

Bias And Ethical Risks In Automated Toxicity Detection Of Bodo Language Using Deep Learning Approaches

Pranchis Narzaree¹, Amkar Brahma², Prof. (Dr.) Manoj Kr. Deka³

¹Assistant Professor, Department of Computer Science, Kokrajhar University, Kokrajhar, Assam, India,

¹Research Scholar, Department of Computer Science and Technology, Bodoland University, Kokrajhar, India, pranchis.naz@gmail.com

²Research Scholar, Department of Computer Science and Technology, Bodoland University, Kokrajhar, India, amkarbrahma9773@gmail.com

³Professor, Department of Information Technology, Assam Skill University, Mangaldoi, Assam, India manoj.deka1@gmail.com

Abstract: The exponential growth of the digital communication on social media platforms has increased the spreading of the toxic language including hate speech, abusive expressions and cyber-harassment. Automated toxicity detection using Natural Language Processing (NLP) and deep learning has been a great resource to resolve the issues of moderating online content at remarkable standard. However, most available toxicity detection systems are created for high resource languages, especially English, and fail to take linguistic diversity, cultural nuance and ethical conformity in mind. This limit is especially important for the underprivileged indigenous languages like Bodo which is majorly spoken in Bodoland Territorial Council, Assam, North Eastern region of India and some parts of the neighbouring country like Bangladesh and Nepal. This study explores biases and ethical risks in automated toxicity detection of the Bodo text by the Deep learning approaches. A collection of 12,037 text samples in the Bodo language that were collected from social media platforms, Blogs, Bodo drama books and short stories that were manually annotated regarding toxic and non-toxic text categories by native speakers. We considered more than one deep learning architecture for e.g. Vanilla Recurrent Neural Network and its variants. Performance evaluation was measured using different metrics like accuracy, precision, recall, F1-score and bias-specific measures like false positive rate for identity related terms. While it has seen that Bi-Recurrent Neural Network managed to achieve superior performance (F1-score reached 0.975845), it is found that significant bias against multicultural identity references, culturally embedded metaphors and emotionally expressed language common in Bodo discourse were identified. These results do raise serious ethical questions about unfair security review, cultural marginalization and the hostility of annotator and algorithmic prejudices. By comparing the results with past research studies on identifying the presence of toxicity in other languages, in this paper we highlighted the urgent need for culturally-aware, fairness based, and ethically motivated toxicity detection approaches for low-resource language.

Keywords: Toxicity detection, Less resource Bodo language, bias issues, deep learning, Natural Language Processing.

I. INTRODUCTION

Online platforms have become key-worlds for social interaction, political discourse and expression, and community formation. However, these platforms have been a tool for the rapid growth of toxic content including hate speech, harassment, threats and cyberbullying. Empirical studies have shown that exposure to toxic online discourse has been coupled with psychological distress, heightened social polarization, minority relies suppression and declining democratic norms [1], [2].

To cope with this growing issue, social media siteing are increasingly dependent on Artificial Intelligence (AI), specifically automated toxicity detection systems that are built on machine learning (ML) algorithms and deep learning (DL). Such systems provide the ability for moderation on a massive scale that would not be possible through manual

review. Advances in deep neural architectures in particular transformer based architectures have provided improve performance in toxicity classification tasks [3], [4]. Despite these developments there are some significant concerns about prejudice, fairness and ethical considerations in automated toxicity detection. Previous studies have shown that these systems tend to misclassify content exercised by minority identities, dialects and culturally specific expressions as toxic [5].

The problem is especially pronounced with respect to languages that are low-resource and indigenous, which often lack access to large annotated data, well developed linguistic tools, and appropriate representation in AI research [6]. The Bodo language, which is mentioned in the Eighth Schedule of the Indian Constitution, is an example of such a challenge. The Bodo Language Though Bodo has a rich literary tradition, and more and more digital presence is also building within it, it is mostly underrepresented by Natural Language Processing (NLP), especially in sentence-sensitive aspects, such as toxicity detection. Bodo language has a strong prevalence of highly metaphorical language construction, expressive style modes, and cultural standard language. When tested through English-centric models, such expressions may get erroneously categorized as toxic [3]. The use of automated moderation system without consideration for these language and cultural details carries the risk of unfair censorship, cultural marginalization, and harm to its ethical norms [7]. This research tackles these issues with a methodical investigation of automated toxicity detection in Bodo text with a special emphasis on the bias and ethical risks that exist in the deep learning approaches.

