Project ID:24\_25J\_195

Project Title: RESAI Toolkit : Framework for cross modality bias detection

1. Introduction
   1. Background:

The rapid advancements in artificial intelligence (AI) and machine learning (ML) have greatly enhanced the capabilities of models in processing and understanding multimodal data, such as audio, video, text, and images. However, these systems are increasingly found to exhibit demographic biases, which can result in unfair outcomes, especially when models are applied to diverse user groups. These biases, which can be based on gender, race, age, and other demographic factors, pose significant challenges in ensuring fairness, equity, and inclusivity in AI systems. As AI and ML models are deployed in sensitive areas like healthcare, finance, and law enforcement, it becomes crucial to understand, detect, and mitigate such biases to avoid unintended discrimination and societal harm.  
  
In multimodal AI systems, bias detection is proven to be challenging as different data types display bias in different ways. As an example Gender and Accents of different people lead inaccurate results in speech recognition systems (Gorrostieta et al., 2020) [1] and in Image and Video data the bias is found in the skin tones of the person in such scenarios underrepresented groups, especially in facial recognition (Buolamwini & Gebru, 2018) [2] face bias . Text data represent historical bias and stereotypes in the language itself leading to discrimination of user of AI systems(Binns et al., 2021) [3]. Though there are many developments in the area it was observed that there is not one framework to detect bias across multiple modalities of data.  
  
Through this research we plan to address the gap in the area of multimodal bias detection frame by firstly creating set of bias detection metrics for each modality (audio, video, image, and text), and then combining them into a framework, this research aims to provide a comprehensive solution for cross-modal bias detection. The goal is to enable AI systems to more accurately identify and mitigate biases, to generate more fair outcomes for all demographic groups across multiple data types and applications.

* 1. Research Problem:

Bias Detection metrics focus on single modality data of Structured data leaving a notable gap in Cross-modality data bias detection.

* 1. Objectives:

(State the primary objectives of the research)

|  |  |  |
| --- | --- | --- |
| It Number | Objective | Objective number |
| IT21380914 | Developing a metrics for detecting Gender bias in RAW audio datasets used for model Building | 1 |
| IT21387562 | Develop a metric to detect contextual bias in image datasets by analyzing object sizes, spatial relationships, and scene-level contexts for gender classification. | 2 |
| IT21208294 | Developing Metrics for Detecting Gender Bias in Motion Patterns of Video Datasets Used for action recognition | 3 |
| IT21183690 | Develop a metric to detect bias in contextual word embeddings in coreference resolution. | 4 |

1. Data Exploration
   1. Data Collection

Data for all the objectives was collected through publicly available datasets to maintain the project’s ethical needs.

* 1. Dataset Description:

(Describe datasets, including sources, size, and key attributes)

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| Data source | Description | Resource | Size | Key attributes |
| Librispeech | Contains 1000 hours of audio of 16KHz English speech of people reading audio books in LibriVox project | V. Panayotov, G. Chen, D. Povey, and S. Khudanpur, "Librispeech: An ASR Corpus Based on Public Domain Audio Books," *2015 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, South Brisbane, QLD, Australia, 2015, pp. 5206–5210 [4] | 60 GB | Speaker IDs, Gender, Audio, Text transcriptions, Duration, Data splits, Language, Speaker Variability |
| Common voice | Contains Multilingual speech corpus with close to 18000 hours of transcribed speech. The dataset was created by Mozilla. | Mozilla Common Voice. (n.d.). "Common Voice Dataset," Mozilla Foundation. Available: <https://commonvoice.mozilla.org/en>. Accessed: Dec. 2, 2024​ [5] | 1.6TB | Audio file, Demographics,  Languages, Speaker Variability, Transcriptions, Recording Conditions |
| COCO dataset | The dataset consists of categories relevant to gender-related stereotypes (e.g., 'handbag,' 'sports ball,' 'umbrella'). | Tsung-Yi Lin, , Michael Maire, Serge Belongie, Lubomir Bourdev, Ross Girshick, James Hays, Pietro Perona, Deva Ramanan, C. Lawrence Zitnick, Piotr Dollár, "Microsoft COCO: Common Objects in Context," 2014.  [https://cocodataset.org/#home](https://cocodataset.org/%23home) | 18 GB | object detection,  captioning,  keypoints detection,  stuff image segmentation,  panoptic,  dense pose |
| UCF101 | The dataset consists of 13,320 video clips, which are classified into 101 categories commonly used for action recognition.. | Soomro, K., Zamir, A. R., & Shah, M. "UCF101: A Dataset of 101 Human Action Classes From Videos in The Wild," 2012.  Available: <https://www.crcv.ucf.edu/data/UCF101.php.> | 7.2  GB | It offers a diverse collection of 13,320 videos spanning 101 action categories, making it ideal for training and evaluating machine learning models. |
| WinoBias  Dataset | The dataset tests for gender biases by providing sentences with gendered pronouns and occupations. | J. Zhao, T. Wang, M. Yatskar, V. Ordonez, and K.-W. Chang, "Gender Bias in Coreference Resolution: Evaluation and Debiasing Methods," *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, New Orleans, LA, USA, Jun. 2018. Available. https://github.com/uclanlp/corefBias/tree/master/WinoBias/wino | 3MB | Gender bias analysis, coreference resolution, NLP evaluation |

