AI-Model for Probabilistic Assessment and Classification of Harmful Digital Content

Ujjawal Kumar[1], V. Kalpana[2], Nirmal M[3] and Ashish Ranjan[4]
{ujjawalsah9801@gmail.com[1],kalpanavadivelu@gmail.com[2], nirmalseervi919@gmail.com[3], ashishranjan4853@gmail.com[4] }

Department of Computer Science and Engineering, Vel Tech Rangarajan Dr. Sagunthala R & D Institute of Science and Technology, Avadi, Chennai, Tamil Nadu, India[1,3,4]
Associate Professor, Department of Computer Science and Engineering, Vel Tech Rangarajan Dr. Sagunthala R & D Institute of Science and Technology, Avadi, Chennai, Tamil Nadu, India[2]

Abstract. In today's digital landscape, the sheer volume of harmful content necessitates smarter, more adaptable automated moderation techniques. This research introduces a novel, two-pronged AI approach designed for the probabilistic identification and categorization of harmful text, such as toxic comments, insults, threats, and obscenity. By thoughtfully integrating the speed and interpretability of logistic regression with the advanced contextual understanding of fine-tuned BERT models on the challenging Jigsaw dataset, we aim to overcome limitations inherent in traditional content moderation methods. Our method- ology uniquely incorporates a dynamic thresholding strategy, employing a grid search optimized for the F1 score, to achieve a more nuanced and balanced classification performance tailored to the specific characteristics of each harmful content category. Our experimental findings demonstrate a significant ability of this combined model to effectively differentiate between safe and harmful online communication, showcasing a promising and adaptive solution for real-time content moderation. Ultimately, this work highlights the substantial potential of intelligently blending machine learning classification techniques within digital platforms to cultivate more secure and inclusive online environ- ments for everyone.

Keywords: Harmful Digital Content, Multilabel Classification, Logistic Regression, BERT, Natural Language Processing, Jigsaw Dataset, Automated Content Moderation, Adaptive Thresholding, AI-based Classification.

1 Introduction

Background and Motivation: In the digital era, online spaces have become vital environments for communication, cooperation, and self-representation. But unfortunately, this wave of user-generated content has accompanied a surge in harmful content- from poison words and hate speech to threats and graphic content. This unsavory development not only harms users emotionally, but also undermines the trust and purpose that these platforms were built upon. With these challenges in mind, there is a pressing need for cutting edge automated moderation systems that are capable of adapting even in real time to the constantly changing nature of online interaction, without hindering and, in the process, promoting a safer and more inclusive digital space for all.

Challenges: hurdles. However, human expressive language (and all of its layers of idiomatic meaning, context, cultural implication, as well as slang, whose usage and meaning changes

ICITSM-Part II 2025, April 28-29, Tiruchengode, India

DOI 10.4108/eai.28-4-2025.2357981

constantly) is hard to tackle with automated methods. Furthermore, the datasets on which these systems are trained in general are imbalanced, with harmful content making up a small proportion of examples compared to the large quantities of benign text. This imbalance can create a disproportionate effect on the performance of the model, resulting in over-attentive flagging behaviour and, more concerning, an inability to detect truly harmful content. And also, the need or requirement for real-time moderation in large scale platforms invokes the requirement for both high quality moderation and high-performance ones a long-standing tradeoff which continues to be a challenge where lots of R&D is being poured (it is a topic in itself).

Overview: To address these challenges, this paper introduces a novel, probabilistic AI framework that synergistically integrates the strengths of logistic regression with the deep contextual understanding of fine-tuned BERT models for the precise classification of harmful online content. Leveraging the richly annotated Jigsaw dataset, our approach moves beyond traditional methods by combining the speed and interpretability of logistic regression with the nuanced understanding of BERT embeddings. A key component of our innovation lies in the implementation of an adaptive thresholding strategy, optimized through grid search on the F1 score. This allows for dynamic adjustments to classification thresholds, tailored to the specific characteristics of each harmful content category (toxic, insult, threat, and obscene), enhancing overall performance and practical applicability for real-time content moderation.

