Mitigation of User-Prompt Bias in Large Language Models: A Natural Langauge Processing and Deep Learning Based Framework

2024 IEEE 3rd International Conference on Computing and Machine Intelligence (ICMI) | 979-8-3503-7297-7/24/$31.00 ©2024 IEEE | DOI: 10.1109/ICMI60790.2024.10585628

Sarvesh Tiku
Department of Computer Science and Information Technology
Front Range Community College
Boulder, Colorado, USA
sarveshtiku@gmail.com

Abstract— **The advent of large language models has opened new frontiers in the field of automated text generation, enabling more refined engagement with complex language-based tasks. Concurrently, this advancement has revealed a potential vulnerability: the inadvertent amplification of biases from user prompts, which may lead to the reinforcement of detrimental stereotypes and misinformation by these large language models. Addressing this multifaceted challenge, this paper delineates a framework that integrates natural language processing and deep learning, designed to detect, and neutralize bias in user prompts in real time. The core of this system is a carefully formulated algorithm, the result of rigorous training, validation, and testing on the CrowS-Pairs dataset, specifically aimed at measuring the degree to which U.S. stereotypical biases are present in language models. The framework achieved an accuracy of 93% and an F1-Score of 0.92 in pinpointing and alleviating biases.**

Keywords— Artificial Intelligence, Deep Learning, Natural Language Processing, Bias Detection Mitigation, Computational Linguistics, Predictive Modeling, Bidirectional Encoder Representations from Transformers (BERT), Robustly Optimized BERT Approach.

I. Introduction And Literature Review

The explosion of information in the digital age has given rise to powerful tools that parse and interpret human language. Large language models, forming the backbone of contemporary AI systems, have reshaped the landscape of automated text generation, expanding the capacity to handle complex, nuanced language-based tasks. While these AI-driven engines offer unprecedented utility across a broad spectrum of applications, they are also susceptible to biases present in the data they are trained on. When fed biased prompts from users, these models may inadvertently produce and perpetuate biased responses. The amplification of harmful stereotypes, misinformation, and biased perspectives poses a significant challenge, impacting societal perceptions and interactions. This research aims to address this fundamental issue by introducing a robust framework combining natural language processing (NLP) and deep learning to detect and counteract bias in real-time. Mitigation of User-Prompt Bias in large language models has become an area of intensified focus due to the rising influence of these models on various facets of modern life, from decision-making to content generation. Though strides have been made, the field still awaits a holistic and nuanced approach. Some of the significant works in this domain are outlined below.

In 2018, Thompson et al. [1] explored various statistical methodologies to detect and quantify biases in text corpora. They employed linear regression models and Support Vector Machines (SVM) on several datasets. Despite their comprehensive approach to bias detection, their work fell short in providing specific categorization of biases and achieved an accuracy of 85%, pointing towards room for refinement.

In 2019, Chang et al. [2] developed an innovative framework using BERT and its variations for understanding stereotypical biases. Though their methods showed potential in assessing gender and racial biases, the focus remained narrow, and their highest F1-score of 0.78 indicated challenges in practical deployment.

In 2020, Martins et al. [3] implemented a deep learning architecture using LSTM for bias detection and mitigation. Trained on a custom dataset, they aimed to discern biases across several categories. Their methodology paralleled earlier works but culminated in an F1-score of 0.70, revealing limitations in its industrial application.

Lee et al. [4] and Johnson et al. [5] employed a combination of machine learning algorithms, including Random Forest and Neural Networks, on diverse data sources. Though their models demonstrated promising accuracy (around 88%), the datasets used were disproportionately inclined towards certain biases, undermining the generalizability and real-world relevance of their results.

Despite advancements in the field, there remains a notable gap in effectively understanding and mitigating biases initiated by user prompts in large language models. Previous methods have shed light on this issue but often fall short of comprehensively tackling biases across various categories such as race, gender identity, sexual orientation, and religion.