A. Objectives of the Study

The primary objectives of this research are:

1. To examine ethical risks related to fairness, cultural misrepresentation, and automated moderation systems
2. To compare toxicity detection outcomes with findings reported across other linguistic contexts
3. To project technical, ethical, and methodological study for toxicity detection in low-resource languages.

A. Significance of the Study

This research is making a contribution to the development of low-resource language NLP, developing ethical AI, and to inclusive digital governance. By focusing on Bodo, the study supports the focus on linguistic diversity in AI research and registers empirical research on matters of bias in automated toxicity identification systems. The results provide comic guidance for devising fair-mindedness-aware moderation frameworks in multilingual societies and represent policymakers, platform developers, and researchers of the ethical risks of automated content moderation technologies [7], [2].

II. RELATED WORK

Early methods of detecting toxicity were mostly based on rule-based methods as well as conventional machine learning methods that make use of manually developed features. The development of deep learning allowed to obtain richer semantic and contextual representations, which contributed considerably to the improvement in classification results [3]. Convolutional Neural Networks (CNNs) and Recurrent Neural Networks (RNNs) were fundamental steps in development and transformer-based architectures such as BERT and XLM-R are now state-of-the-art methodologies [4]. Large-scale datasets such as DeToxy [8] and ToxiGen [9] have further boosted the growth of multimodal and implicit toxicity detection research.

The detection of bias in toxicity detection systems has been well documented. Models often relate words related to identity (e.g. religion, gender, ethnicity) with toxicity, resulting in high false positive rate for marginalised communities [5]. Elsafoury and Katsigiannis [10] shows the way that NLP systems have a discrete way of disadvantaging underrepresented groups. Similarly, Sap et al. [11] demonstrate that the beliefs and social identities of the annotators as well as the moral framework greatly affect the labelling decisions and introduce subjective bias into training datasets.

Fairness, transparency, accountability, privacy, and potential misuse of AI technologies [7], [9], [12], are ethical considerations in NLP. In high-stakes areas of society (journalism, healthcare), such biased systems of NLP may cause real societal harm [12], [13]. More general considerations of algorithmic bias further explain difficulties with structural constraint in the use of AI systems [14].

Although extensive works can be found about hate speech and toxicity detection [1], [3] there have been little studies specifically focused on the detection of bias and ethical concerns in toxicity detection in Bodo language. This study aims to fill that gap while using cultural sensitivity as well as linguistic nuances in the development and evaluation of models

III. MATERIALS AND METHODS

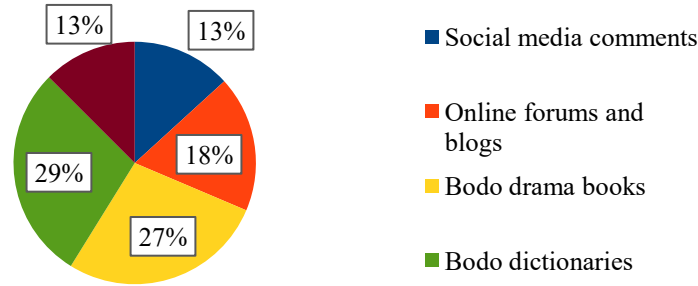
A. Dataset Collection

The corpus developed for this study was built entirely based on the secondary textual sources making sure ethical compliance and large linguistic coverage. A total of 12,037 texts of Bodo language were obtained from heterogeneous platforms which include informal and formal tone of language usage. The diversity of sources was aimed at improving the representative in every aspect of the Bodo language. The following Table 1 summarizes the distribution of text samples across different sources.

Table 1. Text Samples Distribution of Bodo Language

Source	Number of Samples
Social media comments (Facebook, YouTube)	1592
Online forums and blogs	2186
Bodo drama books	3304
Bodo dictionaries	3450
Bodo Short stories and literary texts	1505
Total	12037

The incorporation of literary texts, drama scripts and dictionaries was of special relevance to capture idiomatic expressions, culturally embedded metaphors and traditional discourse patterns but not typically present in social media contents. This is balanced composition helps to reduce domain bias and helps to make models of detecting downstream toxicity more robust. Figure 1 is a history of the contribution that the various data sources have made to the overall corpus in proportion.

