DL4DS Bias Detection Project

Chuqiao Feng, Assylnur Lesken, Jingyuan Liu

December 6, 2024

Abstract

The media has a tremendous impact on the public attitudes of minority communities, and biased reporting can reinforce stereotypes and contribute to systematic discrimination. Bias frequently manifests in subtle forms rather than overtly, often emerging through nonrepresentative distributions and implicit associations that inadequately reflect minority groups' proportional representation within content. This project aims to uncover such biases by developing comprehensive frameworks that evaluate both textual content and visual elements in news media. We implement two approaches: a pipeline-based method that combines BERT (achieving accuracy of 82. 53%, CLIP (94. 70% gender, 71. 49% race detection accuracy) and Large Language Models for integrated analysis; and an end-to-end approach using LLAMA-Vision for simultaneous processing of text and images with enhanced contextual understanding. Evaluation of real-world examples reveals that while the pipeline approach excels in categorical classification, the end-to-end method provides richer contextual analysis. Through this work, we advance the technical capabilities of bias detection systems while highlighting the inherent challenges in identifying media bias, contributing to the broader goal of promoting fair media representation.

1 Introduction

The media plays a key role in shaping public perceptions and social attitudes towards minorities and their communities. Bias in media coverage can, through any form, perpetuate public perceptions, influence public opinion, and contribute to systemic discrimination. Bias frequently manifests in subtle forms rather than overtly; it may emerge through nonrepresentative distributions that inadequately reflect the proportional representation of minority groups within a body of content. These nuances complicate bias detection, as singular cases may appear equitable, yet the overall distribution reveals an inequity. Identifying and confronting these forms of implicit bias is crucial for fostering just and equitable representation.

Our project aims to detect and analyze potential bias in news media using advanced deep learning techniques. By combining state-of-the-art language models (BERT & LLM) with image understanding capabilities (CLIP & Llama-vision), we aim to create a comprehensive framework that can identify both obvious and subtle forms of bias across multiple modalities. The system examines both textual content and visual representations, providing reasoned analysis of potential biases present in news reporting.

2 Related Work

In the realm of computer vision, research has been conducted on detecting bias in visual media. Techniques involve image classification, object detection, and facial analysis to identify and quantify representations of minority groups in images. A typical one is the 'COMPAS' [1] model by Akaash Kambath, used in the criminal justice system to predict recidivism. Kambath discusses issues with COMPAS, notably its tendency to misclassify individuals based on racial bias. For example, research shows it falsely flags Black defendants as high-risk twice as often as white defendants. Kambath argues that COMPAS has not achieved its goal of unbiased justice and highlights the importance of considering alternatives, including the option of non-technological solutions, before implementing predictive tools in sensitive areas like criminal justice.

Methods such as machine learning classifiers have been used to identify negative sentiments associated with minority groups. The article "Context in Informational Bias Detection" [2] explores methods to detect informational bias in news, emphasizing the importance of surrounding context. The study applies context-inclusive models like BERT and RoBERTa with event, article, and domain contexts

to enhance bias detection accuracy. Results show that event-context models outperform sentence-only models, underscoring the role of broader context in identifying subtle informational bias, which is often context-dependent. This contributes to bias detection by developing a more holistic, context-aware approach to processing biased language in news media.

What's more, the paper "Unlocking Bias Detection: Leveraging Transformer-Based Models for Content Analysis" [3] presents the Contextualized Bi-Directional Dual Transformer (CBDT) classifier, which combines two transformers to detect and quantify bias in text. This architecture includes the Context Transformer for sequence-level bias and the Entity Transformer for token-level identification, which helps pinpoint specific biased words or phrases. By fine-tuning with diverse datasets, the CBDT outperforms baseline models in identifying bias across multiple domains.

The advent of Large Language Models (LLMs) like GPT-4 has opened new avenues for understanding nuanced language patterns. These models have shown promise in tasks like sentiment analysis, bias detection, and inferencing implicit information from text. Compared with the previous classical models, we attempt to use LLM to improve the machine's ability to perform bias detection.