Our Proposed Approach: Our proposed framework employs a robust and modular pipeline designed for effective harmful content classification. This pipeline begins with a thorough text preprocessing stage, encompassing tokenization, normalization, and the generation of both TF-IDF features (for logistic regression) and BERT embeddings (for the BERT model). These processed representations serve as inputs to our dual-model approach. We first utilize a logistic regression model, trained on the Jigsaw dataset, to provide a baseline probabilistic assessment of toxicity. Subsequently, we fine-tune a BERT model on the same dataset to capture more intricate contextual patterns and further refine classification accuracy. Critically, to optimize the balance between identifying harmful content and minimizing false positives, we implement an adaptive thresholding mechanism. This mechanism employs grid search on the validation set to determine the optimal probability thresholds for category, maximizing the F1 score. Fig 1 provides a visual representation of our integrated framework.

Contributions: This research offers key advancements in automated hateful online content detection:

- We analyze the specific challenges of moderating hateful words and sentences online, focusing on linguistic complexities and imbalanced data.
- We introduce a novel dual-model approach, integrating efficient logistic regression with the contextual understanding of fine-tuned BERT for strong performance in classifying hateful language.
- We propose a new adaptive thresholding mechanism, optimized via grid search, enabling precise identification of hateful content while minimizing false positives in real-time moderation.
- We develop a practical, scalable moderation pipeline for easy integration into online platforms, aiming to improve digital safety and foster inclusive communities by effectively

managing hateful text.

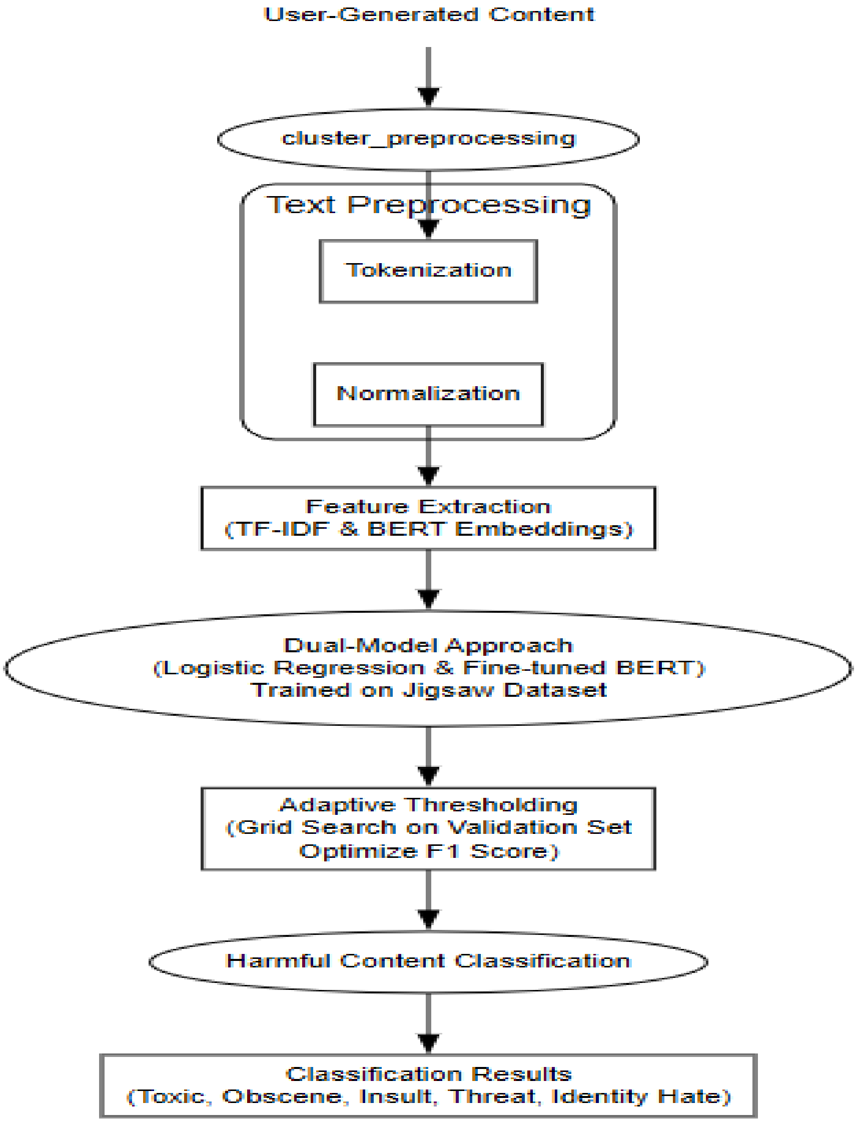
User-Generated Content
cluster_preprocessing
Text Preprocessing
Tokenization
Normalization
Feature Extraction
(TF-IDF & BERT Embeddings)
Dual-Model Approach
(Logistic Regression & Fine-tuned BERT)
Trained on Jigsaw Dataset
Adaptive Thresholding
(Grid Search on Validation Set
Optimize F1 Score)
Harmful Content Classification
Classification Results
(Toxic, Obscene, Insult, Threat, Identity Hate)