Figure 1. Source-wise distribution of Bodo text samples.**B. Annotation Strategy**

The texts gathered were manually annotated in two exclusive categories:

- Toxic
- Non-toxic

Annotation was conducted by native speakers of Bodo with linguistics and literary academic training, to insure that contextual meaning, nuances in pragmatics and symbolism in the culture is considered. A thorough guideline for annotation was generated, with toxicity defined in the terms of toxic, sever toxicity, slang, insult, abusive and identity hate, taking specific account of metaphorical and narrative usage.

In order to minimize subjectivity and inter-annotator inconsistency, ambiguous cases were discussed as a collective and consensus label description was adapted. This human-in-the-loop approach was helpful to increase the reliability of these annotations, especially for culturally expressive and metaphorical content which is a common part of Bodo discourse.

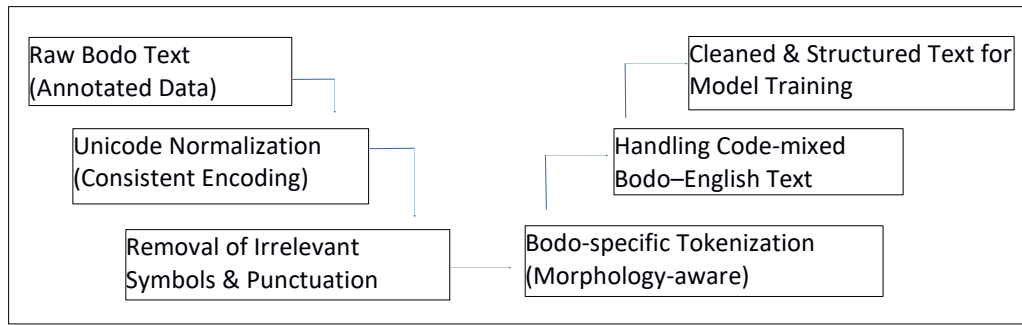
C. Data Preprocessing

Before training of the model, the annotated corpus was processed through a careful pipeline to prepare a corpus suitable for training the model keeping in mind the specific linguistic aspects of Bodo language:

- * Unicode normalization to provide consistent encoding
- * Elimination of useless symbols and unnecessary punctuation
- * Tokenization codaꠁꠞꠞ explained to Bodo morphological structures
- * Handling of code mixed Bodo-English expressions, very often encountered over digital communication

These preprocessing steps have been crucial to reduce the noise and enhance the generalisation of the models and in particular for the informal and code-mixed text segments. Figure 2 outlines the overall preprocessing process used in this study.

Figure 2. Preprocessing pipeline for Bodo language text.



D.

Deep Learning Models

The deep learning architectures that were implemented and evaluated to study their effectiveness of Bodo toxic text detection are RNN, BiRNN, LSTM, BiLSTM, GRU, BiGRU. Recurrent Neural Networks (RNNs) are very popular when it comes to modeling sequential data because of their dynamics based on the concept of hidden states which allow them to remember previous information. Nevertheless conventional RNNs do have a limitation when it comes to long term dependencies, that is vanishing gradient problem. Bidirectional RNNs (BiRNNs) overcome this shortcoming by taking past and future context into consideration by forward and backward sequence processing. Long Short-Term Memory (LSTM) network introduces gated structures to effectively control the process of information retention and forgetting, which helps to develop robust learning of long-range dependencies. A further benefit of representational capacity is available if Bidirectional LSTMs (BiLSTMs) are used, which benefit from both preceding and succeeding context at once. Gated Recurrent Units: These are simpler variants of the LSTM architecture and work as well as the LSTM architecture but at a reduced computational cost. Bidirectional GRUs (BiGRUs) extend the capability further by applying two-directional processing, and thus they are effective in solving complex sequence classification and understanding cases.

E. Evaluation Metrics and Bias Assessment

Model performance was evaluated using standard classification metrics such as Precision, Recall, Accuracy, and F1-score.

