Technical Report

CS5079

Applied Artificial Intelligence

Assignment I: Task II on Evaluation of Different Research Papers

Authors:
Doroteya Stoyanova
Margarita Radeva
Zahary Kwidzinsky
Jakub Pitula
Karol Hetman
Tom Utting

December 13, 2024

Contents

1 Do you MIND? Reflections on the MIND dataset for research on diversity in news recommendations

1.1 Context

This paper by Sanne Vrijenhoek provides a thorough analysis of the MIND dataset composed of items from MSN News, which is currently the largest open-source resource for training and evaluating news recommender systems. The primary focus of the study is to assess the dataset's suitability for studying news diversity, an increasingly critical issue given concerns about filter bubbles and echo chambers in online news consumption.

1.2 Problems

The paper identifies several key limitations with one major issue being the lack of complex metadata within the dataset. Specifically, it lacks annotations for deeper normative diversity metrics such as stance detection or ideological balance. [13]. This absence of comprehensive metadata makes it challenging to accurately measure news diversity, a key element in fostering balanced and inclusive news ecosystems [14].

Another concern is the over-representation of "soft news" in the dataset, described as entertainment and celebrity stories [10]. This comes due to MSN News' anonymity to article aggregation, which posits the possible dominance of soft news contributing to skewed experiments designed to assess diversity. Additionally, the dataset's limited validation set, which only includes data from **a single day (November 1, 2019)**, significantly limits its utility for examining longer-term trends in news diversity. Such a narrow time frame restricts the ability to understand evolving patterns in news consumption and recommendation, although the whole dataset spans over 6 weeks

1.3 Contributions

The paper offers a few notable contributions to the field. It provides a detailed analysis of news type distribution at various stages of the recommendation pipeline, revealing how content distribution evolves as the recommender system processes the data. Furthermore, the paper compares different neural recommender algorithms, demonstrating how variations in algorithmic approaches can lead to differing outcomes, even when accuracy metrics are similar. Another important contribution is the investigation of position bias, exploring how the placement of articles in recommendation lists can influence the diversity and type of content recommended. This analysis highlights the potential biases that can arise in recommender systems and emphasizes the importance of addressing these biases to ensure more equitable and diverse news delivery.

1.4 Discussions and Evaluation

In this section, we will evaluate the aforementioned paper.

1.4.1 Type of Paper

This paper is an empirical research study focusing on the evaluation of the MIND dataset in the context of news recommender systems [11]. It provides both qualitative and quantitative analyses of the dataset's strengths and limitations, offering insights into how well

the dataset supports research into news diversity. This positions the paper as a valuable resource for understanding the implications of biases in news recommendation systems [4].

1.4.2 Research Question Evaluation

The main focus of the paper is on examining if the MIND dataset is appropriate for studying news diversity, as well as the wider impact of biases in recommendation systems on filter bubbles and echo chambers. The article systematically explores these inquiries by analyzing how news categories are distributed at different recommendation process stages and by evaluating results from advanced neural recommender algorithms like LSTUR(Long- and Short-term User Representations) and NRMS(Neural News Recommendation with Multi-Head Self-Attention) [2]-[12].

The results offer valuable information on how dataset qualities and algorithm selections affect diversity results. As an example, the research shows a notable surplus of light news and a restricted amount of serious news in suggestions. This prejudice stems in part from the original makeup of the dataset and is worsened during the selection of candidates and the processing by algorithms. Further examination of position bias reveals how the arrangement of articles in recommendation lists can impact user engagement and the variety of content consumed.

Although the paper effectively answers the research questions, it could have had a greater influence by suggesting practical solutions to reduce these biases. For instance, the writers could have examined techniques for enhancing the metadata of the dataset to allow for more detailed diversity assessments or proposed tactics for guaranteeing equal inclusion of hard and soft news genres. This limitation hinders the research's usefulness in enhancing upcoming recommender systems or datasets.

1.4.3 Methodological Realism

The approach utilized in the paper is strong and based on well-established methods. The analysis gains credibility from utilizing the MIND dataset, which captures authentic user interactions on MSN News. The research uses commonly used neural recommender algorithms and metrics to guarantee the results are relevant and realistic.

