

Project ID:

24-25J-195

1. Topic (12 words max)

ResAI Toolkit: Framework for cross-modality bias detection and mitigation

2. Research group the project belongs to

Centre of Excellence for AI (CEAI)

3. Research area the project belongs to

Artificial Intelligence (AI)

4. If a continuation of a previous project:

Project ID	N/A
Year	N/A

5. Brief description of the research problem including references (200 – 500 words max) – references not included in word count.

Artificial Intelligence (AI) has significantly transformed various sectors by enhancing operational efficiency, personalizing user experiences, and improving decision-making processes. However, the integration of AI systems into these domains raises critical ethical concerns, particularly regarding bias. Bias in AI models can amplify existing disparities and lead to unequal outcomes across different applications. As a result, the identification and mitigation of biases in AI systems are crucial to ensuring fairness and reliability.

Current bias detection and mitigation metrics are often not universally applicable across different domains and modalities. These metrics may lack the precision or relevance needed for specific types of data, making it challenging to achieve consistent and fair AI outcomes. For example, a bias detection metric effective for text data may not be suitable for image, audio, or video data due to the distinct nature of each data type.

Currently available metrics for detecting and mitigating bias in AI may not be very accurate or applicable in all domains. Existing metrics such as Demographic Parity, Equalized Odds, and Disparate Impact, which are commonly used for text datasets, often lack support for different types of datasets, making it difficult to apply them universally. For instance, the “Quantifying Bias in Text-to-Image Generative Models” paper employs metrics like Distribution Bias, Jaccard Hallucination, and Generative Miss Rate for text-to-image datasets.[5] The “AI Fairness 360” toolkit uses Statistical Parity Difference (SPD) and Disparate Impact(DI) for both text and image datasets, though these metrics do not account for true positive and false positive rates, potentially overlooking fairness issues in classification accuracy.[4]

The unique characteristics of each data modality present specific challenges for bias detection and mitigation. Text data is discrete, image data is spatial, audio data is temporal, and video data combines both spatial and temporal elements. Developing metrics that can accurately detect bias across these varied representations is complex. For instance, embeddings used in vision-language models demonstrate a “modality gap,” where embeddings for images and text occupy separate regions in the embedding space, leading to different biases that must be addressed separately.[1]

Feature extraction processes also vary significantly across modalities, complicating the application of a single bias detection method universally. For example, word embeddings for text, pixel values for images, and mel-frequency cepstral coefficients (MFCCs) for audio each require specific methods for effective feature extraction and bias detection. The modality gap in embeddings underscores the need for modality-specific adjustments to improve performance and fairness.[1]

The process of labeling and annotating data differs across modalities, impacting the effectiveness of bias detection metrics. Text annotations may involve tagging parts of speech, while image annotations might include bounding boxes or segmentation masks. Audio annotations could involve transcriptions or speaker identification, and video annotations require frame-by-frame labeling of actions or objects. Aligning annotations across different modalities is challenging, as shown by the detailed annotations needed for both RGB and thermal images in pedestrian detection [2] [3].

Bias manifests differently across modalities, requiring tailored evaluation metrics. Gender bias in text might be detected through pronoun usage, while in images it could relate to demographic representation in certain contexts. In audio, biases may arise from accents or dialects, and in video, from stereotypical portrayals of individuals or groups. Metrics must account for these modality-specific biases to ensure comprehensive detection and mitigation [3].

