Recursive Journal of Informatics
Vol. 4, No. 1, March 2026
p-ISSN 2963-5551 https://journal.unnes.ac.id/journals/index.php/rji e-ISSN 2986-6588

Comparative Analysis of BERT, RoBERTa and ALBERT Model Performance with Text Data Augmentation in Multilabel Toxic Comment Classification

Annisa Kunarji Sari[1]*, Zaenal Abidin[2]
[1,2]Informatics Engineering Study Program, Faculty of Mathematics and Natural Sciences,
Universitas Negeri Semarang, Indonesia

Abstract. Toxic comments on social media pose serious challenges to online safety and moderation efforts. These comments are often multilabel in nature and suffer from class imbalance, making them difficult to classify accurately using standard methods.
Purpose: This study investigates the use of three transformer-based language models, BERT, RoBERTa, and ALBERT, for multilabel toxic comment classification through fine-tuning. The main objective is to address class imbalance and evaluate model performance after data augmentation.
Methods/Study design/approach: The Toxic Comment Classification dataset, consisting of six overlapping labels, was used in this study. A data augmentation strategy was applied using synonym replacement techniques from WordNet and easy data augmentation (EDA) to increase the representation of minority classes. After balancing the data, the dataset was split into training, validation, and testing sets. Each transformer model was fine-tuned using the Hugging Face Transformers library with the same hyperparameter settings. Model evaluation was conducted using accuracy, precision, recall, and both micro and macro F1-scores.
Result/Findings: The RoBERTa model achieved the best performance, with 86.73% accuracy and a micro F1-score of 92.35%, outperforming BERT and ALBERT. The macro F1-score also improved significantly compared to previous studies using imbalanced datasets, indicating better recognition of minority classes such as threat and identity hate.
Novelty/Originality/Value: This study highlights the effectiveness of combining text data augmentation with transformer-based models in handling multilabel classification tasks involving imbalanced data. The use of simple augmentation methods notably improves performance and fairness across labels, contributing to the development of more robust toxic comment detection systems.

Keywords: deep learning, natural language processing, toxic comment, BERT, RoBERTa, ALBERT
Received May 2025 / **Revised** March 2026 / **Accepted** March 2026

INTRODUCTION

The proliferation of toxic comments in online discussions poses a significant threat to mental well-being. These comments can lead to serious psychological consequences, including depression and suicidal ideation among affected individuals [1]. Early mitigation strategies often relied on filtering or censoring specific words; however, such methods are insufficient in capturing the deeper semantic meaning of toxic language, especially when harmful intent is conveyed implicitly or through contextual cues [2]. A more promising direction involves leveraging technology to automatically detect and classify toxic content, including hate speech, threats, and insults, thus enabling platforms to foster constructive, inclusive, and meaningful dialogue [3]. Accordingly, there is an increasing demand for advanced natural language processing (NLP) solutions capable of comprehending and categorizing toxic content with greater accuracy.

Building on this need, text classification, particularly in the context of toxic comments, has become a central focus in NLP research. The task involves assigning one or more predefined categories to textual data based on its content [4]. Initial approaches predominantly relied on traditional machine learning algorithms, which required manual feature engineering [5]. This process proved especially difficult for

[1]*Corresponding author.
Email addresses: annisakunarjisari36@students.unnes.ac.id (Sari)
DOI: 10.15294/rji.v4i1.25436

short and highly contextual texts [6]. Such models often struggled to capture semantic subtleties, as they were limited by handcrafted features and frequency-based word representations that lacked contextual awareness. The emergence of deep learning marked a significant advancement in this domain by enabling automatic feature extraction through multiple layers of nonlinear transformations [7]. Among deep learning methods, long short-term memory (LSTM) networks became particularly influential due to their effectiveness in addressing the vanishing gradient problem in recurrent neural networks (RNNs) [8]. Both LSTM and its variant, bidirectional LSTM (BiLSTM), have been applied successfully to various classification tasks by capturing both forward and backward context [9], [10]. These models rely on converting input text into numerical vectors using word embeddings such as Word2Vec or GloVe, which encode semantic similarity based on word co-occurrence patterns.

