A Predictive Model for Anticipated Hate and Speech Violence in Social Media: Large Language Model Approach

Gimhani Samindika Dissanayake[1], Sudesh Jayathunge Bandara[1], Hemalika T.K. Abeysundara[2]

[1]Postgraduate Institute of Science, University of Peradeniya

[2]Department of Statistics & Computer Science, University of Peradeniya

DOI: https://doi.org/10.51244/IJRSI.2025.12020028

Received: 23 January 2025; Accepted: 31 January 2025; Published: 05 March 2025

ABSTRACT

This study explores the application of Large Language Models (LLMs), specifically a BERT-based architecture, coupled with predictive analytics for proactive hate speech and violence identification and forecasting in social media comment moderation. The research addresses the critical need to shift from reactive to proactive content moderation strategies in order to enhance digital safety and foster inclusivity. Using a dataset of 44,000 public comments from Facebook that are unfiltered and generalized, the methodology includes data collection through the Facebook Graph API, preprocessing phases, and model training using a balanced dataset to improve detection accuracy. The study highlights the importance of ethical deployment in artificial intelligence, noting the role of predictive analytics in identifying patterns and signals that indicate the existence of harmful content before its widespread circulation. From this, the results show that a balanced dataset helped the model to achieve strong performance metrics: accuracy of 82.63%, precision of 82.5%, recall of 82.88%, and F1 score of 82.69%. Results like this have shown the promise that advanced AI technologies hold when integrated into content moderation to successfully handle and pre-empt online hate speech and violence. This research helps to contribute to the larger discussion on ethical AI, responsible digital citizenship, and safer online communities through the promotion of proactive moderation systems.

Keywords: Ethical AI, Hate Speech Detection, Large Language Models, Predictive Analytics, Social Media Moderation

INTRODUCTION

Social media sites like Facebook have a central role in enabling the world to communicate and play an immense role in the exchange of ideas, individual stories, and activism to an unprecedented degree. These sites increase the visibility of silenced voices, enhance inclusivity, and create a globally connected community [1]. At the same time, the very mechanisms that enable cohesion and individual expression also expose users to damaging content, including violence and hate speech, thereby posing considerable dangers to both social cohesion and individual well-being. Hate speech, defined as communication that insults, demeans, or provokes hostility toward individuals based on characteristics such as race, religion, gender, or ethnicity, has become increasingly common on social media sites. This is added to by the fact that information spreads so quickly, allowing for isolated hate incidents to escalate into societal issues of great concern [2]. The harm in this type of content is not limited to individual circumstances, lowering the level of public discourse and undermining the basic values of respect and understanding that underpin peaceful existence [3].

While significant advances have been achieved with content moderation methods, most available methodologies are primarily reactive—only after offensive content has propagated to users is it typically addressed. This strategy falls short in confronting the enormous scale and sped-up pace of interaction on the Internet [4]. Furthermore, perpetrators of hate speech frequently make use of evasive strategies such as coded messages or intentionally misspelled invective for the purpose of evading conventional moderation systems [5].

Recent advances in state-of-the-art artificial intelligence (AI) and machine learning techniques provide encouraging solutions [6]. Large Language Models, such as BERT (Bidirectional Encoder Representations from Transformers), have demonstrated impressive proficiency in natural language comprehension and context understanding, and hence are particularly apt for hate speech detection task [7]. Predictive analytics, that predicts harmful content before it gets published, is the latest development in content moderation from reactive to proactive on the heels of growing demand for safety and inclusivity in online spaces [8].

This paper utilizes a test bed of 44,000 publicly available Facebook comments to find out how effectively the combination of LLMs and predictive analytics is able to identify and foresee hate speech. In this study, we will use a BERT-based model to explore the viability of proactive content moderation methods, balancing online community safety with ethical concerns of privacy, algorithmic bias, and freedom of speech. The study has as its aim a discussion of LLMs' ability to forecast, their performance compared to the conventional methods, and an investigation of the general implications of AI-based content moderation systems [9], [10]. The research aims to unify the content moderation community by illustrating the potential of new technologies to make the online environments more inclusive and secure [11].