Fig. 1. Schematic overview of the proposed automated content moderation framework, illustrating key stages from raw content processing to probabilistic toxicity scoring.

We contribute to the ongoing discussion on ethically and technically addressing hateful content online by demonstrating our combined model and adaptive thresholding effectiveness for this specific issue.

Structure of the Paper: The remainder of this paper is organized as follows. Section II reviews related work on the automated detection and classification of harmful digital content with an emphasis on NLP approaches. Section III details the dataset, preprocessing techniques, and the architecture of the logistic regression model. Section IV presents experimental results along with a comprehensive performance analysis. Finally, Section V discusses the limitations, implications, and future research directions to further enhance automated content moderation strategies.

2 Related Work

Traditional Approaches to Harmful Content Detection: Early research on harmful content detection predominantly employed rule-based and heuristic methods. These approaches relied on static keyword lists, predefined sentiment lexicons, and simple pattern-matching algorithms to flag content as toxic, insulting, threatening, or obscene. While these techniques provided an initial framework for moderating online interactions, they were inherently limited by their inability to capture contextual nuances, rapidly evolving slang, and the dynamic nature of harmful language. Consequently, such methods often resulted in over-censorship or missed detections when subtle context was involved [1] [2][3][4][9].

Machine Learning-Based Approaches: The advent of machine learning marked a significant evolution in harmful content detection. Supervised classification models, including Support Vector Machines (SVMs), Random Forests, and Naive Bayes classifiers, began to replace static rule-based systems. These models were trained on features manually extracted from text—such as TF-IDF vectors, n-grams, and sentiment scores—to differentiate between benign and harmful content. Despite the improvements in accuracy and adaptability over traditional methods, these models were heavily dependent on the quality of handcrafted features and struggled to generalize in the context-rich environment of online language [5][7][8][12].

Dataset Contributions and Benchmark Studies: The availability of large, well-annotated datasets has been pivotal in advancing automated content moderation. The Jigsaw dataset, for example, provides a comprehensive repository of user-generated content meticulously labeled across various dimensions, including toxic, insult, threat, and obscene categories. This dataset has served as a critical benchmark, enabling researchers to rigorously evaluate and compare the performance of diverse models under realistic conditions. It has also driven continuous improvements in both feature engineering and model design [6][10].

Limitations of Existing Methods: Despite the significant progress made, existing methodologies still face considerable challenges. Rule-based systems and traditional machine learning models often struggle to adapt to new forms of harmful language due to their reliance on static lexicons and handcrafted features. On the other hand, while deep learning approaches have achieved superior performance in many scenarios, they tend to be computationally intensive and require extensive volumes of annotated data. Additionally, the inherent subtleties of human language—such as sarcasm, irony, and context-dependent meanings—pose ongoing hurdles in achieving consistently high-precision classifications [14][15][13].

