WoNBias: A Dataset for Classifying Bias & Prejudice Against Women in Bengali Text

Md. Raisul Islam Aupi, Nishat Tafannum, Md. Shahidur Rahman,
Kh Mahmudul Hassan, Naimur Rahman
Shahjalal University of Science & Technology
raisul05@student.sust.edu, nishat29@student.sust.edu, rahmanms@sust.edu,
kh39@student.sust.edu, naimur79@student.sust.edu

Abstract

This paper presents WoNBias, a curated Bengali dataset to identify gender-based biases, stereotypes, and harmful language directed at women. It merges digital sources- social media, blogs, news- with offline tactics comprising surveys and focus groups, alongside some existing corpora to compile a total of 31,484 entries (10,656 negative; 10,170 positive; 10,658 neutral). WoNBias reflects the sociocultural subtleties of bias in both Bengali digital and offline conversations. By bridging online and offline biased contexts, the dataset supports content moderation, policy interventions, and equitable NLP research for Bengali, a low-resource language critically underserved by existing tools. WoNBias aims to combat systemic gender discrimination against women on digital platforms, empowering researchers and practitioners to combat harmful narratives in Bengali-speaking communities.

1 Introduction

To provide essential context for our work, it is crucial first to understand the linguistic landscape of Bengali (Bangla), an Indo-European, Indo-Aryan language primarily spoken in Bangladesh and West Bengal, India. While mutually intelligible, regional variations exist, with this paper focusing on the Bangladeshi variety. Gender bias and negative gender discourse against women on digital platforms in Bengali often escape detection because of the linguistic inability of universal tools and fragmented moderation infrastructure. With its flexible grammar and rich corpus of idioms, Bengali offers a source of subtle stereotypes as well as hate speech that can persist, especially in informal online discourse. There are so far not enough data, specifically to detect bias in language with low resource potential, such as Bengali, which means that online discrimination against women cannot be detected by automated content filters.

Further, we present WoNBias, a dataset of 31,484 annotated texts from social medias, news platforms, blogs, offline surveys, and focus groups. The dataset includes entries in Negative (2), Positive (1), and Neutral(0) categories, respectively, to explore the sociocultural context of Bengali texts. Our study highlights the need for language-specific resources to contribute towards better content moderation, training equitably effective language models in Bengali, and combating discriminatory behavior towards women in social media.

1.1 Bias Statement

In this paper, we identify and analyze bias against women in Bengali text. We define this bias as language that systematically demeans women, perpetuates harmful stereotypes, and erases or fails to recognize their equal status and contributions. This constitutes a **representational harm**(Blodgett et al., 2020).

Such representational harms are damaging because they reinforce restrictive and inappropriate stereotypes about the roles women are expected to perform, such as the notion that (women shouldn't study science) "মেয়েদের সায়েন্স পড়ার দরকার নাই". When automated systems, such as large language models, are trained on data containing this language, they risk perpetuating and even amplifying these societal inequities. This can lead to downstream allocational harms, where systems unfairly limit opportunities for women in areas like professional development, and contributes to the disenfranchisement of women in online spaces.

Our work is based on the normative stance that language should not subordinate individuals based on gender(Blodgett et al., 2020). The WoNBias dataset has been created to directly address this issue. By providing a benchmark for identifying toxic and stereotypical language, WoNBias enables the development of NLP tools that can counteract rather than reinforce existing gender imbalances in Bengali-speaking communities.

2 Related Work

Recent work has focused on identifying and reducing the biases present in large language models (LLMs), with benchmark datasets playing a key role in that effort. One notable example is **BOLD**(Dhamala et al., 2021), a dataset and evaluation framework designed to surface stereotypes in open-ended text generation across domains like gender, race, profession, religion, and politics in English. By comparing model-generated text to Wikipedia-derived prompts, **BOLD** shows that LLMs often produce more biased or toxic content than human writers, highlighting the need for more responsible generative systems.

More recently, **BanStereoSet**(Kamruzzaman et al.,

Proceedings of the 6th Workshop on Gender Bias in Natural Language Processing (GeBNLP), pages 105–110

2024) introduced a culturally grounded benchmark for Bengali, with 1,194 sentences covering nine categories such as race, profession, and religion. This dataset helps reveal how multilingual LLMs carry over or even amplify localized social biases, especially in underrepresented languages. Together, these resources stress the importance of culturally diverse benchmarks when evaluating model fairness.

