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

Anonymous ACL submission

Abstract

In our endeavor, WoNBias: A Dataset for Classifying Bias & Prejudice Against Women in Bengali Text, we focus on identifying and classifying biases, hurtful comments, and stereotypes faced by Bengali-speaking women across various platforms. The dataset was meticulously compiled by scouring tens of online groups, surveying individuals, conducting multiple group sessions, and reviewing over a thousand posts from social sites, blogs, newspaper articles, and existing datasets such as Bengali-paraphrase (Akil et al., 2022). This comprehensive data collection process resulted in a final dataset size of 11,178 entries, categorized into 4028 negative, 3506 positive, and 3644 neutral entries.

Our study underscores the necessity for a standard dataset to train models that detect and combat biases, hateful comments, prejudices, and stereotypes faced by Bengali-speaking women. The WoNBias model achieved a 91% classification accuracy on 11,178 entries, challenging the performance of models like GPT-3.5-turbo. This dataset is effective for content moderation and training universal language models to eliminate biases in Bengali text generation.

1 Introduction

In this paper, our primary aim was to create a complete dataset that encompasses a variety of stereotyping, hateful or discouraging comments, and ideologies toward women. The idea was to train a classification model that properly labels any blatant violation of this caliber. We initially collected a total of 11,178 sentences and paragraphs by targeting Facebook posts triggered by certain events involving women, Facebook groups that are dedicated to inspiring hatred toward any forward-thinking women in general, news articles that covered incidents of discrimination and violations against women, and blog posts where people expressed their ideology on this matter. We manually categorized the collected comments and thoughts into three categories: Neutral (0), Positive (1), and Negative (2).

It was our understanding that the Bengali-speaking population uses Bengali to express hateful comments more as it has less moderation than English.

1.1 Related Work

The LLMs have inherited society's stereotypes due to the biased training data being fed into it. Research has shown that LLMs can introduce and amplify social prejudices(Kurita et al., 2019).

In the large language models, the most conspicuous prejudice is against a specific gender. The systems tend to associate a specific profession with a specific gender and further rationalize their preconceived notions by conjuring illogical reasons. (Kotek et al., 2023). It was observed that preconceived ideas against a specific religion were difficult to overcome even with positive prompting in the GPT-3 Large Language Model (Abid et al., 2021). The functioning monolingual language models are inclined toward Western culture, even when asked in Arabic and contextualized in an Arab cultural situation. To measure this Western bias, CAMeL, a dataset of naturally occurring Arabic prompts spanning eight diverse cultural aspects and an extensive list of 20,504 cultural targets corresponding to Arab or Western culture, has been recently introduced (Naous et al., 2023). A way to mitigate ethnic bias in the models in question, a multilingual model or contextual word alignment of two monolingual models, can be utilized (Ahn and Oh, 2021).

Bhattacharjee et al. (Bhattacharjee et al., 2022) introduce BanglaBERT, a BERT-based Natural Language Understanding (NLU) model, pretrained with 27.5 GB of Bangla data, which is collected by crawling 110 popular Bangla sites. An

annotated sentiment analysis dataset involving informally written Bangla texts is introduced(Islam et al., 2021a) , evaluating that in developing a benchmark classification system, hand-crafted lexical features outperform neural networks and pre-trained language models.

2 Data Collection

2.1 Sources & Participants

During our research, we chose to engage with both the general people (Especially women from different walks of life) and online communities with specific agendas (a combination of pro-feminine, and anti-feminine groups).

2.1.1 Participants

Primarily we released a form[1] on our campus, Shahjalal University of Science and Technology, that asked questions based on different social scenarios about what was some of their bad experiences for being a girl or a woman in those situations. Later we released the form in other universities like Jahangirnagar University, Chittagong University, and Bangladesh University of Professionals.

After receiving a few responses, we started having a few group sessions with people who would be comfortable sharing their experiences with us.