3 Approach

Bias Definition

We focus on bias and stereotypes in news articles and social media posts using Social Bias Frames[4], a framework for representing the biases and offensiveness that are implied in language. In this paper, bias is detected through careful analysis of conversational implicatures and commonsense implications, with particular attention to underlying intent, offensiveness, and power differentials between different social groups. For example, a seemingly neutral statement may carry hidden prejudices that can only be understood by examining the implied power dynamics and social context. This comprehensive analysis of bias through Social Bias Frames allows us to train models to better identify and understand subtle forms of bias in media content. By leveraging this framework's labeled dataset of social media posts with detailed bias annotations, our models can learn to detect not just explicit bias, but also implicit prejudices and stereotypes that may otherwise go unnoticed.

Our project implements and compares two distinct approaches for detecting bias in news media content:

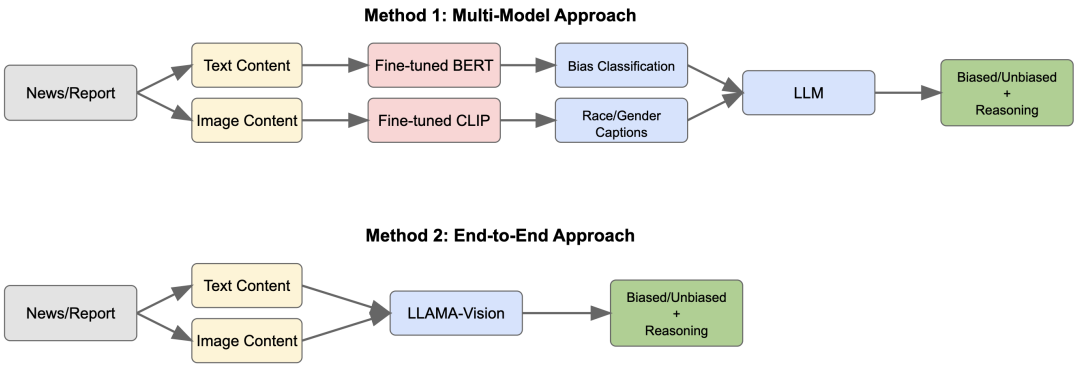

Figure 1: Proposed process of bias detection

Method 1: Pipeline-Based Multimodal Approach

Textual Bias Detection

In this approach, we utilized the BERT model to perform training and predictions for a natural language processing task, focusing on gender and racial bias implication detection.

After selecting all gender- and race-related data from the dataset, we randomly sampled from the remaining non-biased data to construct three balanced datasets containing equal proportions of biased

and unbiased examples. These datasets were then merged to create the final training dataset. For training, we split the data into 0.8 for the training set and 0.2 for the validation set.

Regarding the model, we mainly explored two approaches:

1. Feature Extraction: Using BERT to extract features, followed by classification with Logistic Regression.

2. Fine-Tuning: Fine-tuning the distilbert-base-uncased model on our cleaned dataset due to computational constraints, using AdamW as the optimizer. While this model provides a lighter alternative, it is likely that using the full BERT model would yield better performance.

Both methods adopted batch training, with the sequence maximum length set to 128 to preserve the integrity of the data and a batch size of 32 to ensure efficient computation.

Caption Recognition in Images

In our method, we have applied two methods in total for the image caption task, one is Contractive Language-Image Pretraining (CLIP) to generate bias-related category-based image captions for photographs in the news or post. Also, we have used Llama Vision, where we can directly judge if the photograph make the article consciously or unconsciously conveying biased information.

We chose this because CLIP is a multimodal model, and its visual and textual parts have already been fully pre-trained, so it can easily migrates well to specific tasks. In the fine-tuning process, in order to reduce the amount of computation and efficiently complete the task, we add additional classification heads on the basis of the pre-trained model, and only train the parameters of the classification head, so as to retain the pre-trained knowledge and adapt to the specific task. This means, we freeze the CLIP's visual model, stop its gradient update and only use it to extract image features, then add a separate category header for each target (such as gender and race, or more tasks), mapping the CLIP's visual output to the corresponding number of categories. Here is the prompt we use for our task:

```
gender_prompts = ['Male', 'Female']
race_prompts = ['East Asian', 'White', 'Latino_Hispanic', 'Southeast Asian',
                'Black', 'Indian', 'Middle Eastern']
```

We choose the "ViT-B/32" as the base model, in each training epoch, we traverse the data set and extract the visual features of the image. Input visual features into the corresponding classification head for prediction. Then, calculate categorical losses using real labels and forecasts, then, accumulate losses, backpropagate the loss and update the parameters of the classification headers. In the finetune process, we set epochs=20, batch_size=15, lr=1e-4 to ensure good performance.