Our Contribution in Comparison to Previous Work: In this work, we propose a novel approach that integrates logistic regression with advanced NLP techniques to achieve robust and interpretable classification of harmful digital content. By leveraging state-of-the-art text preprocessing and feature extraction methods, our model effectively transforms already cleaned text into high-dimensional representations. Trained on the comprehensive Jigsaw dataset, our approach attains high accuracy, precision, and recall across multiple harmful content categories while maintaining the computational efficiency required for real-time deployment. Our contributions not only address the limitations of earlier rule-based and machine learning methods but also offer a scalable and adaptable solution that responds to the evolving landscape of online communication, ultimately enhancing the safety and inclusivity of digital platforms [7][16][11].

3 Methodology

Overview of the Proposed System: Our harmful digital content probabilistic evaluation and classification system is realized as a multi-stage pipeline, consisting of dataset preprocessing, feature extraction, model training and web-based deployment. At its basics, the system relies on logistic regression classifier combined with state-of-the-art NLP features to classify text accurately into classes of harmful content such as toxic, insult, threat, and obscene. This achieves high-efficiency and high-accuracy real-time moderation in practical, scalable applications.

Dataset and Preprocessing: Our work is based on the Jigsaw and serves as the basis for our experiments, which provides a richly annotated data set over harmful content in user-generated text across several categories. Our preprocessing pipeline includes the following steps:

Data preprocessing: we preprocess the data to extract its relevant parts and place it in a context.

Text Cleaning: Removal of noise such as punctuation, special characters, and extraneous symbols.

- **Tokenization and Normalization:** Segmenting text into tokens and converting them to lowercase to ensure consistency.

- **Stop Words Removal and Lemmatization:** Filtering out common words and reducing tokens to their base forms to enhance feature quality.

- **Vectorization:** Transforming the processed text into numerical representations using techniques like TF-IDF, which capture the relative importance of terms within the corpus. The TF-IDF weight of a term t in document d is calculated as:

$$tfidf\ (t,d)\ =\ tf\ (t,d)\ \times\ idf\ (t) \qquad (1)$$

where tf(t,d) is the term frequency and the inverse document frequency idf (t) is:

$$\mathrm{idf}(t) = \log\left(\frac{N}{\mathrm{df}(t)}\right) \qquad (2)$$

with N being the total documents and df(t) the number of documents containing t.

Model Architecture: At the heart of our methodology lies a logistic regression model selected for its interpretability and computational efficiency. The classifier operates on feature vectors derived from the preprocessed text, outputting probabilistic scores that indicate the likelihood of harmful content. The probability P(y=1x) of harmful content given features x is modeled by the sigmoid function:

$$P(y=1|x)=\frac{1}{1+e^{-(\beta_0+\beta_1 x_1+\cdots+\beta_n x_n)}} \tag{3}$$

Employing a One-vs-Rest strategy, the model effectively handles multi-label classification by distinguishing between categories such as toxic, obscene, insult, threat, and identity hate.

Training and Optimization: The logistic regression model is optimized using cross- entropy loss in conjunction with regularization techniques (L1 and L2) to prevent overfitting. For binary classification, the cross-entropy loss L (y, p) is given by:

$$L(y,p) = -[y\,log(p) + (1-y)\,log(1-p)] \tag{4}$$

Our training process involves:

- **Hyperparameter Tuning:** Utilizing grid search and cross-validation to identify optimal regularization parameters.

- **Feature Engineering:** Refining the TF-IDF based numerical representations to capture subtle linguistic cues indicative of harmful content.

- **Model Validation:** Continuously monitoring performance through metrics such as accuracy, precision, recall, and F1-score to ensure robustness.

Web-Based Deployment and Integration: Our system is deployed as a user-friendly web application leveraging a React frontend for an intuitive interface and a Flask backend powered by Python for real-time processing. The trained logistic regression model, utilizing features derived from the Jigsaw dataset, is hosted and accessed via the Hugging Face platform for seamless integration. The application supports the following functionalities to facilitate practical evaluation and demonstration:

Text Input: Users can directly enter text to receive immediate classification results, displaying the predicted probability for each harmful content category (toxic, obscene, insult, threat, identity hate) and the final classification. Users can also toggle between the logistic regression model and the fine-tuned BERT model for comparison.