LITERATURE REVIEW

This content-moderation strategy shows movement from responsive manual approaches to automated solutions by using machine learning and NLP technologies. In its earliest terms, moderation relied on users' reports and manual judgment, which soon became outdated against the explosive growth of digital contents [12] , [1]. While the automated alternative brought improvement in terms of scalability and speed, the technology struggled with grasping linguistic nuances. Early ML models were challenged by the ability to distinguish contextually ambiguous terms, which often resulted in errors in moderation decisions [13], [14]. The transformer-based models, like BERT and GPT, have revolutionized content moderation by allowing contextual understanding and intent analysis [15] , [16]. Such models are able to handle the problems litigated by earlier approaches, including linguistic complexity, irony, and sarcasm. However, a stand-alone AI system, no matter how complex, is not free from biases and misjudgements. Research highlights the need for the development of hybrid approaches that marry the efficiency of artificial intelligence with human oversight, ensuring accuracy and reducing the chance of unintended consequences [17]. Human moderators provide critical insight in making complex judgments about creating feedback loops for constant improvement.

Hate speech detection has gone a long way since the advent of Large Language Models (LLMs). In particular, models such as BERT have been able to exemplify the ability to capture complex patterns, contextual expressions, and intent in text data; thus, especially good at recognizing both implicit and explicit forms of hate speech [5] , [18]. For instance, [17] demonstrated the use of automated detection approaches and uncovered the potential for context-aware algorithms. Attempts at multi-modal approaches by combining text analysis with visual and metadata have increased precision in detection [8]. This multi-modal approach elevates the possibility of spotting user intention because it tends to accurately classify. Reduction of the bias found in training datasets, which will contribute to societal prejudices remains challenging. [8] Also highlighted the racial disparity in identifying hate speech, recommending iterating the improvements continuously to assure justice and equality.

Predictive analytics has caused a paradigm shift in content moderation, from reactive to proactive. Using historical data, predictive models can forecast when and where surges of harmful content are likely to occur, enabling timely interventions [19]. Sophisticated machine-learning techniques, coupled with behavioural and temporal analytics, have been shown to identify patterns that can lead to the prediction of future events of hate speech or violent content [7]. Network analysis, as underlined in [9], brings into the spotlight the importance of social network configurations in understanding the diffusion of harmful content. Temporal trends and user behaviour become key indicators on which predictive models can be based to take proactive measures against risks. While such progress has been made, predictive analytics brings about ethical considerations in respect to user privacy and algorithmic transparency. [10] Called for rights and freedoms protection when implementing predictive moderation.

AI-driven content moderation faces several limitations, including data biases, cultural diversity, and dynamic evolution of language. If the machine learning models are trained on biased datasets, then the model may amplify these biases, as revealed by [20]. The complexity of human language in slang, idioms, and variations across cultures adds up to the challenges in developing universally accurate systems [21]. Additionally, there are critical ethical considerations concerning AI in content moderation. The risk of overreach, in which algorithms suppress authentic dialogue, requires a balancing act to be struck between safety and the principle of freedom of expression [1]. As [12] indicates, transparency and accountability are fundamental to gaining trust in artificial intelligence systems.

While LLMs and predictive analytics have contributed much to hate speech detection, there lies a gap in the integration of these technologies for proactive content moderation. Much of the current research is focused on detecting harmful content after it has been published, thereby leaving a wide gap in pre-emptive solutions. This paper fills this gap by employing a BERT-based model that can predict hate speech and violence, hence opening ways for safer and more inclusive digital platforms.

METHODOLOGY

The rapid growth of harmful content on social media platforms presents a significant challenge for maintaining safe and inclusive online environments. Effective interventions require not only the identification of instances of hate speech and violence but also the capability to anticipate their rise in order to prevent further dissemination. This study adopts a comprehensive approach, applying advanced machine learning techniques and Large Language Models (LLMs) for the identification and prediction of harmful content. The methodology is organized into several fundamental phases, including data gathering, preprocessing, model selection and training, evaluation, and integration with predictive analytics to provide a general framework for dealing with the challenges [22].

Data Collection

The foundation of this research is based on the gathering of a high-quality dataset. A total of 44,000 comments were collected from publicly available Facebook posts using the Facebook Graph API. This dataset was carefully collected to include a great variety of interactions, thus reflecting the linguistic and contextual complexities inherent in the discourse of social media. Engagement metrics such as likes, shares, and replies were considered in order to identify discussions that are likely to have a significant social impact. Ethical guidelines were stringently adhered to, ensuring user anonymity and conformity with data protection legislation.

Data Preprocessing

To improve the dataset for analysis, an extensive preprocessing pipeline was developed. This contained:

Noise Removal: The process of removing superfluous data, including URLs, emoji, and special symbols.

Normalization**:** All text data is normalized to lowercase to be aligned with the case-insensitive nature of the BERT variant used.

Labeling and Tokenization: The Gemini tool has been used in labeling and tokenization of the dataset. Gemini worked fine to classify the dataset into relevant classes hate and non-hate and preprocess the text for input into the model by breaking it down into smaller, manageable components.

