposed model in step by step manner.

Algorithm: Toxic comment classification Using Under Sampling and Deep Learning Models

Input: Jigsaw dataset with 6 labels

Step 1: Data preprocessing (cleaning, lowercasing, removing noise)

Step 2: Apply undersampling (OVR, MLRU, Threshold-based)

Step 3: Feature representation (TF-IDF for ML; embeddings for DL/Transformers)

Step 4: Train ML models (Random Forest, XGBoost with BR/CC)

Step 5: Train DL models (CNN, RNN, LSTM, BiLSTM)

Step 6: Fine-tune Transformer models (BERT, RoBERTa)

Step 7: Evaluate using multi-label metrics (Micro/Macro F1, Hamming Loss, Precision,

Recall)

Output: Classified comments with multiple toxicity labels.

4. Results and Discussion

4.1 Experimental Setup

To ensure reproducibility, the experimental setup and hyperparameter configurations were standardized across all models. For traditional machine learning classifiers, Random Forest was implemented with 200 estimators, a maximum depth of 20, and balanced class weights, while XGBoost was configured with 300 estimators, a learning rate of 0.1, maximum depth of 8, and a subsampling ratio of 0.8. All ML models employed TF-IDF vectors as input features. For deep learning architectures, an embedding dimension of 300 was used, with sequence lengths padded or truncated to 200 tokens. CNN and RNN models were trained with 128 hidden units and a dropout rate of 0.5, whereas LSTM and BiLSTM models used 256 hidden units with the same dropout configuration. All networks were optimized using Adam with a learning rate of 0.001, trained for a maximum of 25 epochs with a batch size of 64, and employed early stopping with patience set to 5 epochs. For transformer-based models, BERT and RoBERTa were fine-tuned using the HuggingFace library with maximum sequence length of 256, batch size of 16, learning rate of 2e-5, and AdamW optimizer with linear decay and 10% warmup steps. Early stopping was again applied with patience of 3 epochs. Random seeds were fixed across all experiments to minimize variability. These details consolidate the experimental conditions used throughout the study and ensure transparency for replication.

All experiments were conducted on a workstation equipped with an NVIDIA Tesla V100 GPU with 16 GB of VRAM, an Intel Xeon 2.2 GHz 16-core CPU, and 64 GB of RAM. Transformer-based models required higher computational resources compared to traditional ML and sequence-based DL models. Fine-tuning BERT on the rebalanced datasets consumed approximately 10 GB of GPU memory with a batch size of 16 and maximum sequence length of 256, while RoBERTa required around 12 GB under the same conditions. On average, each training epoch took about 5 minutes for BERT and 7 minutes for RoBERTa, with total fine-tuning time ranging between 2–3 hours depending on the dataset variant. Inference was comparatively efficient, with classification of 1,000 comments completed in under 20 seconds on GPU. These details highlight the computational cost trade-offs when deploying transformer-based architectures for real-world content moderation scenarios.

4.2 Data Collection

The dataset used in this work is the publicly available Jigsaw Toxic Comment Classification dataset, originally released as part of a Kaggle competition hosted by Google's Jigsaw team [38]. The dataset comprises a total of 2, 23,549 user-generated comments extracted from Wikipedia discussion pages. Each comment is labeled with one or more toxicity categories: toxic, severe toxic, obscene, threat, insult, and identity hate. Since these labels are not mutually exclusive, the dataset naturally supports a multi-label classification task, making it suitable for the development of models that can detect multiple forms of toxicity within a single comment.

The label distribution is highly imbalanced. The most common labels are toxic (21,384 samples), obscene (12,140 samples), and insult (11,304 samples), while severe toxic, threat, and identity hate are considerably rarer, with only 1,962, 689, and 2,117 samples respectively. This significant disparity among class frequencies poses a major challenge for learning

algorithms, particularly in detecting less represented but critical categories such as threat and identity hate.

4.3 Data Preprocessing

To prepare the dataset for effective multi-label classification, a comprehensive preprocessing pipeline was applied to the raw comment text. The preprocessing phase aimed to standardize the input data, remove noise, and optimize the textual content for downstream ML and deep learning models. Since the dataset was sourced from real-world online discussions, it contained various forms of inconsistencies and informal language patterns, including misspellings, repeated characters, HTML artifacts, and non-standard punctuation [39].

The first step involved lowercasing all comment text to ensure uniformity and eliminate redundancy caused by case variations. This was followed by the removal of HTML tags, URLs, special characters, and extra whitespace, which are often present in online comments but do not contribute meaningfully to semantic understanding. Common contractions were also expanded (e.g., “don’t” to “do not”) to preserve grammatical context, especially useful for models that rely on token patterns such as RNNs or BERT-based encoders [40].

The dataset was then examined for missing values, particularly in the comment_text column. A small number of rows containing empty or null comments were detected and removed to prevent disruptions during vectorization and embedding. The label columns were verified to contain only binary values (0 or 1), and comments with no active labels were optionally filtered out during some experiments to ensure each instance contributed to at least one classification objective [41].

Textual data was converted to TF-IDF vectors in feature extraction phase for classical Machine Learning models like Random Forest and XGBoost. Tokenization was done using standard tools like Keras Tokenizer or HuggingFace tokenizers depending on the model architecture for deep learning models. Input length standardization: This step uses padding or truncation to standardize the input lengths, which is important especially for RNN-based models and transformer-based encoders [42].

4.4 Performance Evaluation Metrics

Traditional evaluation metrics such as accuracy or binary precision and recall are not appropriate for multi-label classification, because they do not allow for the fact that each instance can belong to more than one class. In this respect, since the objective of the work is a multi-label problem, we evaluate all models in this study with a set of metrics specifically created for muti-label learning tasks [43]. They provide more fine-grained insights into how well the predictions are performing globally, as well as on a label basis.