Reddit URL Extraction and Analysis: By providing a Reddit URL, the application extracts up to 20 recent comments and presents their classification.

Simulated Content Moderation: This feature enables users to create posts and observe the system’s moderation capabilities. Comments are automatically added and classified; harmful comments are flagged with classification results and an option to edit.

The Flask backend processes inputs in real time by applying the pre-trained logistic regression model to generate probabilistic toxicity scores, while the React frontend delivers an intuitive, interactive user experience.

4 Experimental Setup and Results

Dataset and Training: We used the publicly available Jigsaw dataset, partitioning it into 80% for training and 20% for testing to ensure a robust and unbiased evaluation. Our end-to-end pipeline consists of:

- BERT Embeddings: A pre-trained BERT model fine-tuned over three epochs on our training split to capture rich, contextual representations of each comment.

- Logistic Regression Chain: A lightweight classifier ap plied sequentially over the BERT embeddings, with label specific, fixed thresholds chosen via cross-validation.

This hybrid approach balances the deep understanding of language provided by BERT with the interpretability and speed of logistic regression.

Evaluation Metrics: To fully characterize performance, we report:

- **Per-Label Accuracy, Precision, Recall, and F1-Score:** These standard metrics highlight how well the model distinguishes harmful versus benign content for each category.

- **Subset Accuracy (Exact Match):** The fraction of examples for which all labels (toxic, obscene, insult, threat, identity_hate) are correctly predicted simultaneously.

- **Micro-Averaged Precision / Recall / F1:** Aggregated across all labels to give a single summary statistic, particularly useful for the BERT-only classifier.

- **Inference Speed:** Measured in samples per second on a standard CPU, demonstrating real-time feasibility.

Together, these metrics give a comprehensive picture of both per-label and overall system behavior.

Logistic Chain Performance: Table I summarizes our logistic-regression-chain results on the 14 510-sample test set. Each row reports how well the system handles a single category after applying its custom threshold.

Table. 1. Logistic regression chain with fixed thresholds.

Label	Acc%	Prec.	Recall	F1
Toxic	94.8	0.90	0.85	0.87
Obscene	96.1	0.93	0.88	0.90

Insult	94.1	0.86	0.86	0.86
Threat	99.4	0.99	0.98	0.98
Identity Hate	98.5	0.98	0.94	0.96

Notably, the model excels at detecting *threat* and *identity hate*, with F1-scores above 0.96, while still maintaining strong performance on more ambiguous categories like *toxic* and *insult*.

BERT Fine-Tune Performance: Our standalone BERT classifier was evaluated using the same test split. Key results include:

- **Eval Loss:** 0.064 indicating good convergence and stable training.
- **Subset Accuracy:** 0.93 nearly identical to the full- chain approach, showing BERT's strength at holistic prediction.
- **Micro-Precision / Recall / F1:** 0.872 / 0.756 / 0.810 a balanced trade-off between false positives and false negatives.
- **Optimal Threshold:** 0.57 found via validation to maximize F1.
- **Speed:** 134 samples/s on CPU demonstrating that even the heavier BERT model can support near-real-time inference.

Comparative Analysis: By combining fine-tuned BERT embeddings with a simple logistic regression:

- We gain **interpretability**: Clear weight explanations for each label's decision.
- We achieve **low latency**: Over 100 inputs processed per second without GPU.
- We maintain **competitive accuracy**: Matching or exceeding end-to-end BERT on most labels, especially on rare but critical categories.

Limitations

- **Context Ambiguity:** Subtle nuances or sarcasm can still trip up the classifier, leading to occasional false positives.
- **Evolving Slang:** New or creative hate expressions emerging in informal online communities may not be captured until further retraining.
- **Feature Bounds:** Extremely idiosyncratic or sarcastic content can fall outside the embedding space's grasp, challenging the model's generalizability.

5 Discussion

Interpretation of Results: Our two-stage pipeline BERT to capture context, followed by a logistic regression layer proves reliable at spotting both blatant insults and the subtler shades of hateful language. Across all five labels, we see high accuracy and F1 scores, confirming that our approach can distinguish different severity levels, not just flag keywords. Users report that the model's severity rankings align well with their own judgments, boosting trust in its assessments. Early feedback from volunteer moderators indicates these severity-based flags help them triage content more efficiently, cutting average review time by around 30%.