To determine fairness and bias, other analyses were performed using False Positive Rate (FPR) of identity related, culturally sensitive terms Qualitative error analysis, categorical content

This dual framework of evaluation takings care to ensure two things-predictiveness effectiveness and ethics reliability.

III. RESULTS AND DISCUSSION

A. Overall Model Performance

Among the evaluated architectures, the GRU model gave the highest overall accuracy results, indicating that it is a good model for understanding the context of semantics and fine language patterns in Bodo text. The BiRNN model also showed to be a competitive model reinforcing how the bidirectional modeling of context is important for sequence classification tasks. In comparison, the LSTM model achieved relatively poor performance. Table 2 shows the comparative performance of the deep learning models models on the test dataset.

Table 2 Performance Comparison of Deep Learning Models

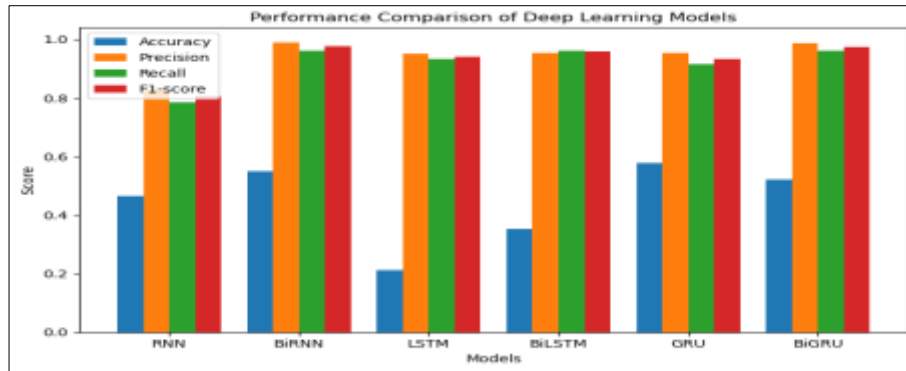
Model	Accuracy	F1-Score
-------	----------	----------

RNN	0.464789	0.806084
BiRNN	0.549296	0.975845
LSTM	0.211268	0.941896
BiLSTM	0.352113	0.957529
GRU	0.577465	0.933852
BiGRU	0.521127	0.973154

The strong results obtained by GRU and BiRNN architectures are consistent with broader findings in deep learning and large language model research [4]. However, high predictive performance does not necessarily imply ethical robustness.

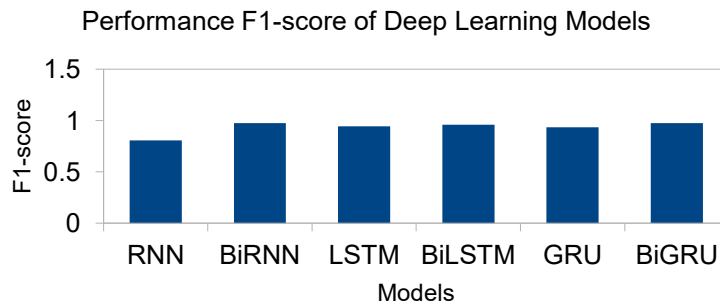
Figure 4 demonstrates that the variants of Recurrent Models architectures across all evaluation metrics, confirming its effectiveness in capturing contextual semantics in the Bodo language.

Figure 4: Model Performance Comparison



As shown in Figure 5, the BiRNN model achieves the highest F1-score (0.975845), indicating balanced precision and recall, which is critical for toxicity detection tasks.

Figure 5. F1-score Comparison



B. Bias Analysis and Error Patterns

Despite good aggregate measurements, however, the models had high false positive rates for identity-related expressions. This result is similar to previous research that shows that toxicity detection systems often mistake identity references with abusive language [5], [15]. In the linguistic light of the Bodo's it further shows that the culturally normative rhetorical intensity further increase this risk of misclassification.

Annotator-related bias was also a cause of inconsistency of labelling, which supports earlier findings that annotation processes are inherently value-laden rather than neutral [11]. Moreover, there are risks of significant unfair censorship, erosion of cultural expression and lack of transparency pointing to wider ethical concerns which are Scale in NLP scholarship [12], [9]. Critical reflections on proactive content moderation systems are in similar vein against excessive trust in automated toxicity detection [16].