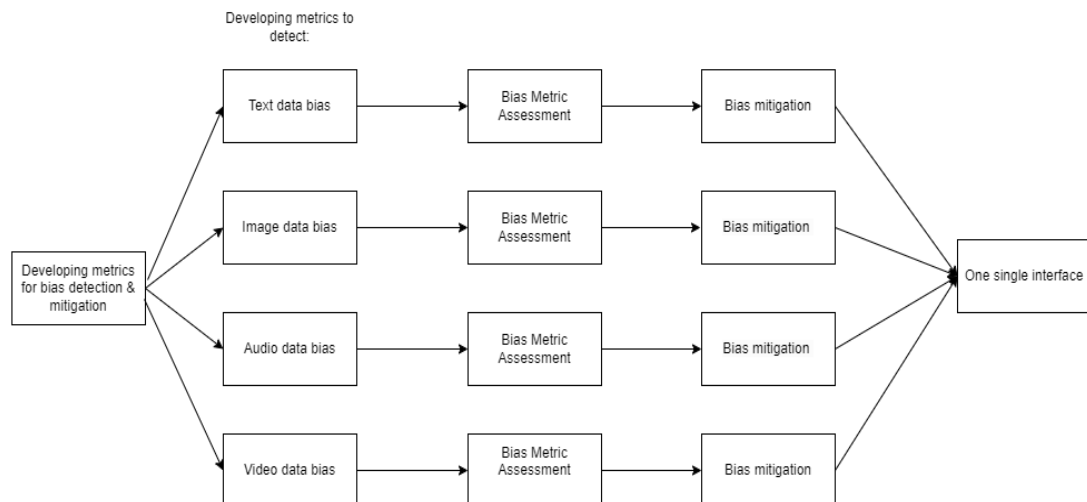
Given these challenges, there is a pressing need to develop new metrics or enhance existing ones to effectively detect and mitigate bias across diverse data formats, such as text, image, audio, and video. This research aims to address this gap by creating a comprehensive framework for cross-modality bias detection and mitigation. By incorporating robust metrics that can be applied to various AI models and datasets, this study seeks to promote ethical and socially responsible AI systems.

References

1. Two Effects, One Trigger: On the Modality Gap, Object Bias, and Information Imbalance in Contrastive Vision-Language Representation Learning, arXiv, 2023 ([ar5iv](#)).
2. MSCoTDet: Language-driven Multi-modal Fusion for Improved Multispectral Pedestrian Detection, arXiv, 2024 ([ar5iv](#)).
3. Causal Mode Multiplexer: A Novel Framework for Unbiased Multispectral Pedestrian Detection, arXiv, 2024 ([ar5iv](#)).
4. Rachel K. E. Bellamy, Kuntal Dey, Michael Hind, Samuel C. Hoffman, Stephanie Houde, Kalapriya Kannan, et al. "AI Fairness 360: An Extensible Toolkit for Detecting, Understanding, and Mitigating Unwanted Algorithmic Bias," arXiv preprint arXiv:1810.01943, Oct. 2018. Available: <https://github.com/ibm/aif360>.

5. J. Vice, N. Akhtar, R. Hartley, and A. Mian, "Quantifying Bias in Text-to-Image Generative Models," arXiv preprint arXiv:2312.13053, 2023.
6. L. H. S. K. S. L. T. B. Booth, "Bias and Fairness in Multimodal Machine Learning: A Case Study of Automated Video Interviews," *ResearchGate*, Oct. 2021. [Online]. Available: https://www.researchgate.net/publication/355379791_Bias_and_Fairness_in_Multimodal_Machine_Learning_A_Case_Study_of_Automated_Video_Interviews.
7. R. A. J. C. M. Schmitz, "Bias and Fairness in Multimodal Machine Learning," *arXiv*, May 2022. [Online]. Available: <https://arxiv.org/abs/2205.08383>.
8. S. A. M. J. F. A. Peña, "Bias in Multimodal AI: Testbed for Fair Automatic Recruitment," *ResearchGate*, Apr. 2020. [Online]. Available: https://www.researchgate.net/publication/340663849_Bias_in_Multimodal_AI_Testbed_for_Fair_Automatic_Recruitment.
9. L. H. S. K. S. L. T. B. Booth, "Bias and Fairness in Multimodal Machine Learning: A Case Study of Automated Video Interviews," *ResearchGate*, Oct. 2021. [Online]. Available: https://www.researchgate.net/publication/355379791_Bias_and_Fairness_in_Multimodal_Machine_Learning_A_Case_Study_of_Automated_Video_Interviews.
10. P. B. J. Wiśniewski, "fairmodels: A Flexible Tool For Bias Detection, Visualization, And Mitigation," *arXiv*, Feb. 2022. [Online]. Available: <https://arxiv.org/abs/2104.00507>.
11. R. K. E. Bellamy, K. Dey, M. Hind, S. C. Hoffman, S. Houde, K. Kannan, P. Lohia, J. Martino, S. Mehta, A. Mojsilovic, S. Nagar, K. Natesan Ramamurthy, J. Richards, D. Saha, P. Sattigeri, M. Singh, K. R. Varshney, and Y. Zhang, "AI Fairness 360: An Extensible Toolkit for Detecting, Understanding, and Mitigating Unwanted Algorithmic Bias," arXiv, 2018. [Online]. Available: <https://arxiv.org/abs/1810.01943>