Underlying all these biases is the data these models are trained on. Studies show that stereotypes in training data often get reinforced by LLMs—such as associating certain professions with specific genders or favoring dominant religious narratives even after corrective prompting (Kotek et al., 2023; Abid et al., 2021). This problem also shows up in multilingual contexts. For example, LLMs tend to default to Western views even when responding to prompts rooted in Arab culture, a bias made clear through the **CAMeL** dataset of culturally grounded Arabic prompts (Naous et al., 2023; Ahn and Oh, 2021).

In low-resource languages like Bengali, progress is being made but challenges remain. **BanglaBERT** (Bhattacharjee et al., 2022), trained on a large, diverse collection of Bengali texts, has improved language understanding, but the model has no focused objective for eliminating discriminatory texts. For instance, the **BanglaSentNoB** (Islam et al., 2021) dataset shows that handcrafted features often outperform deep models when dealing with informal Bengali text. This points to a clear need for richer, context-aware datasets that reflect the diversity of the Bengali language and culture. Without intentional effort, LLMs risk repeating the same biases we wish to move past.

3 Dataset Creation

3.1 Methodology

3.1.1 Expansion Strategy

To scale the dataset to **31,484** entries, we strived to diversify data sources while ensuring representativeness.

- **Sources:** Collected text from Facebook posts and their comment sections (approximately 6,00,000 entries before filtering), regional newspapers native to Bangladesh (e.g., Prothom Alo, Ittefaq), and articles regarding gender issues by the government.

- **Collaboration:** We engaged with female students from a range of universities and colleges in Bangladesh, as well as working professionals, homemakers, and women from various other backgrounds, through conference calls and online surveys. Initially, we distributed a *questionnaire*[1] within our campus community. This form collected negative comments that participants faced in various social contexts, particularly concerning instances where they felt disadvantaged due to their gender. As our understanding of the issue deepened, we updated the questionnaire[2] to better capture the nuances of gender-based bias and discrimination and distributed it to a larger audience.

- **Web Scraping:** To collect textual data from online platforms, we manually gathered public Facebook posts and comments without violating Facebook's terms of service. Our web scraping was limited to publicly accessible content, and conducted strictly for academic research purposes. For websites and blogs (e.g., Prothom Alo, BD News 24), we used tools like *Web Scraper*[3] to extract relevant articles and forum discussions while adhering to standard scraping norms: avoiding large-scale data extraction that might burden servers, respecting copyright (e.g., quoting rather than duplicating full texts).

3.1.2 Annotation Process

- **Annotator Background and Quality Control**

  Our annotation team comprised seven university students (four male, three female) strategically selected to represent diverse geographical and cultural perspectives across Bangladesh's seven divisions: Mymensingh, Barishal, Rangpur, Sylhet, Chittagong, Rajshahi, and Khulna. This regional and gender diversity was essential to capture varied interpretations of **gender bias**, as expressions of bias manifest differently due to local social norms and dialectical variations. All annotators are fluent in standard Bengali with diverse academic backgrounds spanning agriculture, economics, political studies, and software engineering.

  To ensure annotation consistency in this diverse team, we implemented a rigorous **quality control protocol**. A balanced subset of 1,500 entries was labeled by two independent annotators to measure agreement. In cases of disagreement, a third senior annotator served as an arbitrator to resolve the conflict, a process that helped calibrate our annotations and maintain a shared understanding of the labeling criteria.

  We acknowledge several limitations: our annotators, university students aged 22-26, introduce potential generational, socioeconomic, and urban-centric biases. While their diversity aids robustness, it also leads to judgment variability (as seen in our inter-annotator agreement), particularly confusion between Positive and Neutral categories, reflecting subjective cultural and personal interpretations.