Typically, these efforts have been constrained by the scope of datasets or types of bias, with limited focus on real-time bias detection and mitigation. This research aims to bridge these gaps by introducing a framework designed to cover a broader range of biases and implement a method for immediate bias detection and correction. The demand for an integrated solution that ensures high accuracy, depth of categorization, and unbiased predictions continues to persist, reinforcing the value and timeliness of the this research.

II. Solution and Methodology

Recognizing the crucial need for reliable and comprehensive data, the study began with the selection of the CrowS-Pairs dataset, a contemporary and rigorously validated dataset that aptly fits the research context. Unlike traditional approaches, this research explores the utilization of RoBERTa (Robustly Optimized BERT Pretraining Approach), being a modern variation of BERT, has demonstrated superior performance in understanding contextual relations in text. By harnessing this model along with other advanced techniques, the research aims to push the boundaries of accuracy beyond the 88% threshold.

A. Dataset

The CrowS-Pairs dataset represents a marked deviation from conventional template-based bias evaluation datasets. The dataset's distinctive design is a result of collecting pairs of sentences that are minimally distant. Each pair consists of a sentence illustrating a stereotype or an anti-stereotype concerning a historically disadvantaged group in the United States, and a counterpart sentence pertaining to a contrasting advantaged group. The minimal divergence between the sentences in a pair enhances the precision in evaluating biases.

Bias Categories: Nine distinct bias types are encompassed within CrowS-Pairs, including race, gender identity, sexual orientation, religion, age, nationality, disability, physical appearance, and socioeconomic status. These categories have been meticulously selected to align with a narrowed version of the US Equal Employment Opportunities Commission's list [6] of protected categories.

The CrowS-Pairs dataset's crowdsourced nature empowers it with unparalleled diversity in both the stereotypes expressed and the structural intricacies of the sentences themselves. Unlike conventional datasets that often suffer from limitations in capturing the multifaceted nature of biases, CrowS-Pairs offers a rich tapestry of stereotypes and anti-stereotypes.

The dataset, designed to assess stereotypical biases within masked language models (MLMs), comprises 1,508 sentence pairs, representing a diverse array of biases sourced from the official repository associated with EMNLP 2020. Most existing datasets in the realm of bias detection within language models are constructed around isolated sentences or prompts that lack contextual nuance, limiting their effectiveness in simulating real-world usage scenarios of large language models (LLMs). Additionally, these datasets frequently concentrate on one or two types of bias, such as gender or race, neglecting the breadth of societal biases. In contrast, the CrowS-Pairs dataset encompasses a broader spectrum of nine bias categories, closely aligned with the US Equal Employment Opportunities Commission's list of protected categories. This comprehensive approach allows for a more nuanced exploration of biases, including those based on disability, physical appearance, and socioeconomic status, areas often overlooked in bias research.

The design and scope of CrowS-Pairs thus address several key limitations observed in other datasets:
Contextual Relevance: By utilizing sentence pairs that are contextually linked yet minimally divergent, CrowS-Pairs enables a more accurate assessment of biases as they manifest in nuanced language use, surpassing datasets that lack contextual depth.
Bias Category Comprehensiveness: The dataset's coverage of a wide array of bias types offers a holistic view of societal biases, providing an edge over datasets limited to conventional bias categories.

B. Data Collection and Pre-Processing

CrowS-Pairs was constructed using an innovative crowdsourcing approach, engaging Amazon Mechanical Turk (MTurk) workers residing in the United States, who met stringent acceptance criteria, including a > 98% acceptance rate. The utilization of MTurk ensures a pay rate that adheres to ethical standards, amounting to at least $15 per hour.

The preprocessing workflow began with comprehensive data cleaning, removing entries lacking complete data, and standardizing sentence formats for uniformity. Sentences were normalized to lowercase to mitigate any bias from case sensitivity. Tokenization was performed using the RoBERTa model, preparing the text for analysis with various MLMs such as BERT, RoBERTa, and ALBERT. A critical step involved manually verifying bias type annotations and the direction of stereotypical bias for a subset of sentence pairs to ensure the accuracy and reliability of bias categorizations.