6. Brief description of the nature of the solution including a conceptual diagram (250 words max)



Our solution focuses on developing metrics to detect and mitigate bias in AI systems by leveraging various dataset types such as images, text, video, and audio.

- **Data Collection:** Our system utilizes existing datasets containing images, videos, audio, and textual information from a variety of sources.
- **Bias Metrics Development:** For each dataset type, we develop specific metric to identify and quantify biases.
- **Bias Metric Assessment:** We employ our developed metrics to evaluate a diverse range of datasets encompassing images, text, audio, and video. We achieve this by intentionally introducing controlled modifications to these datasets. By observing how our metrics respond to these manipulations, we can verify their effectiveness in capturing bias across various data types. This comprehensive assessment provides robust confirmation of the metrics' ability to detect bias.
- **One single interface :** we will be presenting our metrics and mitigation algorithms using a single interface be it a library or a dashboard.

The interface empowers ML developers to build fair and ethical AI models. It achieves this by providing comprehensive explanations of detected biases in data and algorithms. This empowers developers to proactively address these issues, ultimately promoting fairness and accuracy in AI-based solutions.

7. Brief description of specialized domain expertise, knowledge, and data requirements (300 words max)

Our research on Responsible AI aims to develop metrics for identifying and reducing biases in AI programs, which can have a significant impact on a variety of industries. AI increases productivity and decision-making but also raises ethical concerns about bias, leading to inequality. We focus on creating and validating metrics across datasets, incorporating these into AI performance, and explaining known biases and their impact. Our goal is to build trust in AI technology, ensuring that it is technologically advanced that counts accountability in ethics to better contribute to equalizing outcomes.

Our team is also involved in the processes of leading research issues in responsible AI and unbiased and biased data sets. We leverage comprehensive frameworks and insights from prominent technology companies like Google, Microsoft, AWS, and PWC, among others, to guide our research.

Tools and Frameworks :

- TensorFlow: Know Your Data Tool
Available: [KYD - Documentation \(knowyourdata.withgoogle.com\)](https://knowyourdata.withgoogle.com/)
- Responsible AI Toolbox: Fairness Dashboard:
Available: responsible-ai-toolbox/docs/fairness-dashboard-README.md at [main · microsoft/responsible-ai-toolbox · GitHub](https://main-microsoft-responsible-ai-toolbox.github.io/)
- SageMaker Clarify: Enhances transparency and mitigation of bias in machine learning models.
Available: [Use SageMaker Clarify to explain and detect bias - Amazon SageMaker](https://aws.amazon.com/sagemaker/clarify/)
- AI Fairness 360: Open source toolkit can help you examine, report, and mitigate discrimination and bias in machine learning.
Available: [AI Fairness 360 \(ibm.com\)](https://aif360.github.io/)
- Fairlearn: A toolkit for assessing and improving fairness in AI
Available: [Fairlearn - A Python package to assess AI system's fairness - Microsoft Community Hub](https://fairlearn.org/)
- Holistic AI: Implement AI Governance Across your Enterprise Available: [Holistic AI - AI Governance Platform](https://holisticai.com/)

Data Requirements:

To conduct our research, we need different datasets, including images, text, video, audio, and we use multilevel instead of Binary classification, so we need such data. We get information from a publicly accessible database in, benchmark datasets previously used in similar analysis, and open-source datasets.