[1] Questionnaire: Collection of Personal Experiences Related to Gender Bias

[2] Updated Questionnaire: Collection of Personal Experiences Related to Gender-Based Bias and Discrimination

[3] Web Scraper

Table 1: Annotator Demographics

ID	Gender	Division	Univ./Dept. (Age)
1	Female	Mymensingh	BAU/Agriculture (20)
2	Female	Barishal	BU/Economics (23)
3	Male	Rangpur	SUST/Political Studies (24)
4	Male	Sylhet	SUST/Software Engg. (25)
5	Female	Chittagong	SUST/Software Engg. (26)
6	Male	Rajshahi	SUST/Software Engg. (26)
7	Male	Khulna	SUST/Software Engg. (25)

- **Inclusion Rules**

  1. **Self-Contained Context:** Only sentences that explicitly express bias or affirmation with clear, unambiguous meaning were included. This means we selected sentences where the bias is evident from the text itself without requiring external context for interpretation. For example, we discarded "মেয়েটার পরিণতি ঠিক-ই হয়েছে"[She got what she deserved](ambiguous - requires context about what happened) but kept "মেয়ে জাতটাই খারাপ"[women are the worst as a group](explicit and self-contained).
  2. **Derogatory Language Detection:** Flagged: Direct slurs (e.g., "নারীরা গোলাম"[women are slaves]), Gendered stereotypes (e.g., "মেয়েদের বিজ্ঞান পড়া উচিত নয়"[women shouldn't study science]), Dehumanizing comparisons (e.g., "স্ত্রীজাত গাধার সমান"[wives are like donkeys]).
  3. **Lexical Diversity:** Covered 60+ Bengali feminine terms (e.g., মেয়ে, নারী, স্ত্রী, বউ) and derogatory variants (e.g., অবলা, নষ্টমেয়ে, মাগী).

- **Quality Assurance**

  1. **Deduplication:** Removed identical and near-identical sentences(75% match) but retained paraphrases through script[4] (e.g., "মেয়েরা দুর্বল"[Girls are weak] → "নারীদের শারীরিক শক্তি কম")[Women are physically less capable].
  2. **Source Diversity:** To ensure balance, we collected data from both male-dominated Facebook groups, including anti-feminist forums, and progressive platforms such as women's rights blogs and government policy texts, capturing a wide spectrum of gender-related discourse.

- **Positive & Neutral Data**

  – Positive: Required explicit advocacy (e.g., "নারীরা সকল পেশায় সমর্থ"[women excel in all professions]).

  – Neutral: Excluded any gendered bias (e.g., "বাংলাদেশে শিক্ষার হার বাড়ছে"[literacy rates are rising in Bangladesh]).

- **Data Statement:** The full WoNBias dataset, comprising over 30,000 annotated samples, is publicly available at **gender-bias-bengali/wonbias-complete-dataset**(Aupi et al., 2025). This release supersedes the previously available partial subset and is intended to support further research on gender bias in Bengali NLP tasks.

3.2 Ethical Considerations

1. **Participant Consent**: All questionnaire participants were fully informed about the study's purpose and gave explicit consent, with the freedom to opt out at any time. In-person conversations were held only with their comfort confirmed and conducted respectfully.
2. **Ethical Data Sourcing**: Only publicly accessible content was used. Manual collection avoided mass scraping, and no data was taken from private profiles, closed groups, or paywalled sites. These practices followed the ethical principles outlined in the *Belmont Report* (National Commission for the Protection of Human Subjects of Biomedical and Behavioral Research, 1979).
3. **Anonymization**: Identifiable details such as names, locations, and links were removed early in the cleaning process. Free-text entries were reviewed to avoid accidental identity exposure, and pre-anonymized datasets like BanglaParaphrase (Akil et al., 2022) were used for safe vocabulary expansion.
4. **Mental Health Awareness**: Given the sensitive nature of some content, participants were never pressured to share distressing material. Annotators were provided regular breaks and emotional check-ins to maintain mental well-being during the labeling process.

4 Dataset Analysis

4.1 Statistics

WoNBias demonstrates balanced class distribution across sentiment categories, with each class comprising approximately one-third of the dataset (Table 2). This even distribution ensures unbiased model training and evaluation across all categories.

Table 2: Class Distribution

Sentiment Class	Count	Percentage
Negative	10,656	33.84%
Positive	10,170	32.32%
Neutral	10,658	33.84%
Total	31,484	100.0%

[4]Detect direct or partial duplication in individual dataset label

Lexical diversity analysis[5] reveals substantial vocabulary richness with 52,671 unique tokens in 31,498 texts (Table 3). The high percentage of hapax legomena (61. 60%) indicates extensive lexical variation, while the relatively consistent average text length (11.52-12.31 words) ensures comparable complexity between classes. Negative texts show the highest lexical diversity (24,081 unique tokens), reflecting the varied expressions of bias in Bengali discourse.