While word embedding techniques improved semantic representation, they still lacked contextual sensitivity, that is, the same word would have the same vector regardless of its meaning in different sentences. This limitation led to the emergence of contextualized word representation models such as BERT (bidirectional encoder representations from transformers), which utilize transformer architectures to capture bidirectional context in text [11]. These advancements marked a significant shift in NLP by enabling deeper understanding of language semantics and context, especially for complex tasks such as toxic comment classification. Among these transformer-based models, BERT has gained widespread attention for its ability to generate rich contextual embeddings, making it highly suitable for tasks involving nuanced language understanding.

Toxic comment classification has thus emerged as a critical task in NLP, with various studies exploring methods to enhance contextual understanding of text data. The introduction of transformer-based language models, particularly BERT, has significantly advanced this area. Unlike recurrent neural networks (RNNs) that rely on sequential word inputs, BERT leverages bidirectional attention mechanisms to capture both left and right context simultaneously, enabling more accurate textual representations [12], [13]. The model is pre-trained on a massive corpus, including the BooksCorpus (800 million words) and English Wikipedia (2.5 billion words), enabling BERT to acquire a broad understanding of language that can be adapted to specific NLP tasks through fine-tuning [2]. Unlike RNNs, which require slow, from-scratch training and a large amount of labeled data [14], BERT reduces this dependency significantly. Nevertheless, BERT has its limitations, particularly the need for high computational resources, especially when processing long textual inputs [13].

Previous studies have demonstrated BERT's superior performance across various classification tasks. Aggarwal et al. [15] showed that BERT outperformed LSTM and gradient boosted trees in fake news classification, achieving an accuracy of 97.02%. Tao and Fang [16] also found that BERT delivered strong results in multilabel sentiment analysis tasks, while also offering faster training times compared to XLNet. In non-English language contexts, Yu et al. [17] integrated BERT with a BiGRU architecture for classifying Chinese text, achieving high accuracy, precision, and F1-scores, all above 0.9. These findings confirm that transformer-based models like BERT offer substantial improvements in word representation and classification accuracy compared to conventional methods. Despite these advancements, a major challenge that persists is the class imbalance problem, where certain toxic comment categories are underrepresented. This imbalance causes models to struggle in accurately predicting minority classes, leading to skewed evaluation metrics and diminished overall performance [18], [19]. Some previous studies, such as those by [3], [20], did not incorporate adequate data balancing strategies, thereby limiting the practical applicability of their results in real-world scenarios.

Several studies have attempted to address the task of toxic comment classification using different model architectures and feature representations. van Aken et al. [3] introduced gradient boosting decision tree and compared logistic regression, RNNs, CNNs, LSTM, BiLSTM, BiGRU and attention-based BiGRU. Gradient boosting decision tree achieving an F1-score of 0.79% on the Toxic Comment Classification dataset. Mohammed et al. [20] analyzed the performance of deep learning models using fastText, GloVe, and Word2Vec embeddings, with BiGRU + GloVe attaining an F1-score of 0.85%. Meanwhile, Maslej-Krešňáková et al. [18] compared various deep learning and transformer-based algorithms to evaluate the impact of text preprocessing and representation methods in multilabel text classification, finding that a bare BERT-BASE model achieved a micro average F1-score of 0.68%. Zhao et al. [21] conducted a comparative analysis of BERT, RoBERTa, and XLM on the same dataset, with RoBERTa showing superior performance, achieving a micro F1-score of 0.78%. Kumar and Kanisha [22] applied

classical machine learning methods with balanced datasets but did not provide a detailed explanation of their data balancing technique. Singh et al. [23] utilized an AlexNet-based CNN model and achieved a high accuracy of 98%; however, the F1-score was relatively low at 0.79%, indicating possible imbalances in class distribution. Froste and Hosseini [24] found RoBERTa to outperform traditional models like TF-IDF with logistic regression and BiGRU, achieving an F1-score of 0.80%, but did not address data imbalance. Maity et al. [25] used Bi-LSTM with fastText and achieved strong results (F1 > 0.85%), yet struggled with long-range dependencies. Unlike these studies, the present research focuses on transformer models with data balancing to improve multilabel toxic comment classification. Many prior studies did not properly address data imbalance or fully explore data augmentation for multilabel toxic comment classification. Additionally, while BERT performs well, variants like RoBERTa and ALBERT are still underutilized with data balancing strategies.