2.1.2 Other Sources

We utilized Facebook, pages, and groups to gather data on women's participation and autonomy. Initially, negative comments were collected, but later personal blogs were used to gather views on women's participation and autonomy. The study also examined Bangladesh's government initiatives and the constitution's clauses supporting women. Some of our neutral data was derived from the dataset provided by csebuetnlp/BanglaParaphrase (Akil et al., 2022) from huggingface.

2.2 Data Collection Statistics

After scouring through tens of online groups, surveying people, having multiple group sessions, and reviewing over a thousand online posts from different social sites, blog sites, newspaper articles, and existing datasets on Bengali paraphrase(Akil et al., 2022), we concluded our data collection process. The final size of our dataset stands at 11,178. The data overview is presented in Table 1, showing the number of data on all three of the categories.

[1]https://forms.gle/X7yFCioR6KcfAp4L6

Category	Number of Sentences & Paragraphs
Negative -	4028
Positive -	3506
Neutral -	3644
Total	11,178

Table 1: Data Overview of Complete Dataset

2.3 Data Collection Guidelines

The study aimed to measure the quality of data collected by identifying derogatory and biased sentences and paragraphs. The rules included self-explanatory sentences, avoiding duplication, and including a variety of feminine terms. Positive data should promote women's struggles, while neutral data should be unbiased. Data was collected from Facebook communities, pages, and groups, targeting male-dominated groups. Personal blogs were used to gather views on women's participation, government initiatives, and constitutional clauses supporting women.

3 Methodology

3.1 Mindset Behind the Approach

The research aims to train a robust model to detect and label social biases against women in Bengali-speaking communities. The model is trained by reviewing over a thousand online posts from social sites, blogs, and existing datasets(Akil et al., 2022), and strives for comprehensiveness and diversity.

3.2 Measuring Data Quality

We aspire to create a gold-standard dataset, a trustworthy reference or benchmark for evaluating algorithm or model performance in machine learning, data mining, and natural language processing.

- **High-Quality Annotations:** Researchers and assistants annotate the dataset with accuracy and reliability, ensuring each sentence and paragraph undergoes at least three revisions.
- **Filtering Data:** The dataset is divided into smaller parts for review, removing vague, incomprehensible context data.
- **Representation and Diversity:**Data is gathered from diverse sources to capture various scenarios and sentence types.
- **Consistency:** Standard rules were set for researchers to maintain consistency in the collected data.

3.3 Identifying model to develop

To build an accurate predictive model, we reviewed existing NLP models for the Bengali language, including csebuetnlp/banglabert's sequence classification model (Bhattacharjee et al., 2022), and SetFit for Text Classification (Tunstall et al., 2022) and SetFit Text Classification Hyper-parameter Search(Tunstall et al., 2022). We ultimately chose csebuetnlp/banglabert's model for its specialization in Bengali and its proven effectiveness in classification tasks like "SentNoB: A Dataset for Analyzing Sentiment on Noisy Bangla Texts" (Islam et al., 2021b).

3.4 Identifying model for comparison

To compare our trained model with real-world applications, we introduced the GPT-3.5 Turbo model (OpenAI, 2023) as the second model and tested the same dataset.

We chose GPT-3.5 Turbo, a top natural language processing model, excels in generating contextually relevant text, understanding language semantics and syntax, and providing strong sentiment analysis capabilities.

4 Results & Analysis

4.1 Enactment and Evaluation of Intended Model

Using our dataset, we enacted the csebuetnlp/banglabert's(Bhattacharjee et al., 2022) sequence classification model for double sequence classification to classify sentences and paragraphs with the label ranging from 0 to 2, where 0 stands for neutral, 1 for positive and 2 for negative. The dataset was divided into different ratios for training, validating, and testing. We started with the ratio of -

$$\text{Train} : \text{Validate} : \text{Test} = 90 : 5 : 5 \qquad (1)$$

And after adjusting a couple of times, proceeded with the following ratio -

$$\text{Train} : \text{Validate} : \text{Test} = 79 : 12 : 9 \qquad (2)$$

Our shuffling ensured a balanced distribution of data across all sets. After dividing, the training data stood at 8921, validation at 1261, and testing at 987.