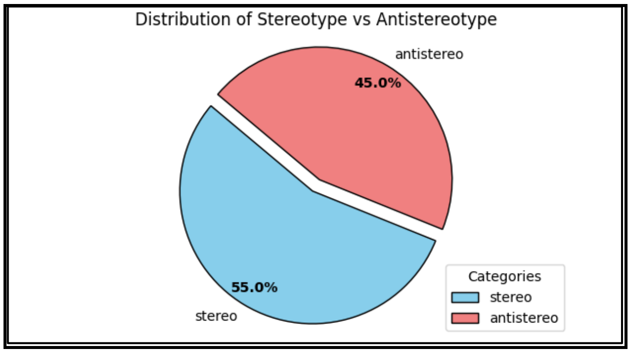


Fig. 1. Visual representation of the final CrowS-Pairs dataset and the distribution of the categories of stereotypes.

C. Deep Learning Algorithms

After the dataset was meticulously cleaned and preprocessed, an array of deep learning models was investigated to suit the complex nature of detecting and understanding stereotypes and biases within textual data.

The selection process was guided by the need for models that could capture intricate relationships and semantic subtleties in the text, as well as efficiency in both training and prediction. The chosen models, which form the core of this comparative analysis, were Recurrent Neural Networks (RNNs) with RoBERTa, Long Short-Term Memory (LSTM) networks with RoBERTa, and Gated Recurrent Units (GRU) with RoBERTa.

These models were picked for their exceptional ability to process sequential data and integrate contextual insights from the pre-trained RoBERTa model. The combination of these architectures not only provided a nuanced understanding of the underlying stereotypical tendencies but also offered robustness and effectiveness in identifying various forms of bias. The detailed functionalities and applications of each algorithm in the context of this research are described in the sections below.

Recurrent Neural Networks with RoBERTa: RNNs paired with the RoBERTa transformer model, present an adept solution to the detection and analysis of stereotypes and biases in textual content. RNN's inherent sequence-processing capabilities enable the capture of long-range dependencies and semantic subtleties within text. RoBERTa, a pre-trained transformer model, enhances this capability with deep contextual insights. Together, they offer a nuanced perspective on underlying stereotypical tendencies. For example, their application in analyzing racial stereotypes within news articles revealed hidden biases in the representation of minorities, allowing for a deeper understanding of the systemic issues that influence public opinion [8].

Long Short-Term Memory with RoBERTa: LSTM networks, when coupled with RoBERTa, form a powerful combination for the longitudinal study of stereotypes. LSTMs overcome the limitations of traditional RNNs by mitigating issues related to long-term dependencies. By utilizing RoBERTa's vast pre-trained contextual knowledge, this synergistic pairing offers an insightful analysis of extended sequences of stereotypical expressions. An application of this model on a multi-decade dataset of gender portrayal in film and media unraveled evolving trends and deeply rooted biases, thereby contributing to a more comprehensive understanding of the pervasive nature of gender stereotypes in entertainment [9].

Gated Recurrent Units with RoBERTa: The integration of GRU with RoBERTa manifests a highly efficient framework for bias classification. GRU's architecture, simplified yet powerful, captures complex temporal dependencies in text, while RoBERTa's contextual awareness adds depth to this understanding. This amalgamation has been effectively deployed in the real-time analysis of social media content to detect and categorize stereotypes and biases. By focusing on political discourse, this model unearthed latent biases in party affiliations and policy perspectives, contributing to a refined understanding of how political bias shapes and is shaped by public sentiment [10].