This study aims to improve multilabel toxic comment classification by combining transformer models with data balancing techniques such as augmentation and undersampling. Two augmentation methods are applied: synonym replacement using WordNet and easy data augmentation (EDA), which introduces variation through simple text manipulations [26], [27]. The research evaluates three language models such as BERT, RoBERTa, and ALBERT. RoBERTa, a robustly optimized variant of BERT trained on larger datasets with improved training strategies, demonstrates superior contextual understanding [28]. Meanwhile, ALBERT, a lite BERT incorporates innovations such as factorized embedding parameterization and cross-layer parameter sharing, significantly reducing model parameters without compromising performance [29]. This study compares their classification performance using accuracy, precision, recall, and both micro and macro F1-scores to determine the impact of data augmentation and model choice in handling multilabel toxic comment classification.

METHODS

This study investigates the effectiveness of fine-tuning three transformer-based language models, including BERT, RoBERTa, and ALBERT, for multilabel toxic comment classification. To address the class imbalance issue commonly found in such datasets, a combination of data augmentation and undersampling techniques is employed. After balancing, the dataset is divided into training and testing sets with a 70:30 ratio, and 10% of the training data is further set aside for validation. Each model is fine-tuned with suitable hyperparameters, trained on processed data, and evaluated using accuracy, precision, recall, and micro/macro F1-scores. The workflow is shown in Figure 1.

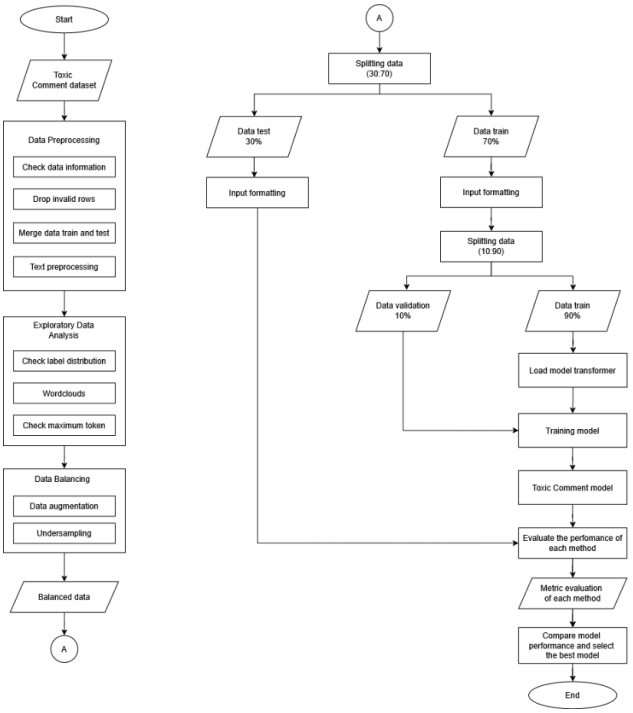

Figure 1. Flowchart of Research

Dataset

The dataset used originates from the Toxic Comment Classification Challenge hosted on Kaggle by Google Jigsaw. It contains a total of 223,549 entries, resulting from the combination of the original training set comprising 159,571 entries and the test set with 63,978 entries. These comments, collected from Wikipedia's talk pages, were annotated by human reviewers and classified into six toxicity categories: toxic, severe toxic, obscene, threat, insult, and identity hate. Each comment can have multiple labels, making this a multilabel classification task. Both subsets were merged into a single dataset to facilitate preprocessing and class balancing. This combination allows for a more flexible handling of label distribution, particularly for minority classes. An overview of the dataset structure is shown in Figure 2, which displays a sample of the data and its multilabel format.

	comment_text	toxic	severe_toxic	obscene	threat	insult	identity_hate
221474	don t post it again or i ll kill you who the h...	1	0	0	1	0	0
162538	tropical cyclone formation alert has been issued	0	0	0	0	0	0
214032	and challenge their very validity	0	0	0	0	0	0
179979	violation in deleting my wikiracist article as...	1	0	1	0	0	0
12475	response why is it funny because i m right you...	1	0	0	0	0	0

Figure 2. Sample of the Toxic Comment Classification Dataset

A closer look at the label distribution is shown in Figure 3, which highlights a severe class imbalance across the dataset. The "toxic" label is the most frequent, followed by "obscene", “insult,” and identity hate”, while categories such as “threat” and “severe toxic” are significantly underrepresented.