Data Splitting

The dataset was divided into two subsets for comparative analysis:

Unbalanced Dataset: The original class distribution was kept with the larger proportion of non-hate comments (40819 for non-hate and 3181 for hate).

Balanced Dataset: This is achieved by performing down-sampling on the non-hate comments such that the number of hate and non-hate instances become equal.

Both datasets were used to train the models and performance metrics were evaluated to determine the most appropriate dataset. The dataset that showed improved precision was selected for subsequent analyses.

Model Selection and Training

We chose a cased BERT model because it has an impressive natural language processing ability, especially when it comes to understanding context, sarcasm, or implicit meaning [23]. The process of training included:

Fine-Tuning: The pre-trained BERT model is fine-tuned for this specific task of classifying social media comments as hate speech or non-hate.

Cross-Validation: The dataset is split into training and testing sets to achieve the generalizability and robustness of the model.

Comparing performances across models trained on balanced and unbalanced datasets revealed that the balanced dataset showed superior precision in hate speech detection.

Evaluation Metrics

The effectiveness of the model was evaluated based on the following metrics:

Accuracy: It is the overall proportion of correct predictions.

Precision: The proportion of true positives to all the predicted positive cases, indicating that the accuracy of violence and hate speech predictions.

Recall: The ability of the model to find all relevant instances of hate speech and violence.

F1 Score: A harmonic mean of precision and recall, providing a balanced measure of performance.

Specificity: measures the true negative rate, i.e., how good the model is at identifying non-harmful content.

Integration with Predictive Analytics

A distinctive aspect of this study is the integration of predictive analytics with the BERT model. Predictive analytics simulated potential future scenarios by analyzing emerging patterns in the data, enabling the model to anticipate hate speech and violence before dissemination. These simulations tested the model’s adaptability to varying data sizes and evolving linguistic patterns, demonstrating its viability for real-time content moderation systems. The approach is proactive, really a huge advance in the methods of content moderation from reactive to predictive and preventive ones.

RESULTS AND DISCUSSION

The uncased BERT model is evaluated on two different datasets: the first one is unbalanced and close to the real-world distribution, where there is a lower prevalence of hate speech compared to neutral comments; the second dataset is balanced, where hate and non-hate comments are uniformly represented. The double assessment of the model provided more insights into its strength and adaptability in dealing with changing conditions.

Results from the Unbalanced Dataset

The unbalanced dataset evaluated the model in a scenario where the dominance of hate speech was low, similar to real-world distributions. Table I summarizes the performance metrics:

Table 1: Performance metrics for unbalanced data

Metric	Value
Accuracy	92.18%

Precision (Hate)	58.97%
Recall (Hate)	43.09%
F1 Score (Hate)	49.49%

The confusion matrix provided further clarity on classification outcomes:

True Positives (Class 1): 287

False Negatives (Class 1 Misclassified as Class 0): 379

True Negatives (Class 0): 6626

False Positives (Class 0 Misclassified as Class 1): 207

The model performed well in the identification of non-hate speech, achieving high accuracy scores overall (Class 0). However, the precision and recall of hate speech were low (Class 1), which is a common challenge in unbalanced datasets where minority classes are underrepresented. The disparity shows the need for strategies like dataset balancing or class-specific adjustments to improve minority class performance.

Results from the Balanced Dataset

With the balanced dataset, the model's performance enhanced significantly across metrics. Table II highlights the results:

Table 2: Performance metrics for balanced data

Metric	Value
Accuracy	82.63%
Precision (Hate)	82.50%
Recall (Hate)	82.88%
F1 Score (Hate)	82.69%

The confusion matrix for the balanced dataset:

True Positives (Class 1): 528

False Negatives (Class 1 Misclassified as Class 0): 109

True Negatives (Class 0): 524

False Positives (Class 0 Misclassified as Class 1): 112

The balancing of the dataset considerably improved the precision, recall, and F1 score in regard to hate speech classification. This balanced dataset encouraged equal performance among the two classes, eliminating the biases of unbalanced datasets and enabling the model to achieve better generalization on minority classes.

Comparative Analysis: Unbalanced vs. Balanced Dataset

A comparative analysis of the two datasets has shown that balancing improves hate speech detection capabilities immeasurably for the model. Indeed, although the unbalanced dataset yields higher general accuracy (92.18%), the balanced dataset outperforms in performance based on hate speech-specific metrics.