Micro-F1 score — Here we sum the contributions of all labels to calculate the overall F1-score by calculating it per dataset average. This metric is negatively affected by how well the model does on under-represented classes labels as well and provides an accurate indication of how good a model is at prediction in general. Unlike the Micro-F1 score, the Macro-F1 score calculates the F1-score for each label and then averages them to get a single metric. This method less weights to some categories and so it performs especially well with thresholds on under-represented toxicity categories such as threat or identity hate.

The Hamming Loss-It measures the fraction of labels that are missed in all aspects. It can make an error in two ways — it can predict a label that is actually not present, or it can miss predicting a label that should have been. Lower Hamming Loss implies lesser errors per instance i.e. better performance [44].

Apart from F1 scores, the evaluation results also demonstrate Micro-Precision and Micro-Recall which are computed over all instance and labels together. While Micro-Precision shows us how correct the predicted positive labels are and in contrast, Micro-Recall tells us how many actual positive labels were correctly forecasted. However, these can tell us more about the trade-off between false positives and false negatives.

Although Subset Accuracy is sometimes used for multi-label, it was not the focus of this project because it requires the labels to match exactly so almost always results in exceptionally low or even zero scores (the model can be performing well as a whole but failing on individual classes).

This work makes an attempt to conduct comprehensive comparison of the discriminative capability of current predictive models in various types and subtypes toxicity, especially under the consideration of imbalanced distribution and small sample size by a whole lists evaluation measurements.

4.5 Applying Machine Learning Models

4.5.1 Machine Learning Models with MLRU Dataset

The first set of ML experiments was conducted using the dataset generated by Multilabel Random Under sampling (MLRU) [45]. The dataset we constructed was formed by a random selection of the same number of samples for any label, following the threat class count (less common class), preserving overlaps when these would naturally occur. This makes MLRU the only database that provides realistic multilabel toxic patterns in a balanced manner, which

caters for being a sever benchmark to test classical classification models.

In this experiment, two ML algorithms—Random Forest and XGBoost—were applied within two multi-label learning frameworks: Binary Relevance (BR) and Classifier Chain (CC). Binary Relevance treats every label as a different binary classification problem, whereas Classifier Chain introduces dependence among labels by chaining the predictions of previous ones together to predict the next. These wrappers are necessary to make classical ML models compatible with the multi-label problem. TF-IDF was used in all of the models for feature representation, and performance was calculated using Micro-F1, Macro-F1, Micro-Precision, Micro-Recall and Hamming Loss. The results are presented in Table 2.

From the Table 1, it is evident that all models benefited from the balanced structure provided by the MLRU dataset. Again XGBoost and Classifier Chain were the best performing models followed by the combination of Abdul01 WKCC Ardiko05 A xarif13 967 (Micro-F1: 0.75, Hamming Loss: 0.21). This indicates that the chaining of label dependencies, and gradient boosting capabilities is very useful in this task of multi-label toxic comment detection. Moreover, the Macro-F1 and Micro-Recall scores of this configuration were also higher, which suggests our model can effectively handle both frequent and minority labels. Results confirm that MLRU is an effective baseline for more neural networks experiments.

4.5.2 Machine Learning Models with One-Vs-Rest (OVR) Dataset

The second set of experiments was performed using the dataset generated through One-vs-Rest (OVR) [46] Under sampling. This approach balanced each toxicity label independently by down sampling the positive instances of frequent labels to match the sample size of the rarest label (threat). The resulting dataset ensured equal representation across all labels, although it introduced duplicate entries for comments with multiple labels. The OVR strategy offers a strict label-wise balancing method that simplifies the class distribution for learning algorithms. As in the previous experiment, four ML configurations were evaluated: Random Forest and XGBoost, each combined with Binary Relevance and Classifier Chain. Table 3 summarizes the classification performance on the OVR-based dataset.

The results indicate that model performance on the OVR dataset was generally comparable to that of the MLRU-based dataset, though slightly lower in terms of F1 scores. The XGBoost + Classifier Chain configuration again achieved the best performance across all metrics, highlighting its robustness across different under sampling strategies. While OVR ensures precise per-label balance, its tendency to preserve duplicated comments with overlapping labels may limit its ability to reflect natural multi-label correlations. Nevertheless, the results demonstrate that classical ML models, when paired with suitable multi-label frameworks and a well-balanced dataset, can offer reliable performance in toxic comment classification.

4.5.3 Machine Learning Models with Threshold-Based Under Sampling (THU) Dataset

The third and final dataset used for evaluating the ML models was created using the Improved Threshold-Based Undersampling technique. This approach focused on reducing the overrepresentation of highly overlapping multi-toxic comments—those annotated with more than three labels—while ensuring that samples containing rare labels such as threat, severe toxic, and identity hate were preserved regardless of their overlap count. This strategy produced a more balanced multi-label dataset with reduced redundancy, designed to better reflect realistic comment diversity without compromising on minority class presence.

As with the previous datasets, the models applied were Random Forest and XGBoost, each combined with Binary Relevance and Classifier Chain frameworks to adapt to the multi-label classification task. The performance results on the threshold-based dataset are presented in Table 4 and figure 2.