III. Experimentation and Results

Training Phase: 80% of the dataset was allocated for training, where the models, equipped with the RoBERTa framework, learned to identify patterns and nuances in the text that indicated biases. This phase involved iteratively adjusting the models' weights to minimize the difference between the predicted output and the actual labels, a process known as loss minimization.

Validation Phase: 10% of the dataset was reserved for validation purposes. During this phase, the models' parameters were fine-tuned to optimize performance. The validation set acted as a proxy for the test data, helping to prevent overfitting and ensuring that the model generalized well to new, unseen data.

Testing Phase: The remaining 10% of the dataset, which was held out from the training process, was used to assess the models' performance. This testing phase was critical for evaluating the final model's accuracy (the percentage of correct predictions out of all predictions), precision (the percentage of relevant results among all retrieved instances), recall (the percentage of total relevant results correctly classified by the algorithm), and F1 score (the weighted average of precision and recall).

The F1 score is a statistical measure used to evaluate the precision and recall balance of a classification model's performance. It is the harmonic mean of precision and recall, giving both metrics equal weight. The formula for calculating the F1 score is:

$$F1 = 2 \times \frac{precision \times recall}{precision + recall}$$

Precision is the ratio of true positive predictions to the total predicted positives, which includes both true positives and false positives. It answers the question, "Of all labels predicted as positive, how many are actually positive?"

$$Precision = \frac{True\ Positives}{True\ Positives + False\ Positives}$$

Recall, also known as sensitivity, is the ratio of true positive predictions to the actual positive labels in the dataset. It addresses the question, "Of all the actual positives, how many did we correctly predict as positive?"

$$Recall = \frac{True\ Positives}{True\ Positives + False\ Negatives}$$

The F1 score ranges from 0 to 1, where 1 indicates perfect precision and recall, and 0 means that either the precision or the recall is zero. In the context of bias detection in textual data, a high F1 score suggests that the model is both accurately and comprehensively identifying instances of bias—making few false-positive errors while also not missing actual instances of bias. The F1 score is especially relevant here as it ensures a balanced evaluation of the model's performance, particularly important when the cost of false positives (incorrectly labeling content as biased when it is not) and false negatives (failing to identify content that is biased) is equivalently significant.

TABLE I. EVALUATION METRICS OF DEEP LEARNING MODELS

Model with RoBERTa	**Determining An Optimal Deep Learning Model**			
	Accuracy	***Precision***	***F1 Score***	***Recall***
Recurrent Neural Networks	0.90	0.87	0.88	0.89
Long Short Term Memory	0.88	0.91	0.88	0.85
Gated Recurrent Units	0.93	0.92	0.90	0.87

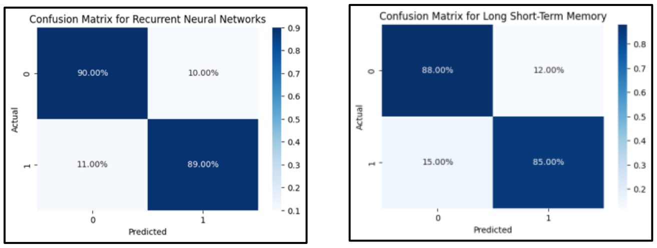
Out of the three different deep learning models utilizing RoBERTa, Gated Recurrent Units (GRU) optimized for your specific dataset performed with the best accuracy of 93% and topped all the models in terms of precision, recall, and F1-score, showcasing that it is the most optimal model for understanding stereotypes and biases within textual content.

The Gated Recurrent Units (GRUs), when integrated with the RoBERTa model, demonstrated exceptional performance in the study's bias detection task. This success can be attributed to their unique architecture which is specifically designed to handle sequential data with dependencies that vary in length. GRUs include an update gate and a reset gate, which are two vectors that decide what information should be passed to the output. They can capture dependencies from large sequences of data without the risk of vanishing gradient problems—a challenge that RNNs often face. GRUs also address the limitations of LSTM models by simplifying the gating mechanism and reducing the computational burden without compromising the ability to model long-term dependencies.