Hate Meter AnalysisThe Hate Meter provided insights into the model's confidence in predicting hate speech:

Unbalanced Dataset: Confidence scores ranged between [1.97, -2.29].

Balanced Dataset: Confidence scores ranged between [-2.82, 1.99].

These scores reflect that the predictions of the balanced dataset are more consistent because this model had less variation in confidence level. That is, the balanced datasets not only improve performance but also make a model's predictions more consistent.

The findings are that we need to balance datasets in order to make hate speech detection models perform optimally. Even though an unbalanced dataset is more reflective of real life, class imbalance limits what is possible for the model to do in hate speech detection. A balanced dataset remedies these problems by making performance more equalized between classes.

The Hate Meter breakdown illustrates how the balance of data can subtly influence the confidence level of the model. Balanced datasets are said to provide a solid basis for confident predictions. This has been substantiated by earlier researches that assert the need to use balanced and representative training datasets for augmenting model fairness as well as model performance.

Implications for Social Media Moderation

The outcomes have far-reaching implications for boosting online safety and user attitudes toward social media websites. A paradigm shift to proactive moderation, made possible by the predictive analytics capability of the model, would be an effective approach to containing online hate speech and violence dissemination. The improved balance between recall and precision on the balanced dataset shows that it require more robust and diverse data when we are training the AI models for content moderation. Furthermore, the BERT model can be further fine-tuned to the balanced dataset. This would indicate that the websites could be enriched with the AI tools trained on fairly balanced instances of harmful and harmless content. This could assist in shaping more respectful and an inclusive online community. This change is consistent with the general aim of making the virtual space more secure and emphasizes the need of utilizing the AI responsibly to guarantee ethical conduct online.

Limitations and Challenges

The findings provide us with some cause for optimism, but we also need to keep in mind the natural limitations and difficulties of the research. Specifically, BERT results are very sensitive to both the quantity and the nature and diversity of the information they are provided with. We can state that the decreased recall and precision on the unbalanced dataset mirror possible biases causing errors in classifying data, resulting in biased predictions by the model. First, it is challenging for the model to understand the complex nature of language and context in online forums. Secondly, there are ethical concerns when applying AI to content moderation. Its conclusions must be designed carefully in such a way that over-reliance on machine systems will not damage user privacy and free speech protection, as it may result in censorship or entrench prejudices.

CONCLUSION AND RECOMMENDATIONS

This research illustrates how the uncased BERT model can identify hate speech and violence on social media more accurately. This research addresses the pressing need to transition from responding to problems to averting them altogether, in an attempt to make online spaces safer and more welcoming. By using a dataset of 44,000 of unfiltered and generalized public Facebook comments, the method involves of collecting data through the Facebook Graph API, going through the preprocessing stages, and training a model based on a balanced dataset to guarantee more accurate detection. The model recorded higher precision, recall, and overall performance, especially for categories of small groups like hate speech, by utilizing balanced datasets and predictive analytics. These findings tell us that we need to move towards a more proactive content management approach. This entails halting and curbing the spread of objectionable content while addressing technical along with social concerns of online hate speech.

The study epitomizes the tremendous progress made; still, it exposes the prevailing limitations and ways to be improved. Future research should center on increasing the model's adaptability within different linguistic and cultural contexts in view of the international nature of social media conversations. Development of multilingual models that could understand context and nuance in multiple languages would significantly extend the applicability of AI-driven content moderation tools.

Finally, real-time implementation is one such area that requires exploration on a continuous basis. Predictive models being deployed in live environments would provide evidence of their performance under dynamic conditions, offering insights into latency, scalability, and practical integration with existing systems. Incorporating user feedback mechanisms will further refine these models, ensuring responsiveness to evolving linguistic trends and user expectations. Similarly, feedback loops could help reduce bias and enhance transparency in AI-driven decision-making.

Ethical considerations shall be the priority in future developments. In this line of thought, it will require equity, protection of privacy, and balance between freedom of expression and user protection that will enable the sustainability of trust in AI-driven systems. This may also give way to joint initiatives in the form of ethical guideposts and practices in using the aforementioned technologies.

In conclusion, this research demonstrates the possibility of combining LLMs with predictive analytics in an effort to create a safer, more inclusive online environment. Addressing the current limitations and seeking future advancements, AI-driven moderation tools can be of great help in fostering healthier digital interactions that align technological innovation with societal well-being.