* 1. Suitability Analysis
     1. Relevance to Individual Research Objectives:

(Explain how well each dataset aligns with your research problem and objectives)

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
|  | 1 | 2 | 3 | 4 |
| Data source 1 | x |  |  |  |
| Data source 2 | x |  |  |  |
| Data source 3 |  | x |  |  |
| Data source 4 |  |  | x |  |
| Data source 5 |  |  |  | x |

1. Methodology
   1. Data Preprocessing:

(Mention data transformation techniques done in each dataset for each objective.)

Ex:

Data Cleaning, Data Normalization, Data Standardization, Data Encoding (e.g., One-Hot Encoding, Label Encoding), Handling Missing Data (e.g., Imputation or Removal),

Data Aggregation, Feature Engineering, Outlier Detection and Handling, Data Scaling,

Data Discretization, Dimensionality Reduction (e.g., PCA), Date/Time Transformation,

Data Integration (Merging or Joining), Data Mapping, Data Type Conversion.

**Objective 1**

|  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- |
|  | Transformation technique | | | | | | | |
|  | Data type conversion | Data mapping |  |  |  |  |  |  |
| Data source 1 | X  {From flac to WAV format} | X  {Maps the audio file with the gender of the speaker} |  |  |  |  |  |  |
| Data source 2 | X  {From MP3 to WAV format} | X  {Maps the audio file with the gender of the speaker} |  |  |  |  |  |  |

**Objective 2**

|  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- |
|  | Transformation technique | | | | | | | |
|  | Transformation Technique | Data Cleaning | Data Type Conversion | Data Mapping | Feature Engineering | Dimensionality Reduction |  |  |
| Data source 3 | X  {Bounding box extraction, object segmentation, and depth map generation using YOLO and Depth-Anything v2.} | X  {Filtered 1,000 images based on gender-stereotypical object categories (e.g., "sports ball," "handbag"). Manually labeled genders for persons. Removed duplicate or irrelevant annotations (non-person objects not meeting criteria).} | X  {Converted segmentation masks and depth maps into usable formats (grayscale PNG).  } | X  {Mapped persons and objects with corresponding metadata (e.g., gender, size, and class labels).} | X  {Computed object sizes, relative distances (XY, Z, 3D), and linked metadata with scene embeddings.} | X  {PCA applied to reduce scene embedding dimensions.} |  |  |
|  |  |  |  |  |  |  |  |  |

**Objective 3**

|  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- |
|  | Transformation technique | | | | | | | |
|  | Subset selection | Data Cleaning | Frame Extraction | Frame Rate Normalization | Feature Engineering | Data Mapping |  |  |
| Data source 4 | Selected only relevant categories for further analysis.(eg: Diving) | Converted video formats (e.g., AVI to MP4) for compatibility with processing libraries. | Extracted and resized frames from videos to 640x480 resolution using OpenCV. | Standardized all videos to 30 FPS for consistent motion calculations. | Calculated motion magnitudes, directions, and speed using the Farneback optical flow algorithm. | Annotated visible persons in frames with inferred gender labels. |  |  |
|  |  |  |  |  |  |  |  |  |

**Objective 4**

|  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- |
|  | Transformation technique | | | | | | | |
|  | Data Cleaning | Data Type Conversion | Data Mapping |  |  |  |  |  |
| Data source 5 | X  {removed incomplete or irrelevant sentences (e.g., missing gendered pronouns or occupations).} | X  {Converted tokenized word lists into readable sentences using Python string manipulation.} | X  {Mapped pronouns (e.g., "he," "she") and occupations (e.g., "nurse," "doctor") to sentence metadata.} |  |  |  |  |  |

* 1. Scalability

(Assess whether each dataset size is sufficient and scalable for analysis.)