Table 2 shows the overall accuracy percentage of the model's prediction on each iteration, indicating a very stable shift on different ratios, ranging from 89.46% to 91.69%. The stat showed that the 3rd iteration had the best accuracy overall. An increase in both the validation dataset size and testing dataset size influenced the accuracy better.

Iterations	Category	Ratio	Accuracy
1st iteration	Training	90%	90.73%
	Validation	5%	
	Testing	5%	
2nd iteration	Training	85%	89.46%
	Validation	10%	
	Testing	5%	
3rd iteration	Training	79%	91.69%
	Validation	12%	
	Testing	9%	

Table 2: Accuracy on different training scenario

As the model with the highest accuracy, the 3rd iteration model is kept for later comparison.

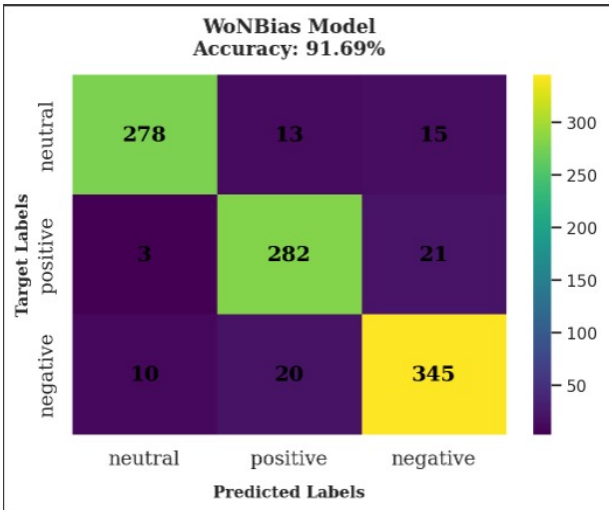

Figure 1: Accuracy of the WoNBias model

4.2 Testing the comparison model

We tested our bias detection dataset against the advanced GPT-3.5-turbo model. After connecting to the GPT-3.5-turbo API, we designed a prompt to guide the model [2]. Then we fed it some of our training data. We then sent chunks of our test data and stored the results for comparison. The GPT-3.5-turbo model achieved a total accuracy of 54.20% against our test dataset.