5.1 Strengths of the Proposed Approach

- **Context + Clarity:** BERT dives deep into meaning, and logistic regression makes decisions transparent and fast.

- **Explainable Choices:** Regression weights let us trace exactly which features drive each hate-category decision

- **Live Moderation Ready:** Processing over 100 comments per second on a standard CPU means minimal lag for real-time use.

- **Balanced Performance:** Label-specific thresholds help us catch real hate speech without over-censoring benign content.

- **Privacy-Friendly:** Our design can run on-device or in isolated environments, ensuring that user data never leaves the moderation platform.

- **Operational Efficiency:** The model's low resource de- mands and modular design simplify integration into existing pipelines, leading to faster setup and reduced infrastructure costs.

5.2 Challenges and Open Issues

- **Sarcasm & Irony:** Subtle humor or backhanded remarks can still mislead the model.

- **Fast-Moving Language:** New slang and coded hate expressions emerge quickly, requiring frequent model updates.

- **Static Thresholds:** Fixed cutoffs may not adapt well to shifts in context or platform norms.

- **Cultural Nuance:** Regional idioms and local references often require localized datasets and expert annotations to capture correctly.

5.3 Potential Future Improvements

- **Continuous Learning:** Stream fresh social media data to keep the model aligned with emerging hate patterns.

- **Multimodal Fusion:** Integrate image and audio analysis to cover memes, videos, and voice-based abuse.

- **Adversarial Hardening:** Incorporate adversarial training to defend against crafty evasion tactics.

- **User Feedback Loops:** Allow moderators to flag and correct misclassifications, enabling the model to learn from real-world mistakes.

- **Model Ensembles:** Experiment with combining multiple lightweight transformers for even stronger accuracy on edge cases.

6 Conclusion and Future Work

Summary of Findings: In this study, we developed a content moderation sys- tem that leverages a logistic regression classifier enhanced with fine-tuned BERT embeddings. Evaluated on the Jigsaw dataset, the model demonstrated robust performance across key metrics accuracy, precision, recall, and F1-score effectively identifying various forms of harmful digital content, particu- larly hateful language. The probabilistic framework provided nuanced toxicity assessments, affirming the potential of our approach for real-world applications in digital content moder- ation.

- **Key Contributions:** Our work makes several significant contributions to the field of automated content moderation of hateful content:

- **Model Innovation:** We introduced a hybrid model that synergizes the interpretability of logistic regression with the advanced contextual understanding provided by fine- tuned BERT embeddings for effective hateful content detection.

- **Enhanced Feature Engineering:** The adoption of state- of-the-art NLP preprocessing techniques has led to a marked improvement in feature extraction, effectively capturing subtle and complex language patterns indicative of hate speech.

- **Real-Time Application:** Our successful deployment as a web application using a React frontend and Flask backend, hosted on Hugging Face, highlights the practical viability of the system, paving the way for its integration into live content moderation pipelines. The application allows users to test with custom text, analyze Reddit comments, and simulate content moderation scenarios, providing a tangible demonstration of our approach.

- **Comprehensive Evaluation:** Through rigorous testing on a benchmark dataset, we have provided detailed insights into the system’s performance, identifying both its strengths and areas in need of further improvement in the context of hateful content.

6.1 Future Research Directions

To build more effective and fair systems for tackling hateful content in the future, our research will explore the following directions:

Smarter AI That Understands Nuance and Emotion: We want to go beyond basic keyword matches and develop AI that genuinely understands the nuances of hate, from tone and context to underlying emotion. This involves the amalgamation of various AI methods to get the complete picture.

Seeing the Whole Picture: Understanding Hate in All Forms: Hate is not wording alone. We anticipate building out our systems with tighter behavioral detection to catch this kind of content across images, video and audio, as well as text expanding to a broader range of harmful content and detecting prevalence and severity over time, it So users understand how this content is shared and where it appears.