Yet, the extent of the study is limited by the validation set that incorporates information from just one day. The short timeframe restricts the analysis of user behaviour and algorithm performance over the long term. Adding longitudinal data to the dataset would offer a more thorough insight into how recommender systems impact diversity over a period of time.

Furthermore, using MIND as the only benchmark dataset limits the ability to apply the findings more broadly. The conclusions of the paper would have been stronger if they had discussed how the results could be relevant to other datasets or recommendation contexts. Nevertheless, this restriction does not lessen the importance of its impact on researching diversity in news recommendation systems.

1.4.4 Conclusion

The research enhances our comprehension of diversity challenges in news suggestions through the examination of biases and algorithmic actions, setting the stage for forthcoming studies on more equitable recommender systems. Concrete solutions and an

expanded examination of datasets and temporal dynamics could enhance its practical significance. In spite of these constraints, it provides valuable insights into how datasets, algorithms, and diversity outcomes in news recommendations interact.

2 Multimodal Composite Association Score: Measuring Gender Bias in Generative Multimodal Models

2.1 Context

Abhishek Mandal, Susan Leavy, and Suzanne Little's study delves into the significant problem of gender bias in sophisticated multimodal AI models like DALL-E [9] and Stable Diffusion [5]. These models, which can create elaborate images based on written instructions, have gained significant importance in the field of content generation. Nevertheless, they carry forward and frequently enhance societal prejudices found within their extensive training datasets sourced from the internet [6]. This research explores the increasing worry about how biases appear in multimodal systems and the constraints of current bias detection techniques, which are usually created for less complex, single-modality models such as convolutional neural networks (CNNs). The paper stresses the importance of implementing strong and scalable methods to identify and address bias effectively in both text and image modalities due to the intricate nature of multimodal models with different data types and stages.

2.2 Problems

The main issue addressed in this research is the insufficiency of current bias evaluation systems for multimodal generative models. Conventional techniques prioritize systems that utilize only one stage and one type of data, frequently not adequately handling the intricacies of generative models that integrate text and images through multiple stages. This deficiency leads to a limited range of tools for identifying and measuring bias in systems such as DALL-E 2, where biases can spread or intensify during various data processing phases [1]. Furthermore, the paper underscores the difficulty of conducting manual bias audits because of the unlimited range of inputs that these models can accommodate. Another important problem is the lack of measurable metrics that can give valuable information about the distribution of biases within the model architecture across various modalities and stages.

2.3 Contributions

The article presents a new measure called the Multimodal Composite Association Score (MCAS) for evaluating gender bias in multimodal generative models [7]. MCAS expands on the abilities of the Word Embedding Association Test (WEAT) to evaluate bias in text and image embeddings by incorporating multimodal data [3]. The framework is made up of four component scores: Image-Image Association Score (IIAS), Image-Text Prompt Association Score (ITPAS), Image-Text Attributes Association Score (ITAAS), and Text-Text Association Score (TTAS) [8]. These parts assess bias at various points in the model, providing a complete perspective on how gender stereotypes appear in produced material.

The writers test the effectiveness of MCAS by using it on DALL-E 2 and Stable Diffusion, examining four types of concepts: jobs, sports, items, and environments. The findings show steady gender stereotypes, with more pronounced biases seen in Stable Diffusion than in DALL-E 2. This in-depth assessment shows how MCAS can effectively detect and correct bias in multimodal AI systems.

2.4 Discussions and Evaluation

2.4.1 Type of Paper

This study introduces and validates a novel metric designed to identify bias in multimodal generative AI models, classifying it as a methodological research paper [**<empty citation>**]. It seamlessly combines theoretical innovation with empirical validation, demonstrating both originality and practical relevance. By addressing a significant gap in the current literature, it contributes a quantifiable and scalable framework for measuring gender bias, positioning itself as a pivotal advancement in the field.