[2]Prompt designed to guide and test the GPT-3.5-turbo model

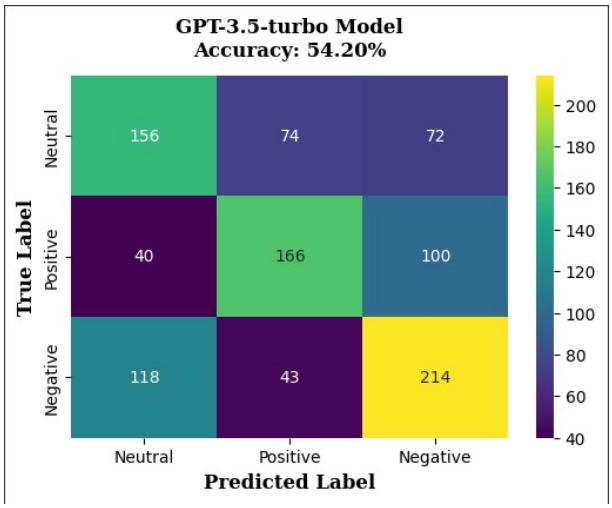

Figure 2: Accuracy of the GPT-3.5-turbo model

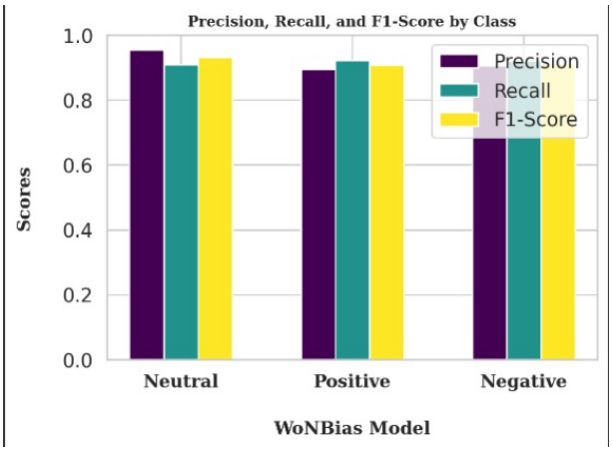

Figure 4: Performance stat for WoNBias model

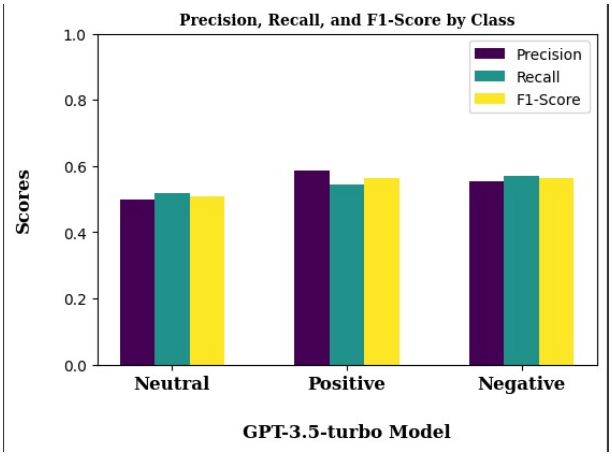

Figure 5: Performance stat for GPT-3.5-turbo model

4.3 Comparative Analysis

Judging by the quality of the response received in individual sentences and paragraphs, it was our understanding that the GPT-3.5-turbo model struggled to classify negative data altogether.

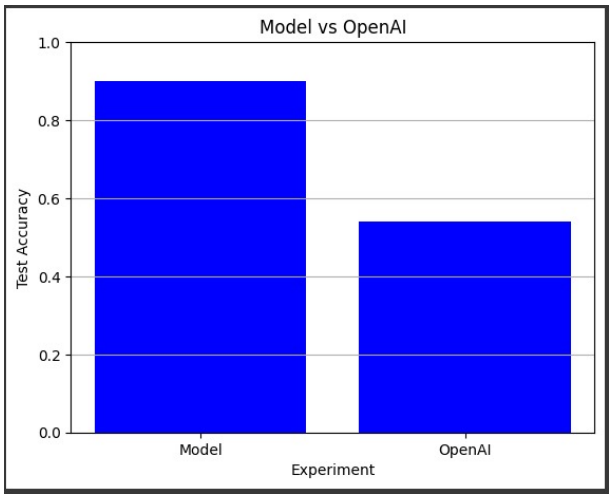

Figure 3: Accuracy difference between GPT-3.5-turbo and WoNBias model

Looking at other performance metrics, such as precision, recall, and f1-score, we saw a significant drop in every criterion for the GPT-3.5-turbo model.

4.4 Limitations

The data collection approach was exhaustive, resulting in less than the targeted amount. The guidelines for derogatory and biased data excluded subtle hate speech. The model struggled to identify feminine terms and flag derogatory sentences as negative, even when unclear about the intended meaning.

5 Conclusion

Our study highlights the critical need for a standardized dataset to train models aimed at detecting and combating biases, hateful comments, prejudices, and stereotypes encountered by Bengali-speaking women. Utilizing the "WoNBias: A Dataset for Classifying Bias & Prejudice Against Women in Bengali Text," our WoNBias model achieved a classification accuracy of 91% on 11,178 entries, surpassing the performance of models like GPT-3.5-turbo. This dataset proves effective for content moderation and training universal language models to eradicate biases in Bengali text generation.

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. *arXiv preprint arXiv:2210.05109*.

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.

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

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

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.

Keita Kurita, Nidhi Vyas, Ayush Pareek, Alan W. Black, and Yulia Tsvetkov. 2019. Measuring bias in contextualized word representations. *CoRR*, abs/1906.07337.

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*.

OpenAI. 2023. GPT-3.5-turbo. Language model.

Lewis Tunstall, Nils Reimers, Unso Eun Seo Jo, Luke Bates, Daniel Korat, Moshe Wasserblat, and Oren Pereg. 2022. Efficient few-shot learning without prompts. *arXiv preprint arXiv:2209.11055*.