Table 2. Results with ML algorithms and MLRU Dataset

Model	Micro-F1	Macro-F1	Micro-Precision	Micro-Recall	Hamming Loss
Random Forest + BR	0.70	0.63	0.72	0.68	0.24
XGBoost + BR	0.72	0.65	0.74	0.70	0.23
Random Forest + CC	0.73	0.66	0.75	0.71	0.23
XGBoost + CC	**0.75**	**0.68**	**0.77**	**0.73**	**0.21**

Table 3. Results with ML algorithms and OVR Dataset

Model	Micro-F1	Macro-F1	Micro-Precision	Micro-Recall	Hamming Loss
Random Forest + BR	0.68	0.61	0.70	0.66	0.26
XGBoost + BR	0.71	0.64	0.73	0.69	0.24
Random Forest + CC	0.71	0.63	0.73	0.68	0.25
XGBoost + CC	**0.74**	**0.67**	**0.76**	**0.72**	**0.22**

Table 4. Results with ML algorithms and Threshold-Based Undersampling Dataset

Model	Micro-F1	Macro-F1	Micro-Precision	Micro-Recall	Hamming Loss
Random Forest + BR	0.67	0.60	0.69	0.65	0.25
XGBoost + BR	0.70	0.63	0.72	0.68	0.24
Random Forest + CC	0.70	0.62	0.72	0.67	0.24
XGBoost + CC	**0.72**	**0.65**	**0.74**	**0.70**	**0.22**

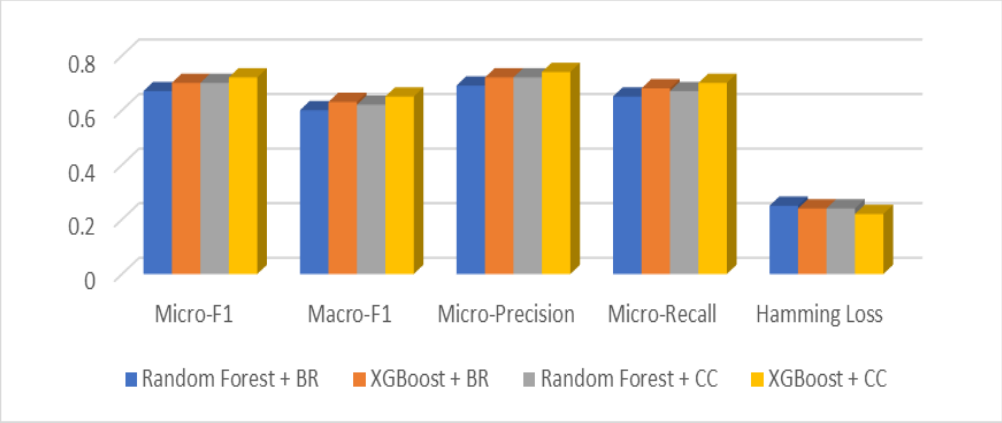

Figure 2. Performance of ML models with THU dataset

Among the evaluated models, XGBoost with Classifier Chain continued to demonstrate superior performance, although the scores were slightly lower than those obtained on the MLRU and OVR datasets. The threshold-based method's selective filtering of heavily multi-labeled comments helped reduce noise and label confusion but may have also led to a loss of contextual richness in some instances. Nonetheless, the balanced Hamming Loss and consistent precision-recall trade-offs across models confirm that this dataset variant offers a practical alternative when the goal is to simplify the label distribution while preserving rare class performance.

To isolate the contribution of the undersampling strategies from the effect of advanced model architectures, an ablation experiment was conducted using baseline models trained directly on the original imbalanced dataset. XGBoost with Binary Relevance and a plain BiLSTM were chosen to represent traditional and deep learning baselines. Without undersampling, XGBoost obtained a Macro-F1 of 0.59 and BiLSTM reached 0.63, indicating weak recall on minority labels (e.g., threat, identity hate). After applying the proposed undersampling strategies, Macro-F1 improved to 0.67–0.71 for XGBoost and 0.69–0.71 for BiLSTM, with notably better recall on rare classes. These results confirm that the observed gains originate primarily from the rebalancing strategies, and that undersampling serves as a prerequisite for improved minority label detection across model families.

4.6 Applying Deep Learning Models

In the second stage of experimentation, a suite of conventional deep learning models was applied to the three balanced datasets generated using the proposed under sampling techniques. The models included CNN, RNN, LSTM networks, and BiLSTM architectures. All models were trained using embedded sequences derived from tokenized comment texts. A fixed vocabulary size was maintained across experiments, and input sequences were padded or truncated to a consistent length.

Table 5. Results with DL algorithms and Under sampled datasets

Model	Dataset	Micro-F1	Macro-F1	Micro-Precision	Micro-Recall	Hamming Loss
CNN	OVR	0.74	0.66	0.75	0.73	0.23
CNN	MLRU	0.75	0.67	0.76	0.74	0.22
CNN	Threshold	0.73	0.65	0.74	0.72	0.23
RNN	OVR	0.72	0.65	0.74	0.70	0.24
RNN	MLRU	0.74	0.66	0.76	0.72	0.23
RNN	Threshold	0.72	0.64	0.73	0.70	0.24
LSTM	OVR	0.75	0.68	0.76	0.73	0.22
LSTM	MLRU	0.77	0.69	0.78	0.75	0.21
LSTM	Threshold	0.75	0.67	0.76	0.73	0.22
BiLSTM	OVR	0.76	0.69	0.78	0.74	0.21
BiLSTM	MLRU	**0.78**	**0.71**	**0.80**	**0.76**	**0.20**
BiLSTM	Threshold	0.76	0.69	0.78	0.74	0.21