* Objective 1 :
  + Both the datasets used for analysis contain more demographics other than gender. Such notable demographics are Age of speaker, Ethnicity. Since these demographics are factors that face bias as well the datasets allow the analysis to be scaled outside of the Gender demographic.
  + The datasets contain respectively 1000 hours and 18000 hours of speech the data size is sufficient for the analysis.
* Objective 2:
* The dataset used for analysis (COCO) contains a wide variety of object categories, scenes, and contexts, making it scalable for research objective.
* The dataset contains over 330,000 annotated images, but for this project, a filtered subset of 1,000 images with gender-object stereotypes was selected. This is sufficient for analyzing spatial and contextual factors while maintaining computational feasibility.
* Objective 3:
* The UCF101 dataset, with its 101 action categories and 13,000 videos, provides a rich diversity of motion patterns and activities, making it a robust choice for analyzing gender bias. Its total size of 7.2 GB is manageable for processing and can scale to larger analyses when combined with enhanced computational resources.
* A subset of 151 videos from Diving category was selected for computational feasibility. This subset is sufficient for detailed analysis while ensuring representation of gender-based motion trends. The methodology remains scalable to include more categories or the full dataset.
* Objective 4:
* The dataset used for analysis (WinoBias) contains diverse sentence structures designed to evaluate gender bias in coreference resolution tasks. It is scalable to analyze biases in other NLP tasks or demographic dimensions.
* The dataset is divided into subsets based on Type 1 (simple resolution) and Type 2 (ambiguous resolution) scenarios, ensuring a diverse range of sentence structures. Each file’s 396 sentences provide a sufficient amount of data for analyzing gender bias while maintaining computational feasibility.
  1. Feature extraction

(Mention steps followed to evaluate the key features and attributes in datasets for each objective)

* Objective 1 :
  + The research is based on the gender bias therefore the feature considered in the dataset was the Gender identity of the speaker mapped with the audio file.
  + The gender of the speaker was initially mapped to the audio file name using the metadata files provided. As each audio file was named with the speakers ID with the help of the metadata provided under each speaker ID the gender mapping was done.
* Objective 2:

The research focuses on understanding how object size, distance, and scene context influence gender bias in datasets.

Key features extracted include:

* Gender identity of the person: This was manually labeled based on visible cues and mapped to each image using metadata.
* Object attributes: Bounding box coordinates, segmentation masks, and object sizes were extracted for measuring object-person relationships.
* Spatial features: Metrics such as normalized distances (XY, Z, and 3D) between persons and objects were calculated using bounding boxes and depth maps.
* Scene context: Scene embeddings and classifications were extracted using Places365 and CLIP and clustered to identify common themes across images.

Steps followed:

* Filtering and Labeling: Selected 1,000 images based on object categories and manually labeled genders for visible persons.
* Bounding Box Extraction: Used YOLO for detecting objects and persons, capturing their sizes and spatial relationships.
* Depth Map Generation: Used Depth-Anything V2 to calculate Z distances between objects and persons.
* Scene Context Extraction: Applied Places365 and CLIP to classify and embed scene-level contexts for each image.
* Objective 3:

Motion Magnitudes:

* Quantified motion intensity between consecutive frames using the optical flow algorithm.
* Averaged pixel-level magnitudes to compute frame-level motion intensity.

Motion Directions:

* Measured angular motion direction using cv2.cartToPolar for angles derived from optical flow.
* Computed mean motion direction for each frame.

Motion Speed:

* Calculated as the rate of change in motion magnitudes across consecutive frames.
* Grouped by video to maintain temporal consistency.

* Objective 4:
  + The research is based on the gender bias therefore the feature considered in the dataset was the contextual Embeddings of the pronouns (e.g., "he," "she") and occupations (e.g., "nurse," "doctor").
  + These embeddings capture the relationship between words in their specific sentence contexts, allowing us to dynamically analyze gender bias.

1. Modelling and Results

* Objective 1 : Since the research solely focuses on datasets used for model building no specific model was used in the project. However, the basic Analysis of the dataset was done to extract gender ratios of factors such as Voice activity, Amplitude, Frequency, Energy levels and Gender counts.
* Objective 2 : Since this research focuses on dataset-level analysis for bias detection, no specific predictive model was built. Instead, the analysis emphasized extracting contextual and spatial metrics to measure potential biases.