Combined Bias Detection LLM

We utilize LLMs to uncover subtle biases that classical NLP methods might miss, introduce the technique of chain of thought for text inference process. Besides, although hallucination of LLM is one of the biggest shortcomings of LLM, we believe that this feature can play an important role in the detection of sensitive information, so we plan to introduce hallucination steps to assist LLM in the judgment of text bias.

We mainly use three methods to define bias in the whole post:

1. **Implicit Contextual Bias**

When the text is initially classified as unbiased but contradicts the demographic information or background context conveyed by the accompanying image, this mismatch may implicitly reinforce stereotypes about the depicted gender or racial group. Such biases arise from the incongruence between textual and visual elements.

2. **Explicit Textual Bias**

If the textual content itself contains explicit bias, and the image aligns with or supports the biased narrative, the post is considered explicitly biased. This form of bias is evident in the language or expression, clearly reflecting discriminatory content.

3. **Hallucination-Driven Assumptive Bias**

Utilizing the hallucination capabilities of large language models, hidden assumptions or latent biases can be uncovered. Even if the text and image do not overtly exhibit bias, generative reasoning can identify implicit stereotypes or conflicts within the underlying context.

Here is the exact prompt we use:

```
"You are a bias detection assistant specializing in identifying racial and gender biases.
Your task is to analyze a social media post that contains both textual content and an
accompanying image. The evaluation involves three steps:

1. Text Bias Analysis: Analyze whether the text is unbiased or biased. If biased, identify
   whether the bias is gender-based, racial, or both.
2. Image Caption Analysis: Review a caption describing the racial group(s) or demographic
   information identified in the accompanying image and assess whether the caption reflects
   implicit or explicit bias.
3. Hallucination-Based Bias Detection: Introduce plausible implicit assumptions or
   stereotypes based on the image and text, even if not explicitly stated. This step helps
   identify potential hidden biases or mismatched contexts.

Based on these inputs, determine whether the post as a whole is biased. A post is considered
biased if it meets any of the following criteria:

- Implicit Bias via Mismatch: The text is classified as unbiased, but the image's demographic
  information or hallucinated assumptions contradict the text's context, implicitly
  reinforcing stereotypes about the depicted racial or gender group.
- Explicit Bias in Text: The text itself is biased, and the image aligns with or
supports the biased narrative.
- Hallucinated Bias Patterns: The hallucination step uncovers hidden assumptions or
stereotypes that suggest the presence of bias.

Provide your evaluation as: 'Gender-biased,' 'Racial-biased,' or 'Unbiased,' along
with a concise explanation highlighting the criteria and hallucination-based observations
that led to your conclusion."
```

Method 2: End-to-End Multimodal Approach

The second method employs a more streamlined approach using LLAMA-Vision, which can process both text and images simultaneously. Since the large model was deliberately designed by the researchers to avoid any answers that might involve discriminatory and biased views during its pre-training process, we need to pay special attention to the logic of the prompt when using the llama-vision. So we applied prompt engineering techniques to enhance the performance of the model so that it can output what we want.

Our prompt logic mainly include three parts:

- Instruction of task: Make definition of the task.
- Text / Image Bias Analysis separately: ask the model to analyze whether the input text itself is unbiased or biased, identify the racial group(s) or demographic information in the given image and analyze whether it is biased.
- Combining Analysis: Determine whether the post as a whole is biased according to our pre-defined bias criteria. (In 8 aspects, selection and presentation of information, word choice and tone, representation, balance, motivation behind the content, subtext and implied messages, cultural and historical background, strategies to avoid bias.)

The model we use is llama3.2-vision:11b, for each trial the token number is about 800 and it will take about 2 minutes on a CPU environment with llama-vision loaded from 'ollama'.

4 Datasets

Since we use LLM to infer whether the text report content is biased, we can directly use the zero-shot model with proper prompt engineering. However, to make our work perform bias detection better on the racial problem specifically, we fine-tuned the LLM with SBIC and BiasCorp datasets, which are labeled with details of text in racism bias. As for image abstraction, since all 'image to text' models are too ambiguous for the area we focused on. We aim to detect the sensitive content from the image directly to generate the image abstraction. Here we found the Ethnicity Recognition Dataset to do this work.