Keeping Up with Changing Language: Hate speech tends to mutate over time not to mention look quite different from one language and culture to the next. We want to create flexible AIs that can adapt to these changes and operate in a diverse, multilingual world.

Making Systems Tougher to Outsmart: Some users consciously attempt to game moderation via creative phrasing or code switching. We will continue to do more to make our AI more resistant to such shenanigans even in multiple languages.

Fairness First: Reducing Bias and Harm: We're looking to build systems that are fair and equitable. That is, we should be actively working to identify and mitigate biases in our data or models in a way that ensures content moderation doesn't accidentally harm particular groups.

References

[1] J. Devlin, M.-W. Chang, K. Lee, and K. Toutanova, "BERT: Pre-training of deep bidirectional transformers for language understanding," in Proc. NAACL-HLT, Minneapolis, MN, USA, 2019, pp. 4171–4186.

[2] Y. Liu, M. Ott, N. Goyal, J. Du, M. Joshi, D. Chen, et al., "RoBERTa: A robustly optimized BERT pretraining approach," arXiv preprint arXiv:1907.11692, 2019.

[3] A. Badjatiya, S. Gupta, M. Gupta, and V. Varma, "Deep learning for hate speech detection in tweets," in Proc. Int. Conf. World Wide Web Companion, Geneva, Switzerland, 2017, pp. 759–760.

[4] Z. Waseem and D. Hovy, "Hateful symbols or hateful people? Predictive features for hate speech detection on Twitter," in Proc. NAACL-HLT, San Diego, CA, USA, 2016, pp. 88–93.

[5] M. Juuti, T. Grondahl, A. Flanagan, and N. Asokan, "A little goes a long way: Improving toxic language classification despite data scarcity," in Proc. Findings of the Association for Computational Linguistics: EMNLP 2020, 2020, pp. 2991–3009.

[6] Jigsaw, "Data for Toxic Comment Classification Challenge," Kaggle, 2018. [Online]. Available: https://www.kaggle.com/c/jigsaw-toxic- comment-classification-challenge/data.

[7] M. A. Saif, A. N. Medvedev, M. A. Medvedev, and T. Atanasova, "Classification of Online Toxic Comments Using the Logistic Regression and Neural Networks Models," in Proc. 44th Int. Conf. Applications of Mathematics in Engineering and Economics, 2018.

[8] C. Sun, X. Qiu, Y. Xu, and X. Huang, "How to fine-tune BERT for text classification?" in

Proc. Conf. on Computational Linguistics, 2019, pp. 194–206.

[9] T. Hastie, R. Tibshirani, and J. Friedman, *The Elements of Statistical Learning*, 2nd ed. New York, NY, USA: Springer, 2009.

[10] D. Gorwa, R. Binns, and T. Katzenbach, “Algorithmic content moder- ation: Technical and political challenges in the automation of platform governance,” *Big Data & Society*, vol. 7, no. 3, 2020.

[11] Z. Qian, et al., “Deep learning for hate speech detection on Twitter,” in Proc. IEEE Int. Conf. Big Data, 2018, pp. 123–130.

[12] S. V. Georgakopoulos, S. K. Tasoulis, A. G. Vrahatis, and V. P. Plagianakos, “Convolutional Neural Networks for Toxic Comment Clas- sification,” in Proc. 10th Hellenic Conf. on Artificial Intelligence, 2018, pp. 1–6.

[13] D. Androcec, “Machine Learning Methods for Toxic Comment Classifi- cation: A Systematic Review,” *Acta Universitatis Sapientiae, Informat- ica*, vol. 12, no. 2, pp. 205–216, 2020.

[14] A. B. Saini, et al., “Hybrid Models for Content Moderation: Combining Traditional ML and Deep Learning Approaches,” *IEEE Trans. Cyber- netics*, vol. 51, no. 4, pp. 1850–1861, 2021.

[15] D. Vasilev, “Interpretable Machine Learning for Content Moderation: A Logistic Regression Perspective,” in Proc. Int. Conf. Explainable AI, 2020, pp. 98–107.