Text, Image, video, audio datasets:

<https://paperswithcode.com/datasets>

8. Objectives and Novelty

Main Objective The primary objective of this research is to develop a comprehensive framework for detecting and mitigating bias in multimodal data, encompassing text, image, audio, and video. While many studies have concentrated on bias in individual data types, this research aims to address bias across multiple modalities where we will compare several bias detection metrics across multimodal data.			
Member Name	Sub Objective	Tasks	Novelty
	Develop or improve metrics to measure text-based data bias. As existing statistical methods often fail to incorporate the contextual meaning of words and phrases, potentially missing subtle biases. Also, the existing methods are not intersectional as they solely focus on a single feature at a time.	<ul style="list-style-type: none"> • Generate or Improve metrics to detect bias in text data. • Develop a novel bias mitigation algorithm or a method to mitigate text data bias. • Test on multiple text-based multiclass classification models. 	<ul style="list-style-type: none"> • Developing/Improving a metric to measure bias in text-based data.

	Develop or improve a metrics to measure Image-based data bias, this will be done as statistical measurements currently can be used to identify high level bias missing out on subtle biases.	<ul style="list-style-type: none"> • Generate/Improve a metrics to detect bias. • Develop a novel bias mitigation algorithm to mitigate image-based data bias. • Test on multiple Image-based classification models. 	<ul style="list-style-type: none"> • Developing/Improving a metric to measure bias in image-based data.
	Develop metrics to measure video data bias as existing bias detection metrics are primarily used to detect bias in Tabular data leading to a lack of metrics solely focused on detecting bias in video data.	<ul style="list-style-type: none"> • Generate metrics to detect bias. • Develop an algorithm to mitigation algorithm to mitigate video-based data bias. • Perform tests on the across multiple datasets to check the accuracy of the new metric and the algorithm. 	<ul style="list-style-type: none"> • Developing a metric to measure bias in video-based data.

	<p>Develop metrics to measure audio data bias. Where the existing measures and tools use audio data processed to a tabular format, the developed metrics will focus solely on detection of bias in audio data.</p>	<ul style="list-style-type: none"> • Generate a metrics to detect bias. • Develop an algorithm to mitigation algorithm to mitigate audio-based data bias. • Perform tests on the accuracy of the new metric and the algorithm. 	<ul style="list-style-type: none"> • Developing a metric to measure bias in audio-based data.
--	--	---	--

9. Supervisor checklist

- a) Does the chosen research topic possess a comprehensive scope suitable for a final-year project?

Yes	√	No	
-----	---	----	--

- b) Does the proposed topic exhibit novelty?

Yes	√	No	
-----	---	----	--

- c) Do you believe they have the capability to successfully execute the proposed project?

Yes	√	No	
-----	---	----	--



- d) Do the proposed sub-objectives reflect the students' areas of specialization?

Yes	√	No	
-----	---	----	--

- e) Supervisor's Evaluation and Recommendation for the Research topic:

Challenging but a good research.

10. Supervisor details

	Title	First Name	Last Name	Signature
Supervisor	Dr.	Prasanna	Sumathipala	
Co-Supervisor	Ms.	Thisara	Shyamalee	
External Supervisor				
Summary of external supervisor's (if any) experience and expertise				

This part is to be filled by the Topic Screening Panel members.

Acceptable: Mark/Select as necessary

Topic Assessment Accepted	
Topic Assessment Accepted with minor changes (should be followed up by the supervisor)*	
Topic Assessment to be Resubmitted with major changes*	
Topic Assessment Rejected. Topic must be changed	

* Detailed comments given below

Comments

The Review Panel Details

Member's Name	Signature

***Important:**

1. According to the comments given by the panel, make the necessary modifications and get the approval by the **Supervisor** or the **Same Panel**.
2. If the project topic is rejected, identify a new topic, and follow the same procedure until the topic is approved by the assessment panel.