REFERENCES

1. T. Gillespie, Custodians of the Internet: Platforms, Content Moderation, and the Hidden Decisions That Shape Social Media., Yale University Press, 2018.
2. Schmid, U. K. a. K{\"u}mpel, A. S. a. Rieger and Diana, “How social media users perceive different forms of online hate speech: A qualitative multi-method study,” New media & society, vol. 26, pp. 2614--2632, 2024.
3. M. Duggan, “Pew Research Center,” 11 July 2017. [Online]. Available: https://www.pewresearch.org/internet/2017/07/11/online-harassment-2017/#:~:text=A%20new%2C%20nationally%20representative%20Pew,these%20behaviors%20directed%20at%20others.. [Accessed 26 12 2024].
4. Gongane, V. U. a. Munot, M. V. a. Anuse and A. D, “Detection and moderation of detrimental content on social media platforms: current status and future directions,” Social Network Analysis and Mining, vol. 12, p. 129, 2022.
5. Nobata, C. a. Tetreault, J. a. Thomas, A. a. Mehdad, Y. a. Chang and Yi, “Abusive language detection in online user content,” in Proceedings of the 25th international conference on world wide web, 2016, pp. 145--153.
6. Devlin and Jacob, “Bert: Pre-training of deep bidirectional transformers for language understanding,” arXiv preprint arXiv:1810.04805, 2018.
7. Schmidt, A. a. Wiegand and Michael, “A survey on hate speech detection using natural language processing,” in Proceedings of the fifth international workshop on natural language processing for social media, 2017, pp. 1--10.
8. Sap, M. a. Card, D. a. Gabriel, S. a. Choi, Y. a. Smith and N. A, “The risk of racial bias in hate speech detection,” in Proceedings of the 57th annual meeting of the association for computational linguistics, 2019, pp. 1668--1678.
9. Ribeiro, M. H. a. Ottoni, R. a. West, R. a. Almeida, V. A. a. M. Jr and Wagner, “Auditing radicalization pathways on YouTube,” in Proceedings of the 2020 conference on fairness, accountability, and transparency, 2020, pp. 131--141.

10. Barocas, S. a. Selbst and A. D, “Big data's disparate impact,” Calif. L. Rev., vol. 104, p. 671, 2016.
11. Burnap, P. a. Williams and M. L, “Us and them: identifying cyber hate on Twitter across multiple protected characteristics,” EPJ Data science, vol. 5, pp. 1--15, 2016.
12. Roberts and S. T, Behind the screen, Yale University Press, 2019.
13. Fortuna, P. a. Nunes and Sergio, “A survey on automatic detection of hate speech in text,” ACM Computing Surveys (CSUR), vol. 51, pp. 1--30, 2018.
14. Waseem, Z. a. Hovy and Dirk, “Hateful symbols or hateful people? predictive features for hate speech detection on twitter,” in Proceedings of the NAACL student research workshop, 2016.
15. Devlin and Jacob, “Bert: Pre-training of deep bidirectional transformers for language understanding,” arXiv preprint arXiv:1810.04805, 2018.
16. S. Rickardsson, “A Hybrid Approach to Hate Speech Detection,” 2023.
17. Davidson, Thomas, D. Warmsley, M. Macy and I. Weber, “Automated hate speech,” in Proceedings of the international AAAI conference on web and social media, 2017.
18. Nikolov, A. a. Radivchev and Victor, “Offensive tweet classification with bert and ensembles,” in Proceedings of the 13th international workshop on semantic evaluation, 2019.
19. Arian, A. a. Yilmaz and Ozgur, “On one-stage recovery for ΣΔ quantized compressed sensing,” arXiv preprint arXiv:1911.07525, 2019.
20. L. a. L. Dixon, J. a. Sorensen, J. a. Thain, N. a. Vasserman and Lucy, “Measuring and mitigating unintended bias in text classification,” in Proceedings of the 2018 AAAI/ACM Conference on AI, Ethics, and Society, 2018.
21. Wiegand, M. a. Ruppenhofer, J. a. Schmidt, A. a. Greenberg and Clayton, “Inducing a lexicon of abusive words--a feature-based approach,” in Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, 2018.
22. Wirth, R. a. Hipp and Jochen, “CRISP-DM: Towards a standard process model for data mining,” in Proceedings of the 4th international conference on the practical applications of knowledge discovery and data mining, 2000, pp. 29--39.
23. Helal, N. A. a. Hassan, A. a. Badr, N. L. a. Afify and Y. M, “A contextual-based approach for sarcasm detection,” Scientific Reports, vol. 14, p. 15415, 2024.
24. Nobata, C. a. Tetreault, J. a. Thomas, A. a. Mehdad, Y. a. Chang and Yi, “Abusive language detection in online user content,” in Proceedings of the 25th international conference on world wide web, 2016.