Key analytical steps include:

* + Calculating gender ratios across object-person interactions.
  + Computing relative sizes and distances (XY, Z, and 3D) between objects and persons.
  + Identifying scene-based trends by clustering similar environments.
* Objective 3 :

No predictive model was developed as the focus was on **dataset-level analysis** of motion patterns.

Extracted motion metrics (magnitudes, directions, and speed) were used to assess gender-specific differences.

Gender labels were inferred for visible persons, enabling demographic analysis.

* Objective 4: Used BERT (bert-base-uncased) as the contextual embedding model to generate token and sentence-level embeddings. Set the model to evaluation mode to ensure that embeddings are generated without updating model weights. Using Contextual embeddings extracts the pronouns and occupations.
  1. Key Insights:
* Objective 1: Key insights include the values of the factors Voice activity, Amplitude, Frequency, Energy levels and Gender counts of the audio files used. The below provided images of graphs are based on the Librispeech Dataset split of Train-360.

A graph of a number of people

Description automatically generatedA graph of different colored rectangles

Description automatically generatedA graph showing a number of people

Description automatically generated

* Objective 2 :
* Object-Person Relationships: Gendered patterns emerged, such as "bicycle" being more associated with males and "handbags" with females.

A graph of different colored squares

Description automatically generated

* Spatial Factors: Objects closer to persons had a higher impact on the association with gender.  
  A graph of a bar graph

  Description automatically generated

A graph of a bar chart

Description automatically generated

* Scene Contexts: Scene clustering revealed biases tied to specific settings, such as kitchens being more associated with females and sports fields with males.