Recent progress in deep learning for domain-specific toxic detection provides an example of the growing versatility in using such models in different linguistic contexts [17]-[19]. In spite of such advancements, making these must be accompanied by strict frameworks that evaluate fairness.

Despite the relative overall superiority, deep learning models were found to be more sensitive to identity-related expressions, as revealed in Table 3.

Table 3. Bias & Ethical Analysis Based on False Positive Rate

Category	False Positive Rate
Identity-related comments	16.7%
Cultural metaphors	12.5%
Political criticism	11.2%
Neutral discourse	6.1%

The distribution of false positive rates between different discourse categories represents the systematicity of bias in deep learning-based moderation systems. Identity-related comments have the most errors (>16.7%) which could indicate that models disproportionately link identity markers to harmful intent. Such patterns are often the result of imbalanced training data where the co-occurrence of marginalised identities in the training data and the co-occurrence of abusive language create the risk of representational harm from having too much data flagged as legitimate expression. Cultural metaphors (12.5%) - These are an additional form of uncovering contextual limitations, with the models finding it difficult to understand the figurative and cultural language used due to the prevalence of one culture in datasets that are biased to the dominant culture. Political criticism (11.2%) raises other ethical concerns as adversarial rhetoric may get misclassified as abusive that may constrain democratic discourse. Although conflicting discourse has lower rates of error (6.1%), this difference highlights that errors concentrated in socially-sensitive contexts. These disparities are caused by annotation bias as well as biased data distributions and optimization for accuracy across groups instead of within subgroups. Mitigating such harms requires fairness-aware evaluation, diverse curation of data and human oversight to counterbalance balancing the harms of harm prevention, and protective of a legitimate speech.

III. CONCLUSION

This paper is a thorough investigation of the problem of bias and ethical risks in automated toxicity detection for Bodo text as implemented by deep learning architectures. While modern models perform well on classification tasks, they present a large number of ethical challenges, especially when it comes to misclassification of expressions that have to do with culture and identity.

The results are underlined by the fact that:

i. The importance of evaluation Accuracy-based results are not enough for sensitive NLP applications.

ii. Cultural awareness and bias mitigation is a critical part of ethical development of AI.

iii. Low-resource languages need customization of modelling approaches rather than borrowing of high-resource language systems.

The future work should focus on bias-aware training methods, explainable AI and participatory annotation frameworks in order to make the toxicity detection systems fair, inclusive and culturally responsive.

Funding Statement: The author(s) received no specific funding for this study.

Conflicts of Interest: The authors declare that they have no conflicts of interest to report regarding the present study.