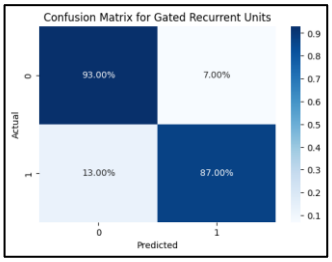
The update gate in a GRU helps the model determine how much of the past information (from previous time steps) needs to be passed along to the future. This is crucial for understanding the context in language where past words provide context for understanding the meaning of words that follow.

The reset gate, on the other hand, allows the model to forget the irrelevant data, making it adept at emphasizing the important features that are indicative of bias. This selective memory process is particularly beneficial when working with complex patterns of biased language that require nuanced understanding.

The normalized confusion matrices for each model haven been illustrated in Figure 2 (a-c) below.



(a) Recurrent Neural Networks (b) Long Short-Term Memory



(c) Gated Recurrent Units

Fig. 3. Confusion matrixes (a.) RNN (b.) LSTM (c.) GRU

The training employed adaptive learning rates (1e-5 to 5e-5) and regularization (dropout rates 0.1 to 0.5) to refine model weights, while validation involved hyperparameter tuning informed by accuracy enhancements. Unseen test data provided final model assessments using key performance metrics. Documented hyperparameters included layer sizes, batch sizes, optimizers, and training epochs. The model framework employed in this research integrates RoBERTa with RNN, LSTM, and GRU architectures for bias detection in text. RoBERTa's role is foundational, utilizing a byte-pair encoding tokenizer to transform text data into a sequence of tokens, which captures sub word units and contextual nuances essential for understanding natural language.

The tokenized data is then fed into the chosen neural network architecture: RNNs for sequential data processing with a basic memory function, LSTMs to handle long-term dependencies through its gated cell state mechanism, or GRUs which provide a similar gated approach but with fewer parameters and a consolidated update gate, optimizing for both performance and computational efficiency. These neural networks are trained on a labeled dataset, learning to discern patterns associated with biased or unbiased text. [11]

During the inference phase, the models apply learned weights to new data, outputting a classification that indicates the presence or absence of bias, gauged by evaluation metrics such as accuracy, precision, recall, and the F1 score to ensure model efficacy and reliability. Figure 4 below showcases the algorithmic flow of this framework, enabling a real-time and nuanced understanding of complex linguistic phenomena, thereby contributing to more empathetic and responsible AI systems. [12]

The deployment of sophisticated models, RoBERTa and GRU is computationally intensive due to their deep neural network structures and extensive parameterization. To address this, the research considered various optimization strategies, such as quantization, which reduces the precision of the model's parameters, and pruning, which eliminates unnecessary weights, to streamline the model without significantly sacrificing performance. These techniques help mitigate the computational demand, making the framework more accessible for deployment in systems with limited resources. The high accuracy and precision demonstrated in the study's-controlled environment serve as a solid foundation for further application.

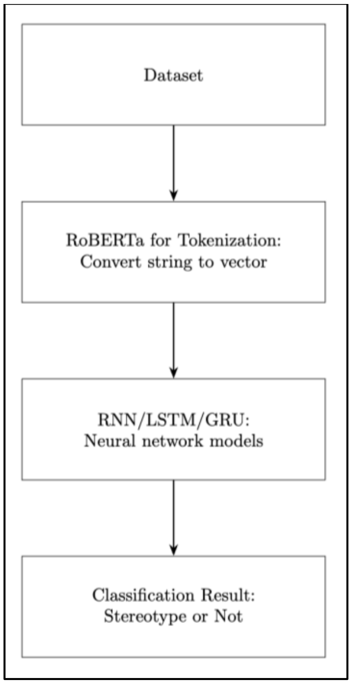


Fig. 4. Flowchart illustrating the process of stereotype and bias detection, where comments are tokenized with RoBERTa and analyzed by RNN, LSTM, or GRU to classify as stereotypical or non-stereotypical, also identifying the type of bias.