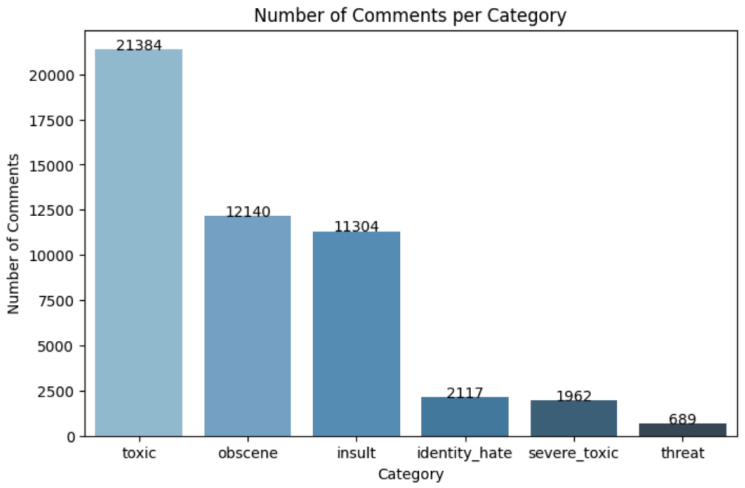

Figure 3. Label Distribution in the Toxic Comment Dataset

Although not visualized in the bar chart, the dataset also contains 201,081 comments with no toxic labels (*non-toxic* class). These entries were included in the dataset and later reduced through undersampling to mitigate the imbalance between toxic and non-toxic instances.

This imbalance highlights the necessity of applying class balancing techniques prior to model training. Without such measures, the model may become biased toward the majority classes and fail to accurately identify instances from minority classes, especially in multilabel settings where overlapping label distributions can obscure the learning process [30], [31].

Data Preprocessing

The preprocessing phase began by removing rows with invalid or missing label values to ensure the dataset's integrity. Text normalization steps were then applied, including converting all text to lowercase and eliminating unnecessary elements such as hashtags, numbers, user mentions, URLs, special

characters, and excessive whitespace. These operations aimed to standardize the input and reduce noise, thereby improving the effectiveness of downstream classification tasks.

Data Balancing

The study combines data augmentation and undersampling to address the class imbalance. Augmentation is first applied to minority classes to match the number of samples in the toxic class. This is performed through synonym replacement based on WordNet, followed by EDA techniques. The EDA techniques include synonym replacement (SR), where selected words in a sentence are replaced with their synonyms, random insertion (RI), which involves adding new synonyms into random positions within a sentence, random swap (RS), where two words in a sentence are randomly exchanged, and random deletion (RD), which removes words randomly from a sentence. These techniques introduce semantic and structural variation without altering the underlying meaning of the text.

Once minority classes were sufficiently expanded, undersampling was applied to the non-toxic class to match the number of toxic samples, ensuring a balanced class distribution for model training. This technique reduces the number of majority class samples in the dataset to achieve a more even distribution across classes. It is performed by adjusting the number of samples between minority and majority classes through random or systematic removal of majority class instances [32].

Splitting Data

After balancing, the dataset is divided into 70% for training and 30% for testing, following the recommendation by Xu and Goodacre [33]. To support model tuning, the training set is further partitioned, with 90% used for actual training and the remaining 10% for validation, as adopted by Wang et al. [34]. This strategy ensures that models can be properly tuned and validated before final evaluation.

Prediction Modeling Algorithm

After completing the data preparation and splitting steps, the study implemented, fine-tuned, and evaluated three transformer-based language models: BERT, RoBERTa, and ALBERT. Each model is fine-tuned independently. Tokenization is conducted using the appropriate tokenizer: bert-base-multilingual-cased for BERT, roberta-base for RoBERTa, and albert-base-v2 for ALBERT. BERT and ALBERT apply WordPiece tokenization with special tokens '[CLS]' and '[SEP]', while RoBERTa uses byte-pair encoding (BPE) with '<s>' and '</s>' tokens. Padding ensures uniform input length, and each input is formatted into token IDs and attention masks, with segment IDs specifically used for BERT and ALBERT. Fine-tuning is performed with hyperparameters adapted from Devlin et al. [12], optimizing settings such as learning rate, batch size, epochs, and optimizer choice. Binary cross-entropy loss is employed to minimize classification error across multiple labels. After training, models predict probabilities for each label using a sigmoid activation function. The trained models are evaluated based on accuracy, precision, recall, micro and macro F1-scores to compare their effectiveness in multilabel classification tasks.

RESULTS AND DISCUSSION

This section presents the experimental results of each research phase, including data preprocessing, balancing, and model evaluation. The discussion focuses on interpreting these results and their implications for the research objectives.