Table 3: Lexical Diversity Metrics

Metric	All	Neg	Pos	Neu
Texts	31,484	10,656	10,170	10,659
Unique tokens	52,671	24,081	15,426	29,441
Total words	371,781	-	-	-
Avg words/text	11.80	12.31	11.57	11.52
Hapax legomena	32,447	-	-	-
Hapax %	61.60%	-	-	-

4.2 Quality Metrics

To ensure annotation consistency, two independent annotators labeled a balanced subset of 1,500 entries (500 per class) from the WoNBias dataset. The following contingency matrix was created to reflect their agreement and disagreement, particularly highlighting confusion between the positive and neutral categories.

Table 4: Contingency Matrix Between Coder A and Coder B

	Neg	Neu	Pos	Total
Neg	446	32	22	500
Neu	27	398	75	500
Pos	12	89	399	500
Total	485	519	496	1,500

Observed Agreement (P_o):

$$P_o = \frac{446 + 398 + 399}{1500} = \frac{1243}{1500} \approx 0.8287$$

Expected Agreement (P_e):

$$\begin{aligned} P_e &= \sum_{i=1}^{3} \left(\frac{\text{Row}_i \cdot \text{Col}_i}{N^2} \right) \\ &= \frac{500 \cdot 485}{1500^2} + \frac{500 \cdot 519}{1500^2} + \frac{500 \cdot 496}{1500^2} \\ &= \frac{242500 + 259500 + 248000}{2250000} \\ &= \frac{750000}{2250000} = 0.3333 \end{aligned}$$

Cohen's Kappa (κ):

$$\begin{aligned} \kappa &= \frac{P_o - P_e}{1 - P_e} \\ &= \frac{0.8287 - 0.3333}{1 - 0.3333} \\ &= \frac{0.4954}{0.6667} \approx 0.7431 \end{aligned}$$

Interpretation: The inter-annotator agreement yields $\varkappa$= 0.74 (95% CI [0.71, 0.77]), indicating substantial agreement according to Landis & Koch's benchmark ($\kappa > 0.61$ = substantial) (Landis and Koch, 1977). Some key observations emerge from the contingency matrix:

1. **High-Reliability Categories:** The negative class showed the strongest agreement, with 89.2% pairwise precision. This was due to the presence of clear lexical markers of bias, such as slurs and explicit comparisons.

2. **Positive/Neutral Ambiguity:** 16.4% of positive/neutral cases were contested — 75 out of 500 neutral cases were labeled as positive, and 89 out of 500 positive cases were labeled as neutral. Disagreements arose from sentences containing implicit praises & context-dependent sentiment.

3. **Adjudication Protocol:** The third annotator's arbitration, based on agreed-upon labeled data, was introduced to resolve conflicted entries.

4.3 Error Analysis

Common Mislabels:

- False Neutral: Sarcasm (e.g., নারীরা তো সবজান্তা! [Women know everything!]).
- False Positive: Neutral praise (e.g., মেয়েদের স্কুলে যাওয়া ভালো ["Girls going to school is good"]).
- Edge Cases: Code-mixed insults (e.g., মাইয়া পুরাই আন্টি আন্টি লাগে [aunt, derogatory]).

5 Applications

- In bias mitigation, WoNBias serves as:
  1. A **filter corpus** to remove gendered bias from pre-training data (e.g., for BanglaBERT(Bhattacharjee et al., 2022)).
  2. A **benchmark** to support in evaluating bias in existing dataset like BanStereoSet(Kamruzzaman et al., 2024).
  3. Similar to BOLD for English (Dhamala et al., 2021), WoNBias may help quantify bias in generative outputs like মেয়েরা [occupation] হতে পারে না [women can't be [occupation]]).
- For content moderation, WoNBias can help in real-time hate-speech detection on social platforms.

[5] Analyzing lexical diversity

- Regarding policy making, WoNBias can inform gender-sensitive AI policies in Bangladesh with the help of authorities like Bangladesh ICT Ministry.