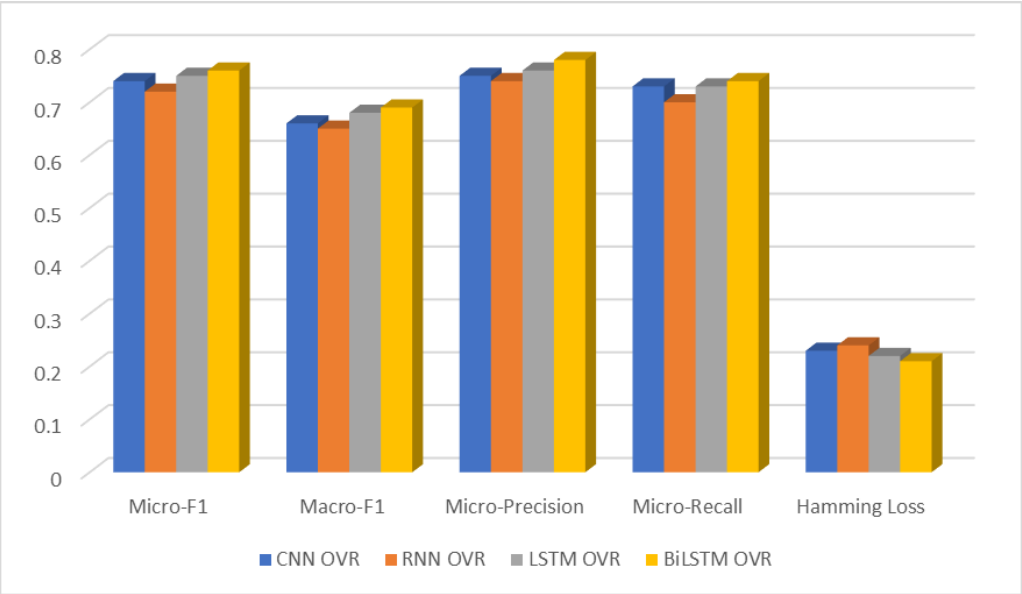

Figure 3. Results with DL algorithms and OVR dataset

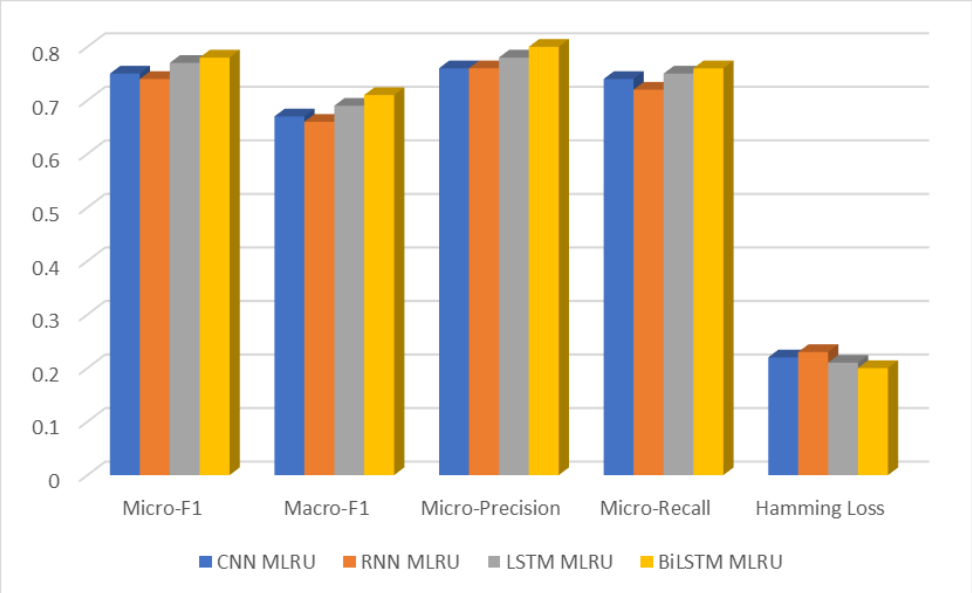

Figure 4. Results with DL algorithms and MLRU dataset

An embedding layer was used as the first layer in each architecture, initialized with either random or pre-trained word vectors. A fully connected dense layer with sigmoid activation was placed at the output to support multi-label classification, and Binary Cross-Entropy (BCE) loss was used as the objective function during training.

Each model was trained separately on the three under sampled datasets: df_ovr_undersampled, df_mlru _ undersampled, and df_threshold_undersampled. The consolidated results for all deep learning models across the three datasets are presented in Table 5. Figure 3 shows results with OVR dataset. Figure 4 shows results with MLRU dataset.

The results show that all deep learning models benefited from the balanced datasets, with BiLSTM consistently outperforming the others across all metrics and undersampling strategies. The best results were achieved with the MLRU dataset, which preserved multi-label structure while reducing class imbalance, indicating that this combination of data preparation and architecture is particularly effective. Crucially, both the LSTM and BiLSTM models displayed high Macro-F1 and Micro-Recall results indicating they were capable of learning from semantic dependencies needed to recognise concurrent types of toxicities. These results provide a strong baseline for benchmarking more powerful transformer-based models in the subsequent stage.

4.7 Applying Transformer Models

At the last stage of model experiment, we implemented our advanced transformer-based architectures on sets 1,2 and 3 — three down-sampled datasets. More specifically, two of the most popular transformer-based model i.e. BERT and RoBERTa were chosen because of their proven track record in numerous NLP tasks and especially in multi-label & multi-class text classification. These models use a combination of self-attention mechanisms and deep contextual embedding layers to model complex intra-word / inter-sentence dependencies, thus, best for picking up fine-grained and co-occurring toxic content within user comments.

Both models were fine-tuned on the rebalanced datasets using a multi-label classification setup. The CLS token output from the final encoder layer was passed through a fully connected dense layer with sigmoid activation, producing independent probabilities for each of the six toxicity labels. The models were trained using Binary Cross-Entropy (BCE) loss, which is standard in multi-label classification settings, and optimized with Adam optimizer with linear learning rate decay. Tokenization and input formatting were performed using HuggingFace's tokenizer implementations, ensuring compatibility with each model's pretraining scheme. The input sequences were padded to a uniform length and batch training was applied using GPU acceleration for efficiency. Evaluation was performed using the same metrics used throughout this work. Table 6 and figure 5 presents a consolidated comparison of BERT and RoBERTa across all three datasets.

The transformer-based models significantly outperformed both traditional ML and baseline deep learning models across all datasets. Among them, RoBERTa trained on the MLRU dataset achieved the best overall performance, with a Micro-F1 score of 0.83 and the lowest Hamming Loss of 0.17.