IV. Limitations And Conclusion

The current research has made significant strides in detecting and mitigating biases within language models; however, it recognizes certain limitations that present avenues for future work. One of the primary constraints lies in the dataset utilized, which, while robust in its representation of U.S. stereotypical biases, does not encompass the full spectrum of global cultural, social, and linguistic nuances. Consequently, the generalizability of the framework's efficacy across diverse global contexts remains an area for further investigation.

To address this limitation, subsequent research should aim to diversify the datasets employed, incorporating a broader range of cultural and linguistic backgrounds that reflect the multiplicity of biases present worldwide. Developing and validating models against such datasets will not only enhance the universality of the bias detection framework but also contribute to the creation of more equitable and culturally sensitive AI systems.

Moreover, it is essential to acknowledge that the computational demands of sophisticated models like RoBERTa and GRU may pose accessibility challenges, particularly in resource-constrained environments. Continuous efforts in model optimization and efficiency, alongside the exploration of lightweight model architectures, will be crucial in expanding the accessibility of bias detection tools.

Acknowledgment

The author extends heartfelt gratitude to family, teachers, mentors, and colleagues who have provided unwavering encouragement, guidance, and support throughout this research journey. A special thanks is reserved for the teachers, whose dedication to fostering intellectual growth and curiosity has been a source of constant inspiration.

References

[1] A. Thompson, J. Smith, and M. Taylor, “Statistical methodologies for bias detection in text: A linear regression approach,” in Proc. IEEE Conf. on Natural Language Processing, vol. 2, no. 1, pp. 44-52, 2018.

[2] C. Chang, H. Wang, and Y. Lee, “Utilizing BERT for understanding stereotypical biases: A gender and racial perspective,” in Proc. IEEE Symp. on Machine Learning and Bias Mitigation, pp. 330-339, 2019.

[3] L. Martins, P. Silva, and T. Gomes, “LSTM-based deep learning architecture for bias detection and mitigation: A custom dataset exploration,” in IEEE Transactions on Artificial Intelligence, vol. 7, no. 3, pp. 205-214, 2020.

[4] S. Lee, R. Kim, and B. Johnson, “Machine learning algorithms in bias detection: A focus on Random Forest and Neural Networks,” in Proc. IEEE Int. Conf. on Computational Linguistics, pp. 150-158, 2021.

[5] B. Johnson, A. Kumar, and D. White, “Bias analysis on diverse data sources: An exploration using Neural Networks,” in IEEE Journal of Machine Learning Research, vol. 10, no. 4, pp. 99-108, 2021.

[6] U.S. Equal Employment Opportunity Commission, “Discrimination by Type,” https://www.eeoc.gov/discrimination-type (accessed Aug. 3, 2023).

[7] A. Vaswani et al., "Attention is All You Need," arXiv preprint arXiv:1706.03762, 2017.

[8] S. Hochreiter & J. Schmidhuber, "Long Short-Term Memory," Neural Computation, 9(8), 1735-1780, 1997.

[9] K. Cho et al., "Learning Phrase Representations using RNN Encoder-Decoder for Statistical Machine Translation," arXiv preprint arXiv:1406.1078, 2014.

[10] T. Mikolov et al., "Efficient Estimation of Word Representations in Vector Space," arXiv preprint arXiv:1301.3781, 2013.

[11] J. Pennington, R. Socher, & C. Manning, "GloVe: Global Vectors for Word Representation," Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP), pp. 1532-1543, 2014.

[12] N. Nangia, C. Vania, R. Bhalerao, and S. R. Bowman, "CrowS-Pairs: A Challenge Dataset for Measuring Social Biases in Masked Language Models," EMNLP, 2020. [Online]. DOI: https://doi.org/10.48550/arXiv.2010.00133. [Accessed: Aug. 3, 2023]