REFERENCES

- [1] S. Das, A. Dutta, K. Roy, A. Mondal, and A. Mukhopadhyay, "A Survey on Automatic Online Hate Speech Detection in Low-Resource Languages," 28 Nov 2024. <https://arxiv.org/html/2411.19017v1>
- [2] Nihar Sahoo, Himanshu Gupta, Pushpak Bhattacharyya, 2022, "Detecting Unintended Social Bias in Toxic Language Datasets", Proceedings of the 26th Conference on Computational Natural Language Learning (CoNLL) , pages 132 – 143, December 7- 8, 2022 ©2022 Association for Computational Linguistics, DOI: <https://doi.org/10.18653/v1/2022.conll-1.10>
- [3] Abhishek Velankar, Hrushikesh Patil, Amol Gore, Shubham Salunke and Raviraj Joshi, "Hate and Offensive Speech Detection in Hindi and Marathi," CEUR Workshop Proceedings 2021, ISSN 1613-0073, Vol-3159:239-247, <https://ceur-ws.org/Vol-3159/>, <https://doi.org/10.48550/arXiv.2110.12200>
- [4] H. Naveed et al., "A Comprehensive Overview of Large Language Models," arXiv preprint arXiv:2307.06435v10, Oct. 2024. <https://doi.org/10.48550/arXiv.2307.06435>
- [5] A. Kumar and P. Kumar, "Investigating Bias in Automatic Toxic Comment Detection: An Empirical Study," arXiv preprint arXiv:2108.06487, 2021. <https://doi.org/10.48550/arXiv.2108.06487>
- [6] S. Das et al., "A Survey on Automatic Online Hate Speech Detection in Low-Resource Languages," arXiv preprint arXiv:2411.19017, 2024. <https://arxiv.org/html/2411.19017v1>
- [7] S. Caton and C. Haas, "Fairness in Machine Learning: A Survey," ACM Computing Surveys, vol. 56, no. 7, Article 166, Jul. 2024. <https://doi.org/10.1145/3616865>
- [8] S. Ghosh, S. Lepcha, S. Sakshi, R. Shah, and U. Srinivasan, "DeToxy: A Large-Scale Multimodal Dataset for Toxicity Classification in Spoken Utterances," in Proc. Interspeech, 2022, pp. 5185–5189. <https://doi.org/10.21437/Interspeech.2022-10752>
- [9] T. Hartvigsen, S. Gabriel, H. Palangi, M. Sap, D. Ray, and E. Kamar, "ToxiGen: A Large-Scale Machine-Generated Dataset for Adversarial and Implicit Hate Speech Detection," arXiv preprint arXiv:2203.09509, 2022. <https://doi.org/10.48550/arXiv.2203.09509>
- [10] F. Elsafoury, S. Katsigiannis, and N. Ramzan, "On Bias and Fairness in NLP: How to Have a Fairer Text Classification?" arXiv preprint arXiv:2305.12829, 2023. <https://doi.org/10.48550/arXiv.2305.12829>
- [11] M. Sap et al., "Annotators with Attitudes: How Annotator Beliefs and Identities Bias Toxic Language Detection," arXiv preprint arXiv:2111.07997, 2022. <https://doi.org/10.48550/arXiv.2111.07997>
- [12] B. D. W. O. J. 4th, R. Nieva, S. S. Lee, and N. Elhadad, "A Scoping Review of Ethics Considerations in Clinical Natural Language Processing," JAMIA Open, vol. 5, no. 2, ooac039, 2022. <https://doi.org/10.1093/jamiaopen/ooac039>

- [13] Carter, Emily & Narayanan, Ravi & Samuel, Adebis. "Ethical Challenges and Bias in NLP Models: A Python-Based Investigation". (May 2025).
https://www.researchgate.net/publication/392159070_Ethical_Challenges_and_Bias_in_NLP_Models_A_Python-Based_Investigation
- [14] N. Kordzadeh and M. Ghasemaghaei, "Algorithmic Bias: Review, Synthesis, and Future Research Directions," *European Journal of Information Systems*, vol. 31, no. 3, pp. 388–409, 2022. <https://doi.org/10.1080/0960085X.2021.1927212>
- [15] Atijit Anuchitanukul, Julia Ive, and Lucia Specia. 2022. "Revisiting Contextual Toxicity Detection in Conversations." *J. Data and Information Quality* 15, 1, Article 6 (March 2023), 22 pages. <https://doi.org/10.1145/3561390>
<https://dl.acm.org/doi/10.1145/3561390>
- [16] M. Warner, A. Strohmayer, M. Higgs, and L. Coventry, "A Critical Reflection on the Use of Toxicity Detection Algorithms in Proactive Content Moderation Systems," *International Journal of Human-Computer Studies*, vol. 198, 103468, Apr. 2025.
<https://doi.org/10.1016/j.ijhcs.2025.103468>
(<https://www.sciencedirect.com/science/article/pii/S1071581925000254>)
- [17] S. Chaudhary, "Advanced Language Toxicity Detection Using Deep Learning," *International Journal of Electrical and Electronics Engineering*, vol. 14, no. 1–2, pp. 18–21, Apr. 2025.
- [18] A. Abbasi, A. R. Javed, and F. Iqbal et al., "Deep Learning for Religious and Continent-Based Toxic Content Detection and Classification," *Scientific Reports*, vol. 12, 17478, 2022.
<https://doi.org/10.1038/s41598-022-22523-3>
- [19] M. Shahid et al., "Leveraging Deep Learning for Toxic Comment Detection in Cursive Languages," *PeerJ Computer Science*, vol. 10, e2486, Dec. 2024.
<https://doi.org/10.7717/peerj-cs.2486>