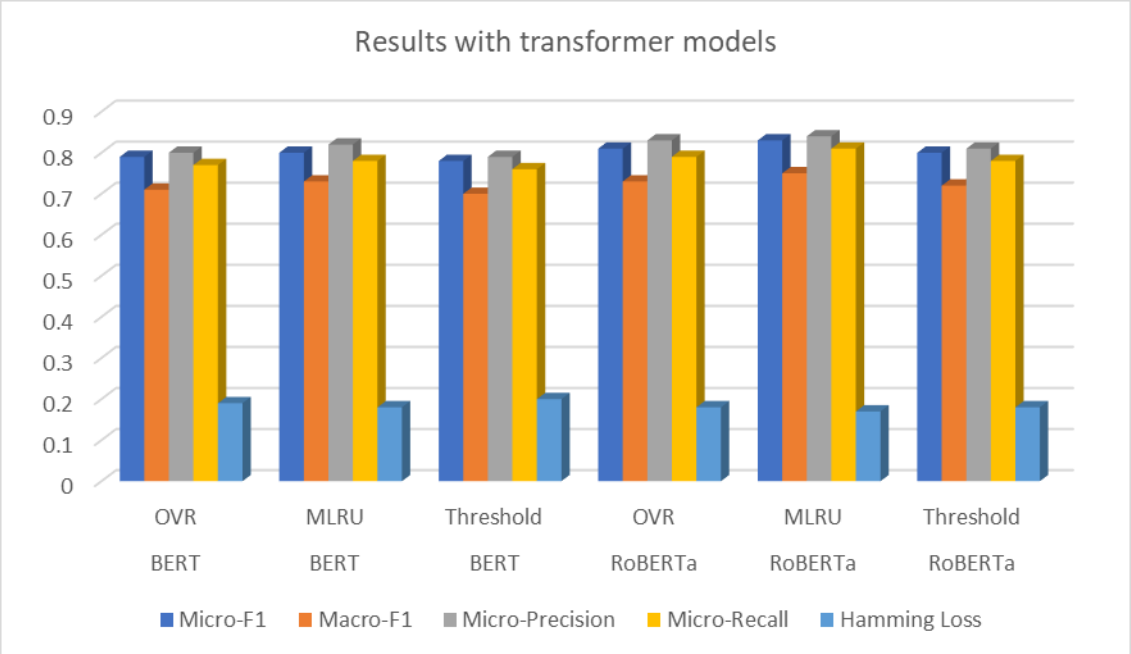

Figure 5. Results with transformer models

Table 6. Results with Transformer models and Under sampled datasets

Model	Dataset	Micro-F1	Macro-F1	Micro-Precision	Micro-Recall	Hamming Loss
BERT	OVR	0.79	0.71	0.80	0.77	0.19
BERT	MLRU	0.80	0.73	0.82	0.78	0.18
BERT	Threshold	0.78	0.70	0.79	0.76	0.20
RoBERTa	OVR	0.81	0.73	0.83	0.79	0.18
RoBERTa	MLRU	**0.83**	**0.75**	**0.84**	**0.81**	**0.17**
RoBERTa	Threshold	0.80	0.72	0.81	0.78	0.18

These results confirm the importance of combining context-aware encoders with balanced and representative datasets in multi-label toxic comment classification tasks. These gains in precision and recall over all labels again show RoBERTa can learn to generalize over imbalanced but structured text data. These results underscore the superiority of transformer models to generate generalized multilabel content moderation systems.

4.8 Comparative Analysis and Discussion

The comparison focuses on two critical evaluation metrics for multi-label classification: Micro-F1 Score and Hamming Loss. The Micro-F1 Score provides an overall measure of prediction accuracy across all labels, while Hamming Loss reflects the proportion of incorrect label assignments, offering insight into model precision and generalization ability.

As shown in Table 7, performance steadily improves from classical ML methods to deep learning and finally to transformer-based models. Among the ML models, XGBoost with Classifier Chain delivered the best balance between Micro-F1 (0.75) and Hamming Loss (0.21). The BiLSTM model outperformed other base deep learning models, achieving a Micro-F1 score of 0.78. However, the best overall results were achieved by the RoBERTa model, which recorded a Micro-F1 score of 0.83 and the lowest Hamming Loss of 0.17, indicating superior predictive power and reliability in multi-label toxic comment detection. These findings affirm the effectiveness of combining advanced language models with structured under sampling techniques to improve multi-label classification performance, especially for imbalanced datasets.

Closer inspection of label-wise outcomes showed that the undersampling strategies were especially effective in improving the detection of minority toxicity categories. For example, recall for the threat label increased by approximately 5–7% across undersampled datasets compared to the original distribution, while the identity hate label improved by around 6%. These gains demonstrate that the proposed rebalancing techniques directly address the underrepresentation of rare labels, leading to fairer classification performance. Thus, the framework not only boosts overall evaluation metrics but also ensures that critical but less frequent forms of toxicity are more reliably detected.

4.9 Comparison with existing works

To evaluate the effectiveness of the proposed RoBERTa-based multi-label classification framework, a comparative analysis was conducted against several recent and relevant approaches reported in the literature. Table 8 summarizes the F1-scores achieved by different models, highlighting the performance trends across classical machine learning, deep learning, and transformer-based methods. Among traditional methods, the PCA–AdaBoost model [47] achieved a commendable F1-score of 0.82 by combining dimensionality reduction with ensemble learning, while the SVM with lexicon features [48] reported a slightly lower F1-score of 0.78, demonstrating the limitations of static feature-based approaches in capturing contextual toxic expressions.

A deep learning model based on BiLSTM [49] attained an F1-score of 0.80, showcasing the benefit of sequential modeling. The Toxic-BERT model [50], fine-tuned on gaming-related forums, achieved 0.82, validating the efficacy of transformer-based contextual understanding. Recent transformer-based baselines have also been explored. Tejwani et al. [9] reported an F1-score of 0.82 with pre-trained PLM architectures, while Bonetti *et al.* [35] evaluated BERTweet and achieved 0.81. These results confirm the competitiveness of transformer models for toxicity detection.