6 Model Training and Results

We fine-tuned the *BanglaBERT-model* model[6] for bias classification on our dataset comprising three classes: *Neutral*, *Positive*, and *Negative*. The evaluation metrics focused on per-class recall (accuracy), as shown in Figure 1. The model achieved the highest recall for the *Neutral* class (0.962), followed by *Positive* (0.881), and *Negative* (0.824).

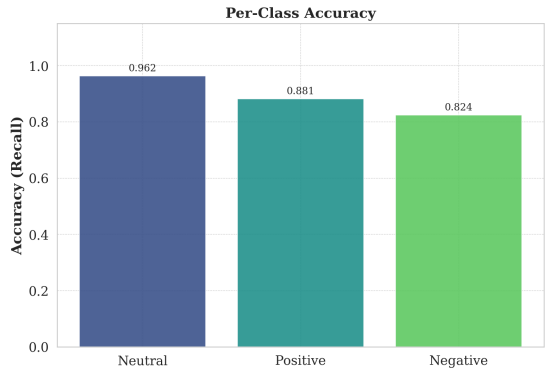

Figure 1: Per-class accuracy (recall) for the sentiment classifier.

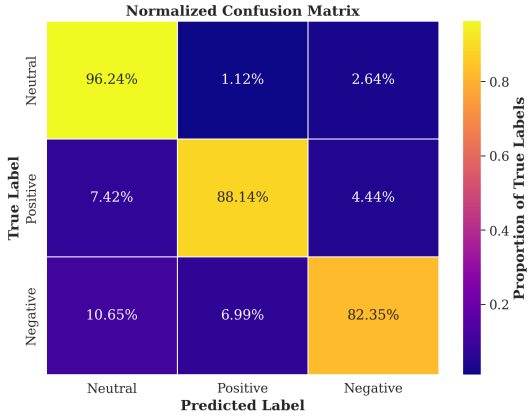

Figure 2: Normalized confusion matrix for the sentiment classifier.

To further analyze model performance, we provide the normalized confusion matrix in Figure 2.

The classifier shows relatively stronger performance in distinguishing *Neutral* and *Positive* classes, while the *Negative* class exhibits more confusion—most notably being misclassified as *Neutral* (10.65%).

While the overall performance is promising, we acknowledge that the classifier struggles more with the *Negative* class. This version of the model serves as a foundational baseline for further improvements in the classification of bias against women in Bengali. Future work will explore class imbalance handling, richer contextual embeddings, and domain-specific fine-tuning to mitigate these limitations.

[6] BanglaBERT-WoNBias-GenderBiasAndPrejudiceClassifier

7 Limitations & Future Plans

The dataset presents several limitations: it primarily focuses on **binary gender bias**, overlooking non-binary identities and intersectional discrimination, thus limiting broader applicability. Furthermore, its **lack of contextual bias detection** means keyword-based methods struggle with implicit or culturally coded biases like sarcasm. Lastly, the absence of **onomastic analysis** prevents distinguishing gendered names or analyzing related biases, limiting insights into subtle job associations.

In future work, we aim to pursue several avenues, including **Cross-Linguistic Expansion** of WoNBias to other South Asian languages (e.g., Urdu, Hindi) for comparative gender bias analysis. We also aim for enhanced **Dialect Coverage**, incorporating local dialects (e.g., Sylheti, Chittagonian) to explore bias variations across linguistic subcultures. Further, developing a **Bias Severity Scale** to classify intensity (mild stereotypes to hate speech) would enable targeted content moderation. Finally, Model Benchmarking on WoNBias would assess various language models' effectiveness in addressing gender bias.

8 Conclusion

This paper presents **WoNBias**, a comprehensive 31,484-entry annotated Bengali text dataset for detecting gender bias against women in digital discourse. Sourced diversely (social media, news, blogs, direct participant engagement), we have created a resource that captures the complex linguistic patterns of gender bias specific to the Bengali language and culture. The dataset's balanced distribution across the categories provides a solid foundation for training and evaluating bias detection systems. This paper addresses a critical gap in low-resource language NLP by providing a culturally grounded benchmark for bias detection in Bengali.