The preprocessing phase was conducted to ensure data quality and consistency. This step involved text normalization and removal of noise, which contributed to more reliable results in subsequent augmentation and classification stages.

Following preprocessing, class imbalance was addressed using a combination of data augmentation and undersampling. Augmentation was conducted for minority labels using WordNet-based synonym replacement and EDA techniques. The “severe toxic” label was excluded from the augmentation process due to its frequent co-occurrence with other labels, which made independent augmentation impractical and less meaningful.

Undersampling was subsequently applied to the non-toxic class by reducing its sample count to match that of the toxic class, resulting in 21,384 entries per class. To ensure the representational quality of the

non-toxic class was preserved, the samples were randomly selected while maintaining content diversity. The number of samples before and after augmentation and undersampling is summarized in Table 1.

Table 1. Sample Distribution Before and After Balancing

Label	Before augmentation	After augmentation
Toxic	21.384	21.384
Obscene	12.140	21.388
Insult	11.304	21.388
Identity hate	2.117	21.385
Severe toxic	1.962	-
Threat	689	21.385

After the balancing process, the dataset consisted of 98,696 entries with a more evenly distributed set of labels. The dataset was then partitioned into training and testing sets using a 70:30 split ratio. Subsequently, 10% of the training set was further set aside as a validation set. The final data distribution across the training, validation, and testing sets is summarized in Table 2.

Table 2. Final Dataset Split After Balancing

Train	**Validation**	**Test**
62.178	6.909	29.609

Following the data preparation and splitting process, the next stage involved implementing, fine-tuning, and evaluating three transformer-based models, BERT, RoBERTa, and ALBERT. Each model was integrated and fine-tuned using the Hugging Face Transformers library, which provides a streamlined pipeline for working with pre-trained language models. The implementation workflow included tokenization, model configuration, training, and evaluation.

For tokenization, model-specific tokenizers such as BertTokenizer, RobertaTokenizer, and AlbertTokenizer were employed to convert raw text into numerical inputs consisting of input IDs and attention masks, with appropriate handling of padding and truncation to a maximum sequence length of 100. These inputs were then transformed into PyTorch tensors to be processed by the models during training. Each model was trained using a batch size of 16, a learning rate of 2e-5, and three training epochs, utilizing the AdamW optimizer.

Following preprocessing and input formatting, model training and evaluation were conducted using four key metrics: accuracy, precision, recall, and F1-score reported in both micro and macro averages to account for multilabel characteristics. The evaluation results reveal that RoBERTa consistently outperformed the other models, achieving the highest overall performance with an accuracy of 86.73%, precision of 91.00%, recall of 93.00%, and F1-score of 92.35%. BERT closely followed, demonstrating an accuracy of 86.58% and an F1-score of 92.10%, highlighting its strong generalization capability.

In contrast, ALBERT achieved the highest precision, 94.00%, indicating a cautious prediction behavior that minimizes false positives. However, this came at the cost of a lower recall, 85.00%, and F1-score, 88.93%, suggesting a trade-off due to its parameter reduction strategy and lightweight architecture. These findings imply that while ALBERT excels in precision, it may overlook some relevant instances, making it less suitable in critical recall scenarios. A detailed summary of each model's performance across the evaluation metrics is presented in Table 3.

Table 3. Evaluation results for BERT, RoBERTa, and ALBERT

Model	Accuracy	Precision	Recall	F1-score
BERT	86.58	93.00	91.00	92.10
RoBERTa	86.73	91.00	93.00	92.35
ALBERT	82.50	94.00	85.00	88.93

To comprehensively evaluate model performance, the proposed approach was assessed using accuracy, micro F1-score, and macro F1-score, and the results were compared with previous studies utilizing the same toxic comment dataset. While accuracy is often reported in classification tasks, it alone may not adequately capture the effectiveness of multilabel classification, particularly under class imbalance

conditions. In this study, BERT and RoBERTa achieved competitive accuracy scores of 86 percent, while ALBERT achieved 82 percent.

More importantly, BERT and RoBERTa attained superior micro and macro F1-scores of 0.92%, surpassing those reported in prior research. For instance, Maslej-Krešňáková et al. [18] reported micro and macro F1-scores of 0.69% and 0.55%, respectively, while Zhao et al. [21] achieved approximately 0.78% and 0.65%. Although Singh et al. [23] reported a notably high accuracy of 98 percent using AlexNet with fastText, their micro F1-score reached only 0.79%, reinforcing the notion that accuracy alone may be misleading in multilabel contexts.