SBIC

We selected SBIC [4] as our texual bias detection dataset. SBIC, Social Bias Frames is a new way of representing the biases and offensiveness that are implied in language. For example, these frames are meant to distill the implication that "women (candidates) are less qualified" behind the statement "we shouldn't lower our standards to hire more women." It contains 150k structured annotations of social media posts, covering over 34k implications about a thousand demographic groups.

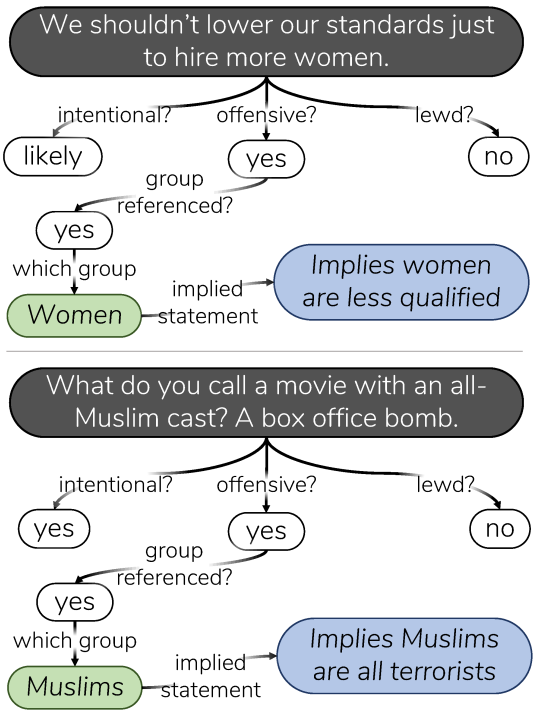

Figure 2: Example of social bias frames

Considering the initial imbalance in the data distribution, we performed two main types of cleaning on the dataset. First, we selected all entries classified as gender or racial bias. Second, we randomly sampled an equal number of non-biased entries to match the number of biased examples, resulting in a balanced and restructured dataset.

Class	**Sample Count**	**Avg Length (chars)**
unbiased	28919	111.82
gender-biased	4521	108.34
racial-biased	5479	107.68
Total	38919	109.28

Table 1: Data Distribution Before Balancing

Class	**Sample Count**	**Avg Length (chars)**
unbiased	5479	112.69
gender-biased	4521	108.35
racial-biased	5479	107.68
Total	15479	109.57

Table 2: Data Distribution After Balancing

FairFace

We utilize FairFace [5], a balanced face dataset containing 108,501 face images with demographic annotations. The dataset was specifically constructed to mitigate racial bias in face image datasets, as

most existing datasets are heavily skewed toward White faces. FairFace provides balanced annotations across 7 race groups (White, Black, Indian, East Asian, Southeast Asian, Middle Eastern, and Latino), along with gender labels.

		Train	**Test**
Gender	Male	45986	5792
	Female	40758	5162
Race	White	16527	2085
	Latino_Hispanic	13367	1623
	Indian	12319	1516
	East Asian	12287	1550
	Black	12233	1556
	Southeast Asian	10795	1415
	Middle Eastern	9216	1209
Total		86744	10954

Table 3: Train and Test Data Distribution by Gender and Race

5 Evaluation

Textual Bias Detection BERT model

The BERT's performance using feature abstracting was evaluated on a validation dataset with the proportion of 0.2 on the SBIC dataset, achieving an overall accuracy of 0.75. A detailed classification report, including precision, recall, F1-score, and support for each class, is presented in Table 5. The macro and weighted averages indicate a balanced performance across classes, with macro F1-score of 0.73 and weighted F1-score of 0.75. These results suggest that the feature extraction method can effectively capture relevant characteristics for biased implication detection, although there is room for improvement, particularly in classifying gender-biased classes.