Our annotation process achieved substantial inter-annotator agreement ($\kappa = 0.74$), demonstrating the reliability of the dataset despite challenges in distinguishing between subtle forms of bias, particularly in the positive-neutral boundary cases. The extensive lexical diversity captured in WoNBias, with 52,671 unique tokens and over 60 Bengali feminine terms, ensures comprehensive coverage of gender-related discourse.

While acknowledging limitations, we are hopeful that our future work will incorporate dialect-specific annotations, develop nuanced bias severity classifications, and enhance contextual understanding capabilities to detect increasingly subtle forms of linguistic discrimination.

References

Abubakar Abid, Maheen Farooqi, and James Zou. 2021. Persistent anti-muslim bias in large language models. In *Proceedings of the 2021 AAAI/ACM Conference on AI, Ethics, and Society*, AIES '21, page 298–306, New York, NY, USA. Association for Computing Machinery.

Jaimeen Ahn and Alice Oh. 2021. Mitigating language-dependent ethnic bias in BERT. *CoRR*, abs/2109.05704.

Ajwad Akil, Najrin Sultana, Abhik Bhattacharjee, and Rifat Shahriyar. 2022. BanglaParaphrase: A high-quality Bangla paraphrase dataset. In *Proceedings of the 2nd Conference of the Asia-Pacific Chapter of the Association for Computational Linguistics and the 12th International Joint Conference on Natural Language Processing (Volume 2: Short Papers)*, pages 261–272, Online only. Association for Computational Linguistics.

Md. Raisul Islam Aupi, Nishat Tafannum, Md. Shahidur Rahman, Kh Mahmudul Hassan, and Naimur Rahman. 2025. Wonbias: A bengali dataset for gender bias detection. Available on Hugging Face Datasets.

Abhik Bhattacharjee, Tahmid Hasan, Wasi Ahmad, Kazi Samin Mubasshir, Md Saiful Islam, Anindya Iqbal, M. Sohel Rahman, and Rifat Shahriyar. 2022. BanglaBERT: Language model pretraining and benchmarks for low-resource language understanding evaluation in Bangla. In *Findings of the Association for Computational Linguistics: NAACL 2022*, pages 1318–1327, Seattle, United States. Association for Computational Linguistics.

Su Lin Blodgett, Solon Barocas, Hal Daumé III, and Hanna Wallach. 2020. Language (technology) is power: A critical survey of "bias" in NLP. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 5454–5476, Online. Association for Computational Linguistics.

Jwala Dhamala, Tony Sun, Varun Kumar, Satyapriya Krishna, Yada Pruksachatkun, Kai-Wei Chang, and Rahul Gupta. 2021. Bold: Dataset and metrics for measuring biases in open-ended language generation. In *Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency*, FAccT '21, page 862–872, New York, NY, USA. Association for Computing Machinery.

Khondoker Ittehadul Islam, Sudipta Kar, Md Saiful Islam, and Mohammad Ruhul Amin. 2021. Sentnob: A dataset for analysing sentiment on noisy bangla texts. In *Findings of the Association for Computational Linguistics: EMNLP 2021*, pages 3265–3271.

Mahammed Kamruzzaman, Abdullah Al Monsur, Shrabon Das, Enamul Hassan, and Gene Louis Kim. 2024. Banstereoset: A dataset to measure stereotypical social biases in llms for bangla. *Preprint*, arXiv:2409.11638.

Hadas Kotek, Rikker Dockum, and David Sun. 2023. Gender bias and stereotypes in large language models. In *Proceedings of The ACM Collective Intelligence Conference*, CI '23, page 12–24, New York, NY, USA. Association for Computing Machinery.

J. Richard Landis and Gary G. Koch. 1977. The measurement of observer agreement for categorical data. *Biometrics*, 33(1):159–174.

Tarek Naous, Michael J Ryan, and Wei Xu. 2023. Having beer after prayer? measuring cultural bias in large language models. *arXiv preprint arXiv:2305.14456*.

National Commission for the Protection of Human Subjects of Biomedical and Behavioral Research. 1979. The belmont report: Ethical principles and guidelines for the protection of human subjects of research. `https://www.hhs.gov/ohrp/regulations-and-policy/belmont-report/index.html`. U.S. Department of Health and Human Services.