Despite slightly lower F1-scores, ALBERT still performed respectably, achieving 0.89% for both micro and macro F1, while offering the advantage of parameter efficiency due to its compact architecture. A summary of comparative results between the proposed models and prior studies is presented in Table 4.

Table 4. Comparison of Model Performance with Previous Studies

Reference	Dataset	Model	Accuracy	F1-score	
				Micro	Macro
[3]	unbalanced	gradient boosting decision tree	-	-	0.79
[20]	unbalanced	bidirectional GRU + GloVe embedding	-	-	0.85
[18]	unbalanced	bare BERT-BASE uncased	0.89	0.69	0.55
[21]	unbalanced	BERT	-	0.78	0.63
		RoBERTa	-	0.78	0.65
[22]	balanced	random forest + TF-IDF	0.84	-	-
[23]	unbalanced	AlexNet + fastText	0.98	0.79	-
[24]	unbalanced	RoBERTa	-	-	0,80
Proposed method	balanced	BERT	0.86	0.92	0.92
		RoBERTa	0.86	0.92	0.92
		ALBERT	0.82	0.89	0.89

In a further label-wise comparison with the BiLSTM and fastText model by Maity et al. [25], the proposed RoBERTa model demonstrated superior performance in less frequent categories such as threat and identity hate, achieving F1-scores of 0.99% and 0.96%, respectively. These scores significantly surpass those of the baseline model, which obtained 0.86% and 0.88% in the same categories. Conversely, the BiLSTM and fastText models performed better in more frequent labels like toxic, obscene, and insult. This variation highlights that while RoBERTa excels at capturing subtle toxic expressions in underrepresented categories, performance can still vary depending on label frequency and characteristics. The detailed F1-score comparison for each label is presented in Table 5. Overall, the use of balanced data combined with transformer-based architectures in this study yielded strong and consistent results, especially in the context of multilabel classification.

Table 5. Comparison of F1-scores

Label	RoBERTa	BiLSTM + fastText [25]
Toxic	0.90	0.93
Obscene	0.90	0.94
Threat	0.99	0.86
Insult	0.87	0.96
Identity hate	0.96	0.88

CONCLUSION

This study demonstrated the effectiveness of text data augmentation in enhancing the performance of transformer-based models such as BERT, RoBERTa, and ALBERT for multilabel toxic comment classification. By addressing class imbalance through synonym replacement using WordNet and EDA techniques, the representation of minority classes was significantly improved, resulting in better learning outcomes. RoBERTa achieved the highest performance, with 86.73% accuracy and a micro F1-score of 92.35%, confirming its strong capability in capturing nuanced toxic language. The improved macro F1-scores compared to prior studies on imbalanced datasets further highlight the benefits of balanced data for recognizing all label categories fairly. This study highlights how thoughtful preprocessing and model selection can meaningfully contribute to online content moderation efforts. Future research may explore hyperparameter tuning techniques, such as grid search or Bayesian optimization, to optimize model performance further. In addition, testing models across different domains or platforms would help assess

generalizability and robustness in varied real-world contexts. These directions can support the development of more adaptive and high-performing toxic comment classification systems.

REFERENCES

[1] C. Maurya, T. Muhammad, P. Dhillon, and P. Maurya, "The effects of cyberbullying victimization on depression and suicidal ideation among adolescents and young adults: a three year cohort study from India," *BMC Psychiatry*, vol. 22, no. 1, Dec. 2022, doi: 10.1186/s12888-022-04238-x.

[2] R. Rivaldo, A. Amalia, and D. Gunawan, "Multilabeling Indonesian toxic comments classification using the bidirectional encoder representations of transformers model," in *2021 International Conference on Data Science, Artificial Intelligence, and Business Analytics, DATABIA 2021 - Proceedings*, Institute of Electrical and Electronics Engineers Inc., 2021, pp. 22–26. doi: 10.1109/DATABIA53375.2021.9650126.

[3] B. Van Aken, J. Risch, R. Krestel, and A. Löser, "Challenges for toxic comment classification: an in-depth error analysis," in *2nd Workshop on Abusive Language Online - Proceedings of the Workshop, co-located with EMNLP 2018*, Association for Computational Linguistics (ACL), 2018, pp. 33–42. doi: 10.18653/V1/W18-5105.