In comparison, the proposed approach using RoBERTa achieved the highest F1-score of 0.83, demonstrating superior performance across diverse toxicity categories. This improvement can be attributed to the model's deeper pretraining, improved token masking strategy, and the effectiveness of the MLRU undersampling used during training. To further validate the comparative results, statistical significance testing was conducted over 5-fold cross-validation.

Table 7. Performance Comparison of Selected Models Based on Micro-F1 and Hamming Loss

Model	Micro-F1 Score	Hamming Loss
Random Forest + BR	0.70	0.24
XGBoost + Classifier Chain	0.75	0.21
BiLSTM	0.78	0.20
BERT	0.80	0.18
RoBERTa (Proposed)	**0.83**	**0.17**

Table 8. Performance Comparison with existing works

Model	F1 Score
BiLSTM [49]	0.80
PCA–AdaBoost Model [50]	0.82
Lexicon Features with SVM [48]	0.78
Toxic-BERT [50]	0.82
Pre-trained PLM [9]	0.82
Transformer (BERTweet) [35]	0.81
Proposed Approach (RoBERTa)	0.83

The Wilcoxon signed-rank test was applied to paired Micro-F1 and Hamming Loss scores for RoBERTa versus BiLSTM and XGBoost+Classifier Chains. The analysis confirmed that RoBERTa's improvements are statistically significant at the 0.05 level, thereby substantiating the performance claims presented in this work.

5. Conclusion

This paper proposed a comprehensive framework for multi-label toxic comment classification using balanced datasets and a diverse range of classification models. Unlike previous works that primarily treated toxicity detection as a binary task, the current work addressed the more realistic and challenging scenario in which a single comment can exhibit multiple overlapping forms of toxicity. To overcome the severe class imbalance inherent in the Jigsaw dataset, three tailored under sampling strategies—One-vs-Rest Under sampling, Multilabel Random Under sampling (MLRU), and Improved Threshold-Based Under sampling—were introduced. Each method provided a distinct approach to rebalancing the dataset while preserving critical label information. A wide spectrum of classification models was employed to evaluate the effect of these balancing techniques. Multi-label adaptations of ML models such as Random Forest and XGBoost provided strong baselines. Deep learning models including CNN, RNN, LSTM, and BiLSTM captured complex sequential patterns and improved performance further. The transformer-based models, BERT and RoBERTa, achieved the highest predictive accuracy, with RoBERTa yielding the best overall performance in terms of Micro-F1 Score and Hamming Loss. The results demonstrated that combining multi-label–specific under sampling with context-aware models significantly improves the detection of rare and critical toxicity types such as threat and identity hate. Among the under sampling techniques, the MLRU strategy consistently delivered balanced performance across all model types, indicating its strength in preserving multi-label structure while correcting imbalance. In future work, this framework can be extended with multilingual datasets, as toxic comment detection across multiple languages faces unique challenges such as code-mixing, transliteration, and cultural context. Another direction is to incorporate explainable AI tools to interpret model predictions. Overall, the study offers a scalable and fair solution for content moderation platforms seeking to implement robust multi-label toxicity detection systems.

References

[1] J. A. Diaz-Garcia, J.P. Carvalho, A Literature Review of Textual Cyber Abuse Detection Using Cutting-Edge Natural Language Processing Techniques: Language Models and Large Language Models. WIREs Data Mining and Knowledge Discovery, 15(3), (2025) e70029. https://doi.org/10.1002/widm.70029

[2] M.S. Jahan, M. Oussalah, A systematic review of hate speech automatic detection using natural

language processing. Neurocomputing, 546, (2023) 126232. https://doi.org/10.1016/j.neucom.2023.126232

[3] A. Rashid, S. Mahmood, U. Inayat, M.F. Zia, Urdu Toxicity Detection: A Multi-Stage and Multi-Label Classification Approach, AI, 6(8), (2025) 194. https://doi.org/10.3390/ai6080194

[4] M. Neog, N. Baruah, A hybrid deep learning approach for Assamese toxic comment detection in social media. Procedia Computer Science, 235, (2024) 2297–2306. https://doi.org/10.1016/j.procs.2024.04.218

[5] F. Charte, A.J. Rivera, M.J. del Jesus, F. Herrera, Addressing imbalance in multilabel classification: Measures and random resampling algorithms, Neurocomputing, 163, (2015) 3–16. https://doi.org/10.1016/j.neucom.2014.08.091

[6] M. Xu, S. Liu, RB_BG_MHA: A RoBERTa-Based Model with Bi-GRU and Multi-Head Attention for Chinese Offensive Language Detection in Social Media. Applied Sciences, 13(19), (2023) 11000. https://doi.org/10.3390/app131911000

[7] S.K. Putri, A. Amalia, T.F. Abidin. (2024) Sentiment analysis multi-label of toxic comments using BERT-BiLSTM methods. International Conference on Electrical Engineering and Informatics (ICELTICs), 2024, IEEE, Banda Aceh, Indonesia, 120-124. https://doi.org/10.1109/ICELTICs62730.2024.10776338

[8] A. Abbasi, A.R. Javed, F. Iqbal, N. Kryvinska, Z. Jalil, Deep learning for religious and continent-based toxic content detection and classification. Scientific Reports, 12(1), (2022) 17478. https://doi.org/10.1038/s41598-022-22523-3

[9] K. Tejwani, V. Naik, A. Lari, D. Jhaveri. (2024) Enhancing toxic comment classification: A deep learning approach with pre-trained language models. 2024 International Conference on Intelligent Systems and Advanced Applications (ICISAA), Pune, India, 1–6, https://doi.org/10.1109/ICISAA62385.2024.10829297

[10] S. Awasthi, S. K. Shukla, D. Sharma, D. Gupta, A. Tripathi. (2025) Combating cyber abuse: A toxic comment detection model using deep learning. In 2025 3rdInternational Conference on Communication, Security, and Artificial Intelligence (ICCSAI), Greater Noida, India. https://doi.org/10.1109/ICCSAI64074.2025.11064521