2.4.2 Research Question Evaluation

The research queries are well-established and extremely relevant in the current landscape of AI advancement. The crucial challenge of measuring and quantifying bias in multimodal generative systems like DALL-E 2 and Stable Diffusion is the central question due to their rapid adoption. The authors investigate this question by creating the MCAS framework, which assesses prejudice in various facets and phases of these models. The experimental results reveal consistent gender stereotypes across both models. For example, "CEO" is strongly associated with men, while "beautician" and "housekeeper" are associated with women [8].Yet, the paper's concentration on bias against only two genders restricts its wider relevance. In the future, further studies may expand on the framework to cover additional aspects of bias, like race, age, and intersectionality, for a more thorough examination.

2.4.3 Methodological Realism

The method is strong and well-planned, utilizing the popular WEAT framework and modifying it for multimodal situations. Utilizing established generative models such as DALL-E 2 and Stable Diffusion enhances the credibility of the results, showcasing the practicality of MCAS. The scalability and modular structure of the framework enable in-depth examination of bias at various points, offering valuable perspectives for developers and researchers. Nevertheless, there are some restrictions in the research. Large-scale image generation and feature extraction require high computational resources, which can be a barrier to adoption in resource-constrained environments. Furthermore, relying on a few models and categories may cast doubt on the generalizability of the results. Broadening the validation to encompass a wider variety of models and bias dimensions would enhance the framework's relevance.

2.4.4 Conclusion

In general, the paper provides a substantial contribution to the area of detecting bias in AI by presenting a thorough and adaptable framework for assessing gender bias in multimodal generative systems. Although it effectively answers its research questions

and offers valuable insights, the narrow concentration on binary gender bias and the restricted range of examined models indicate potential areas for further research.

References

[1] Tosin Adewumi et al. "Fairness and bias in multimodal ai: A survey". In: *arXiv preprint arXiv:2406.19097* (2024).

[2] Mingxiao An et al. "Neural news recommendation with long-and short-term user representations". In: *Proceedings of the 57th annual meeting of the association for computational linguistics*. 2019, pp. 336–345.

[3] Gemma Bel-Enguix et al. "Wan2vec: Embeddings learned on word association norms". In: *Semantic Web* 10.6 (2019), pp. 991–1006.

[4] Saumya Bhadani. "Biases in recommendation system". In: *Proceedings of the 15th ACM conference on recommender systems*. 2021, pp. 855–859.

[5] Ali Borji. "Generated faces in the wild: Quantitative comparison of stable diffusion, midjourney and dall-e 2". In: *arXiv preprint arXiv:2210.00586* (2022).

[6] Sasha Luccioni et al. "Stable bias: Evaluating societal representations in diffusion models". In: *Advances in Neural Information Processing Systems* 36 (2024).

[7] Abhishek Mandal, Susan Leavy, and Suzanne Little. "Generated Bias: Auditing Internal Bias Dynamics of Text-To-Image Generative Models". In: *arXiv preprint arXiv:2410.07884* (2024).

[8] Abhishek Mandal, Susan Leavy, and Suzanne Little. "Multimodal composite association score: Measuring gender bias in generative multimodal models". In: *arXiv preprint arXiv:2304.13855* (2023).

[9] Gary Marcus, Ernest Davis, and Scott Aaronson. "A very preliminary analysis of DALL-E 2". In: *arXiv preprint arXiv:2204.13807* (2022).

[10] Carsten Reinemann et al. "Hard and soft news: A review of concepts, operationalizations and key findings". In: *Journalism* 13.2 (2012), pp. 221–239.

[11] University of La Verne. *Empirical Articles*. https://laverne.libguides.com/empirical-articles. Accessed: December 13, 2024. 2024.

[12] Chuhan Wu et al. "Neural news recommendation with multi-head self-attention". In: *Proceedings of the 2019 conference on empirical methods in natural language processing and the 9th international joint conference on natural language processing (EMNLP-IJCNLP)*. 2019, pp. 6389–6394.

[13] Fangzhao Wu et al. "Mind: A large-scale dataset for news recommendation". In: *Proceedings of the 58th annual meeting of the association for computational linguistics*. 2020, pp. 3597–3606.

[14] Marcia Lei Zeng and Jian Qin. *Metadata*. American Library Association, 2020.