A screenshot of a computer code

Description automatically generated

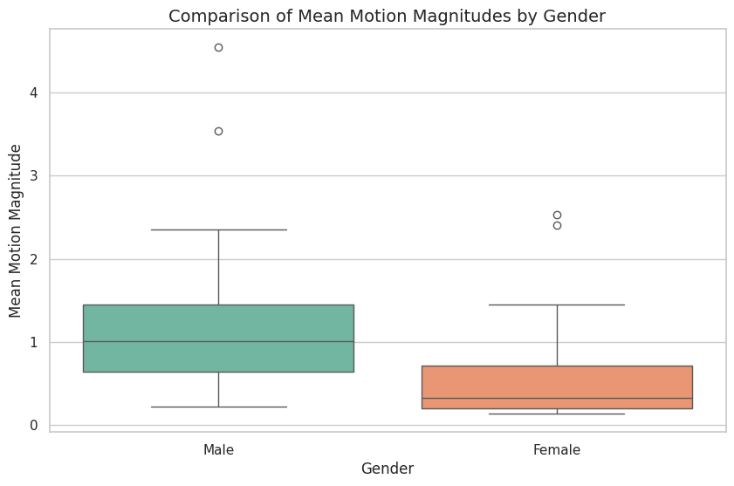
A list of sports activities

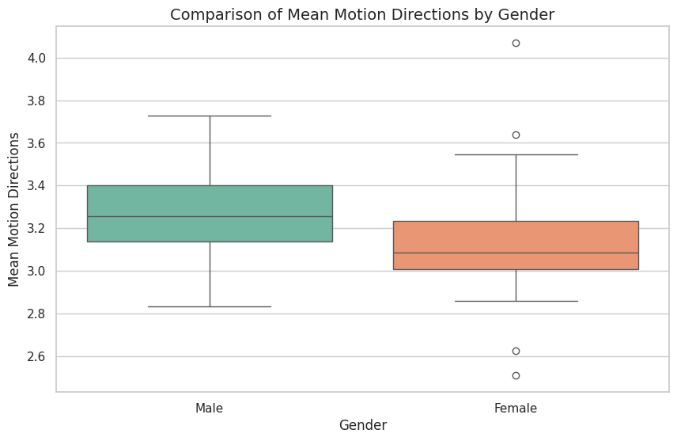
Description automatically generated

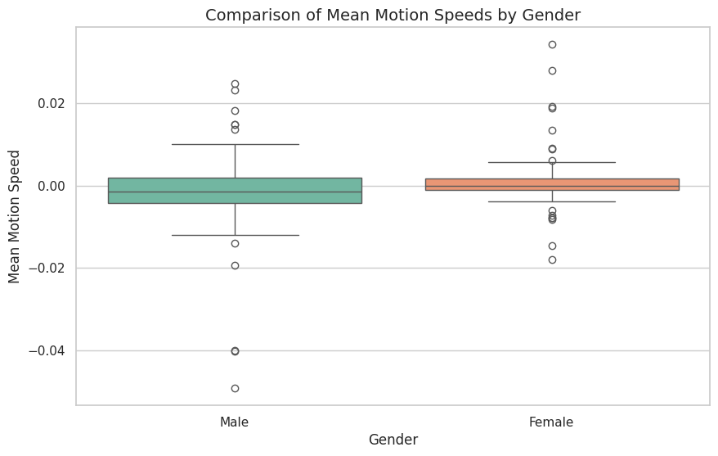
* Objective 3:  
  **Males**: Show higher motion magnitudes, greater variability in directions, and more abrupt speed changes, reflecting dynamic and diverse motion patterns.

**Females**: Exhibit smoother, more consistent motion across all metrics, with lower variability.

**Dataset Bias**: Gender imbalance amplifies male-dominant trends, highlighting the need for balanced datasets for fair motion analysis.







* Objective 4 : Key insights include the contextualized embeddings of sentences reconstructed using BERT, alignment values between pronouns and occupations measured via cosine similarity, and co-occurrence patterns quantified through PMI.

A graph of different colored lines

Description automatically generated with medium confidence

A diagram of embeddings of occupations

Description automatically generated

* 1. Challenges Faced During Data Analysis:

Objective 1: The computational capabilities to handle the amount of data that needed to be processed was the biggest challenge faced during the data analysis.

Objective 2: Since COCO does not provide gender labels, labeling visible persons manually was time-intensive.

Processing large datasets for depth maps, embeddings, and spatial calculations required significant computational resources.

Obective 3: The UCF101 dataset lacks demographic labels, requiring manual annotation or inference using pretrained gender classification models, which was time-consuming and prone to errors.

Objective 4: Ensuring accurate alignment of embeddings for diverse occupations and pronouns was computationally demanding. Visualizing multi-dimensional relationships between embeddings required significant preprocessing and dimensionality reduction.

1. References

# References

|  |  |
| --- | --- |
| [1] | C. Gorrostieta, R. Lotfian, K. Taylor, R. Brutti and J. Kane, "Gender de-biasing in speech emotion recognition," *IEEE Access,,* pp. 158749-158758, 2020. |
| [2] | J. Buolamwini and T. Gebru, "Gender shades: Intersectional accuracy disparities in commercial gender classification," in *1st Conference on Fairness, Accountability and Transparency*, 2018. |
| [3] | R. Bins and e. al, "Discrimination in online ad delivery," in *2021 CHI Conference on Human Factors in Computing Systems*, 2021. |
| [4] | V. Panayotov, G. Chen, D. Povey and S. Khudanpur, "Librispeech: An ASR Corpus Based on Public Domain Audio Books," in *2015 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP),*, South Brisbane, QLD, Australia, 2015. |
| [5] | M. Foundation, "Common Voice Dataset," [Online]. Available: https://commonvoice.mozilla.org/en. |
| [6] | Tsung-Yi Lin, , Michael Maire, Serge Belongie, Lubomir Bourdev, Ross Girshick, James Hays, Pietro Perona, Deva Ramanan, C. Lawrence Zitnick, Piotr Dollár, "Microsoft COCO: Common Objects in Context," 2014. |
| [7] | Bolei Zhou; Agata Lapedriza; Aditya Khosla; Aude Oliva; Antonio Torralba, "Places: A 10 Million Image Database for Scene Recognition," IEEE, 2017. |
| [8] | A. R. Z. a. M. S. K. Soomro, UCF101: A Dataset of 101 Human Action Classes From Videos in The Wild, Orlando, FL, USA, 2012. |
| [9] | A. R. Z. a. M. S. K. Soomro, UCF101: A Dataset of 101 Human Action Classes From Videos in The Wild, Center for Research in Computer Vision (CRCV), Orlando, FL, USA, 2012. |
| [10] | J. Zhao, T. Wang, M. Yatskar, V. Ordonez, and K.-W. Chang, "Gender Bias in Coreference Resolution: Evaluation and Debiasing Methods," in *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics*, 2018. |