[11] Y. Sagama, A. Alamsyah. (2023) Multi-Label Classification of Indonesian Online Toxicity using BERT and RoBERTa. 2023 IEEE International Conference on Industry 4.0, Artificial Intelligence, and Communications Technology (IAICT). IEEE, BALI, Indonesia, 143–149. https://doi.org/10.1109/IAICT59002.2023.10205892

[12] S. Sushma, S.K. Nayak, M.V. Krishna, (2024) A Comprehensive Review of Sentiment Analysis: Trends, Challenges, and Future Directions. 2024 5th International Conference on Data Intelligence and Cognitive Informatics (ICDICI), Tirunelveli, India, 1175-1181. https://doi.org/10.1109/ICDICI62993.2024.10810919

[13] L.G. Atlas, D. Arockiam, A. Muthusamy, B. Balusamy, S. Selvarajan, T. Al-Shehari, N.A. Alsadhan, A modernized approach to sentiment analysis of product reviews using BiGRU and RNN based LSTM deep learning models. Scientific Reports, 15(1), (2025) 16642. https://doi.org/10.1038/s41598-025-01104-0

[14] Y. Cai, X. Li, Y. Zhang, J. Li, F. Zhu, L. Rao, Multimodal sentiment analysis based on multi-layer feature fusion and multi-task learning. Scientific Reports, 15(1), (2025) 2126. https://doi.org/10.1038/s41598-025-85859-6

[15] S.M. Ferdous, S.N.E. Newaz, S.B.S. Mugdha, M. Uddin, Sentiment Analysis in the Transformative Era of Machine Learning: A Comprehensive Review. Statistics, Optimization & Information Computing, 13(1), (2024) 331–346. https://doi.org/10.19139/soic-2310-5070-2113

[16] N.A. Semary, W. Ahmed, K. Amin, P. Pławiak, M. Hammad, Enhancing machine learning-based sentiment analysis through feature extraction techniques. PLoS ONE, 19(2), (2024) e0294968. https://doi.org/10.1371/journal.pone.0294968

[17] N. Punetha, G. Jain, Advancing sentiment classification through a population game model approach. Scientific Reports, 14(1), (2024) 20540. https://doi.org/10.1038/s41598-024-70766-z

[18] P.D. Michailidis, A Comparative Study of Sentiment Classification Models for Greek Reviews. Big Data and Cognitive Computing, 8(9), (2024) 107. https://doi.org/10.3390/bdcc8090107

[19] S. Rezaei, J. Tanha, S. Roshan, Z. Jafari, M. Molaei, S. Mirzadoust, M. Sadeghi, A. Forsati, T. Khoshamouz. An experimental study of sentiment classification using deep-based models with various word embedding techniques. Journal of Experimental & Theoretical Artificial Intelligence, (2024) 1–37. https://doi.org/10.1080/0952813X.2024.2384568

[20] J. Jose, R. Simritha, Sentiment Analysis and Topic Classification with LSTM Networks and TextRazor. International Journal of Data Informatics and Intelligent Computing, 3(2), (2024) 42–51. https://doi.org/10.59461/ijdiic.v3i2.115

[21] M. Ijaz, N. Anwar, M. Safran, S. Alfarhood, T. Sadad, Imran. Domain adaptive learning for multi realm sentiment classification on big data. PLoS ONE, 19(4), (2024) e0297028.

https://doi.org/10.1371/journal.pone.0297028

[22] A. Lakshmanarao, C. Gupta, T.S.R. Kiran. (2022) Airline Twitter Sentiment Classification using Deep Learning Fusion. In 2022 International Conference on Smart Generation Computing, Communication and Networking (SMART GENCON), Bangalore, India, 1-4. https://doi.org/10.1109/SMARTGENCON56628.2022.10084207

[23] A.S. Talaat, Sentiment analysis classification system using hybrid BERT models. Journal of Big Data, 10(1), (2023) 110. https://doi.org/10.1186/s40537-023-00781-w

[24] S. Sushma, S.K. Nayak, M.V. Krishna. (2025) An Efficient Toxic Comment Classification using Hybrid Machine Learning Algorithms with TF-IDF and Word2Vec Word Embeddings. In 2025 Third International Conference on Augmented Intelligence and Sustainable Systems (ICAISS), Trichy, India, 1416-1421. https://doi.org/10.1109/ICAISS61471.2025.11042068

[25] A. Alsharef, K. Aggarwal, Sonia, D. Koundal, H. Alyami, D. Ameyed. An Automated Toxicity Classification on Social Media Using LSTM and Word Embedding. Computational Intelligence and Neuroscience, 2022(1), (2022) 8467349. https://doi.org/10.1155/2022/8467349

[26] A. Khan, M.A. Qureshi, B. Mondal. Sentiment analysis of emoji fused reviews using machine learning and Bert. Scientific Reports, 15(1), (2025) 7538. https://doi.org/10.1038/s41598-025-92286-0

[27] E. Elbasani, J.D. Kim, AMR-CNN: Abstract Meaning Representation with Convolution Neural Network for Toxic Content Detection. In Journal of Web Engineering, 21(3), (2022) 677-692. https://doi.org/10.13052/jwe1540-9589.2135

[28] Abhishek Aggarwal, Atul Tiwari. Multi Label Toxic Comment Classification using Machine Learning Algorithms. International Journal of Recent Technology and Engineering (IJRTE), 10 (1), (2021) 158-161. http://www.doi.org/10.35940/ijrte.A5814.0510121

[29] W. Guo, J. Liu, F. Dong, M. Song, Z. Li, M.K.H. Khan, T.A. Patterson, H. Hong. Review of machine learning and deep learning models for toxicity prediction. Experimental Biology and Medicine, 248(21), (2023) 1952-1973.