[4] M. Thangaraj and M. Sivakami, "Text classification techniques: a literature review," *Interdiscip. J. Information, Knowledge, Manag.*, vol. 13, pp. 117–135, 2018, doi: 10.28945/4066.

[5] X. Zhang and Y. LeCun, "Text understanding from scratch," *CoRR*, vol. abs/1502.01710, 2015, [Online]. Available: http://arxiv.org/abs/1502.01710

[6] J. H. Wang, T. W. Liu, X. Luo, and L. Wang, "An LSTM approach to short text sentiment classification with word embeddings," *Proc. 30th Conf. Comput. Linguist. Speech Process. ROCLING 2018*, pp. 214–223, 2018.

[7] L. Deng and D. Yu, "Deep learning: methods and applications, " vol. 7, no. 3–4. 2013. doi: 10.1561/2000000039.

[8] J. L. Huan, A. A. Sekh, C. Quek, and D. K. Prasad, "Emotionally charged text classification with deep learning and sentiment semantic," *Neural Comput. Appl.*, vol. 34, no. 3, pp. 2341–2351, Feb. 2022, doi: 10.1007/s00521-021-06542-1.

[9] A. Graves and J. Schmidhuber, "Framewise phoneme classification with bidirectional LSTM and other neural network architectures," in *Neural Networks*, Jul. 2005, pp. 602–610. doi: 10.1016/j.neunet.2005.06.042.

[10] A. Aziz Sharfuddin, M. Nafis Tihami, and M. Saiful Islam, "A deep recurrent neural network with BiLSTM model for sentiment classification," in *2018 International Conference on Bangla Speech and Language Processing, ICBSLP 2018*, IEEE, Nov. 2018.

[11] J. Alghamdi, Y. Lin, and S. Luo, "A comparative study of machine learning and deep learning techniques for fake news detection," *Inf.*, vol. 13, no. 12, Dec. 2022, doi: 10.3390/info13120576.

[12] J. Devlin, M. W. Chang, K. Lee, and K. Toutanova, "BERT: pre-training of deep bidirectional transformers for language understanding," *NAACL HLT 2019 - 2019 Conf. North Am. Chapter Assoc. Comput. Linguist. Hum. Lang. Technol. - Proc. Conf.*, vol. 1, pp. 4171–4186, Oct. 2018, Accessed: May 17, 2023. [Online]. Available: https://arxiv.org/abs/1810.04805v2

[13] H. Lu, L. Ehwerhemuepha, and C. Rakovski, "A comparative study on deep learning models for text classification of unstructured medical notes with various levels of class imbalance," *BMC Med. Res. Methodol.*, vol. 22, no. 1, Dec. 2022, doi: 10.1186/s12874-022-01665-y.

[14] M. Hadikhah Mozhdehi and A. M. Eftekhari Moghadam, "Textual emotion detection utilizing a transfer learning approach," *J. Supercomput.*, 2023, doi: 10.1007/s11227-023-05168-5.

[15] A. Aggarwal, A. Chauhan, D. Kumar, M. Mittal, and S. Verma, "Classification of fake news by fine-tuning deep bidirectional transformers based language model," *EAI Endorsed Trans. Scalable Inf. Syst.*, vol. 7, no. 27, pp. 1–12, 2020, doi: 10.4108/eai.13-7-2018.163973.

[16] J. Tao and X. Fang, "Toward multilabel sentiment analysis: a transfer learning based approach," *J. Big Data*, vol. 7, no. 1, Dec. 2020, doi: 10.1186/s40537-019-0278-0.

[17] Q. Yu, Z. Wang, and K. Jiang, "Research on text classification based on bert-bigru model," in *Journal of Physics: Conference Series*, IOP Publishing Ltd, Jan. 2021. doi: 10.1088/1742-6596/1746/1/012019.

[18] V. Maslej-Krešňáková, M. Sarnovský, P. Butka, and K. Machová, "Comparison of deep learning models and various text preprocessing techniques for the toxic comments classification," *Appl. Sci.*, vol. 10, no. 23, p. 8631, Dec. 2020, doi: 10.3390/app10238631.