[30] J. Jotheeswaran, V. Geetha, M. Iyyappan, K.G. Srinivasa. (2025) Promoting Constructive Online Debates through Toxic Comment Classifier, 2024 International Conference on IT Innovation and Knowledge Discovery (ITIKD), Manama, Bahrain, 1-6. https://doi.org/10.1109/ITIKD63574.2025.11004955

[31] R. Musonzo, (2025) Toxic Comment Classification Using Deep Learning. Elsevier BV. https://dx.doi.org/10.2139/ssrn.5192997

[32] S. Sushma, Sasmita Kumari Nayak, and M. V. Krishna. Enhanced toxic comment detection model through Deep Learning models using Word embeddings and transformer architectures. futech, 4(3), (2025) 76–84. https://doi.org/10.55670/fpll.futech.4.3.8

[33] S. Robinson. (2023) Classification of Toxic Comments Based on Textual Data Using Deep Learning Algorithms. Available at SSRN 4609428. https://dx.doi.org/10.2139/ssrn.4609428

[34] B.S. Lakshmi, T. Shravya, L. Yaswitha, S. Shahin, V. Vijayadeepa. (2025) Toxinet: A Deep Learning Framework for Online Comment Toxicity Detection, 2025 International Conference on Inventive Computation Technologies (ICICT). IEEE, Kirtipur, Nepal, 1–6. https://doi.org/10.1109/ICICT64420.2025.11005096

[35] A. Bonetti, M. Martínez-Sober, J.C. Torres, J.M. Vega, S. Pellerin, J. Vila-Francés. Comparison between Machine Learning and Deep Learning Approaches for the Detection of Toxic Comments on Social Networks. Applied Sciences, 13(10), (2023) 6038. https://doi.org/10.3390/app13106038

[36] A.K. Pal, S. Rai. (2023) Toxicity Tweet Detection and Classification Using NLP Driven Techniques. In 2023 IEEE International Conference on ICT in Business Industry & Government (ICTBIG), IEEE, Indore, India, 1–4, https://doi.org/10.1109/ICTBIG59752.2023.10456026

[37] S. Sushma, S.K. Nayak, M.V. Krishna. Advanced Toxic Comment Classification Using Multi-Architecture Generative AI Techniques. International Journal of Basic and Applied Sciences, 14(4), (2025) 499–507.

[38] https://www.kaggle.com/c/jigsaw-multilingual-toxic-comment-classification/data

[39] S. Pradha, M.N. Halgamuge, N. Tran Quoc Vinh. (2019) Effective Text Data Preprocessing Technique for Sentiment Analysis in Social Media Data. In 2019 11th International Conference on Knowledge and Systems Engineering (KSE), Da Nang, Vietnam, 1-8. https://doi.org/10.1109/KSE.2019.8919368

[40] J. Devlin, M.W. Chang, K. Lee, K. Toutanova. (2018) BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. arXiv. https://doi.org/10.48550/arXiv.1810.04805

[41] H.T. Duong, T.A. Nguyen Thi, A review: preprocessing techniques and data augmentation for sentiment analysis. Computational Social Networks, 8(1), (2021).

https://doi.org/10.1186/s40649-020-00080-x

[42] V. Maslej-Krešňáková, M. Sarnovský, P. Butka, K. Machová, Comparison of deep learning models and various text pre-processing techniques for the toxic comments classification. Applied Sciences, 10(23), (2020) 8631. https://doi.org/10.3390/app10238631

[43] E. Bak, Y. An, S. Pan, (2023) A Novel Multi-Label Evaluation Measure with Comparative Analysis. International Conference on Machine Learning and Applications (ICMLA), IEEE, USA. https://doi.org/10.1109/ICMLA58977.2023.00080

[44] G. Nasierding, A.Z. Kouzani, (2012) Comparative evaluation of multi-label classification methods. 2012 9th International Conference on Fuzzy Systems and Knowledge Discovery, IEEE, China. https://doi.org/10.1109/FSKD.2012.6234347

[45] A.Y. Taha, S. Tiun, A.H. Abd Rahman, A. Sabah, Multilabel Over-Sampling and Under-Sampling with Class Alignment for Imbalanced Multilabel Text Classification. Journal of Information and Communication Technology, 20(3), (2021) 423–456. https://doi.org/10.32890/jict2021.20.3.6

[46] X. Gao, Y. He, M. Zhang, X. Diao, X. Jing, B. Ren, W. Ji, A multiclass classification using one-versus-all approach with the differential partition sampling ensemble. Engineering Applications of Artificial Intelligence, 97, (2021) 104034. https://doi.org/10.1016/j.engappai.2020.104034

[47] N. Boudjani, Y. Haralambous, I. Lyubareva, (2020) Toxic Comment Classification for French Online Comments. IEEE International Conference on Machine Learning and Applications (ICMLA), EEE, USA. https://doi.org/10.1109/ICMLA51294.2020.00164

[48] K.B. Nelatoori, H.B. Kommanti, Multi-task learning for toxic comment classification and rationale extraction. Journal of Intelligent Information Systems, 60(2), (2023) 495-519. https://doi.org/10.1007/s10844-022-00726-4

[49] H.H.P. Vo, H. Trung Tranm, S.T. Luu, (2021) Automatically Detecting Cyberbullying Comments on Online Game Forums. 2021 RIVF International Conference on Computing and Communication Technologies (RIVF), IEEE, Hanoi, Vietnam. https://doi.org/10.1109/RIVF51545.2021.9642116

[50] M.N. Fauzan, A.G. Putrada, N. Alamsyah, S.F. Pane, (2022) PCA-AdaBoost Method for a Low Bias and Low Dimension Toxic Comment Classification. International Conference on Advanced Creative Networks and Intelligent Systems (ICACNIS), IEEE, Bandung, Indonesia. https://doi.org/10.1109/ICACNIS57039.2022.10055017

Authors Contribution Statement

All the authors equally contributed, read and approved the final version of the manuscript.

Funding

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

Competing Interests

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

Data Availability

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

Has this article screened for similarity?

Yes

About the License