[19] M. Ibrahim, M. Torki, and N. El-Makky, "Imbalanced toxic comments classification using data augmentation and deep learning," in *Proceedings - 17th IEEE International Conference on Machine Learning and Applications, ICMLA 2018*, Institute of Electrical and Electronics Engineers Inc., Jan. 2019, pp. 875–878. doi: 10.1109/ICMLA.2018.00141.

[20] H. H. Mohammed, E. Dogdu, A. K. Gorur, and R. Choupani, "Multilabel classification of text documents using deep learning," in *2020 IEEE International Conference on Big Data (Big Data)*, IEEE, Dec. 2020, pp. 4681–4689. doi: 10.1109/BigData50022.2020.9378266.

[21] Z. Zhao, Z. Zhang, and F. Hopfgartner, "A comparative study of using pre-trained language models for toxic comment classification," in *Companion Proceedings of the Web Conference 2021*, New York, NY, USA: ACM, Apr. 2021, pp. 500–507. doi: 10.1145/3442442.3452313.

[22] K. A. Kumar and B. Kanisha, "Analysis of multiple toxicities using ML algorithms to detect toxic comments," *2022 2nd Int. Conf. Adv. Comput. Innov. Technol. Eng. ICACITE 2022*, pp. 1561–1566, 2022, doi: 10.1109/ICACITE53722.2022.9823822.

[23] I. Singh, G. Goyal, and A. Chandel, "AlexNet architecture based convolutional neural network for toxic comments classification," *J. King Saud Univ. - Comput. Inf. Sci.*, vol. 34, no. 9, pp. 7547–7558, 2022, doi: 10.1016/j.jksuci.2022.06.007.

[24] M. Froste and M. Hosseini, "Multilabel toxic comment classification using machine learning : an in-depth study," 2023.

[25] A. Maity, R. More, A. Patil, J. Oza, and G. Kambli, "Toxic comment detection using bidirectional sequence classifiers," *2nd Int. Conf. Intell. Data Commun. Technol. Internet Things, IDCIoT 2024*, pp. 709–716, 2024, doi: 10.1109/IDCIoT59759.2024.10467922.

[26] Princeton University, "About wordnet," WordNet. Accessed: Jan. 03, 2025. [Online]. Available: https://wordnet.princeton.edu/

[27] J. Wei and K. Zou, "EDA: easy data augmentation techniques for boosting performance on text classification tasks," in *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, K. Inui, J. Jiang, V. Ng, and X. Wan, Eds., Hong Kong, China: Association for Computational Linguistics, Nov. 2019, pp. 6382–6388. doi: 10.18653/v1/D19-1670.

[28] Y. Liu *et al.*, “RoBERTa: a robustly optimized bert pretraining approach,” *CoRR*, vol. abs/1907.11692, Jul. 2019, [Online]. Available: http://arxiv.org/abs/1907.11692
[29] Z. Lan, M. Chen, S. Goodman, K. Gimpel, P. Sharma, and R. Soricut, “Albert: a lite bert for self-supervised learning of language representations,” *8th Int. Conf. Learn. Represent. ICLR 2020*, pp. 1–17, 2020.
[30] G. Aguiar, B. Krawczyk, and A. Cano, "A survey on learning from imbalanced data streams: taxonomy, challenges, empirical study, and reproducible experimental framework," vol. 113, no. 7. Springer US, 2024. doi: 10.1007/s10994-023-06353-6.
[31] P. Vuttipittayamongkol, E. Elyan, and A. Petrovski, “On the class overlap problem in imbalanced data classification,” *Knowledge-Based Syst.*, vol. 212, p. 106631, 2021, doi: 10.1016/j.knosys.2020.106631.
[32] W. Chen, K. Yang, Z. Yu, Y. Shi, and C. L. P. Chen, "A survey on imbalanced learning: latest research, applications and future directions," vol. 57, no. 6. 2024. doi: 10.1007/s10462-024-10759-6.
[33] Y. Xu and R. Goodacre, “On splitting training and validation set: a comparative study of cross-validation, bootstrap and systematic sampling for estimating the generalization performance of supervised learning,” *J. Anal. Test.*, vol. 2, no. 3, pp. 249–262, 2018, doi: 10.1007/s41664-018-0068-2.
[34] B. Wang, G. Wang, J. Huang, J. You, J. Leskovec, and C. C. J. Kuo, “Inductive learning on commonsense knowledge graph completion,” *Proc. Int. Jt. Conf. Neural Networks*, vol. 2021-July, 2021, doi: 10.1109/IJCNN52387.2021.9534355.