Class	**Precision**	**Recall**	**F1-score**	**Support**
unbiased	0.76	0.82	0.79	2031
gender-biased	0.65	0.59	0.62	880
racial-biased	0.79	0.75	0.77	1089
Accuracy		0.75 (4000 total)		
Macro Avg	0.74	0.72	0.73	4000
Weighted Avg	0.75	0.75	0.75	4000

Table 4: Classification report of the Feature-Abstracting BERT

For better performance on textual bias detection, we also finetuned a distilbert-base-uncased model using our SBIC dataset, which achieved much better result The model achieved a training loss of 0.3357 and a validation accuracy of 0.82. The classification report (Table 6) shows strong performance for the "notbiased" and "racial" categories, with F1-scores of 0.85 and 0.85, respectively. However, the "gender" category underperformed slightly, with an F1-score of 74.02

To address overfitting, we adopted a strategy of saving the model with the highest validation accuracy during training. Analysis of accuracy fluctuations revealed that the model peaked at epoch 2, after which overfitting became apparent. This overfitting likely stems from the limited size of the dataset, which restricts the model's ability to generalize beyond the training data.

Image Annotation with CLIP model

For the CLIP model, we first use the baseline model to test the fairface dataset, and calculate the accuracy of image annotation on 'Gender' and 'Race'. After finetuned, our new model performances on these two tasks are below here.

Class	Precision	Recall	F1-score	Support
Not Biased	0.8720	0.8321	0.8516	2031
Gender	0.7028	0.7818	0.7402	880
Racial	0.8523	0.8476	0.8499	1089
Accuracy		0.8253		4000
Macro Avg	0.8090	0.8205	0.8139	4000
Weighted Avg	0.8294	0.8253	0.8266	4000

Table 5: Classification Report for the Fine-Tuned BERT Model

	Gender	Race
Accuracy	54.72%	36.26%

Table 6: Performance of CLIP base model

	Gender	Race
Accuracy	94.70%	71.49%

Table 7: Performance of fine-tuned CLIP model

Pipeline and Llamavision performance

In our evaluation, we tested our two models using two sets of combined image and textual data posts. The first set consisted of a photograph of a Chinatown paired with two texts: one neutral and the other biased. The biased text employed a subtly ambiguous tone, yet its prejudiced nature remained recognizable to the majority of readers. Individuals with relevant training in detecting bias were expected to identify it quickly and accurately.

Figure 3: chinatown image post

This is the neutral post, simply describing in a positive and observational tone, focusing on the lively atmosphere and cultural diversity without making any judgments or assumptions:

```
" Stumbled upon this super colorful street today|lanterns everywhere and signs in all
kinds of languages! The vibe here is so lively, with people chatting and exploring
little shops and restaurants. Definitely one of those places that feels like its own
little world. Might have to come back for some dim sum #CityAdventures
#ChinatownVibes "
```

While the biased post subtly expresses negative implications about the neighborhood. Phrases like "quieter" and "decorations looking a little dated" suggest a sense of decline, while the statement "where all the noise at night in our area comes from—now I think I know" unfairly associates the community with being a source of disturbance. These implications rely on stereotypes and could lead to negative perceptions of the Chinatown and its residents.

```
"Today I came across a Chinatown near our neighborhood. It felt a bit quieter, with
fewer people around and some of the decorations looking a little dated. I've always
wondered where all the noise at night in our area comes from|now I think I know."
```

Both our Multi-modal and End-to-End models performed well on these two tests. They both accurately identified whether the posts contained biased content. Furthermore, the end-to-end model

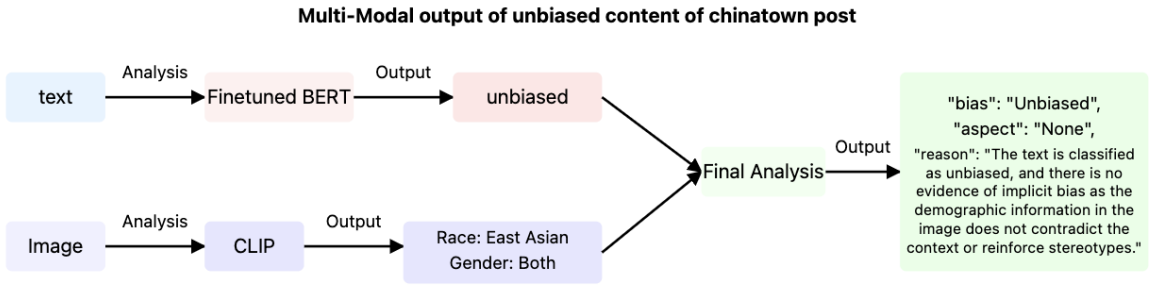

Figure 4: Multi-Modal output of unbiased chinatown post

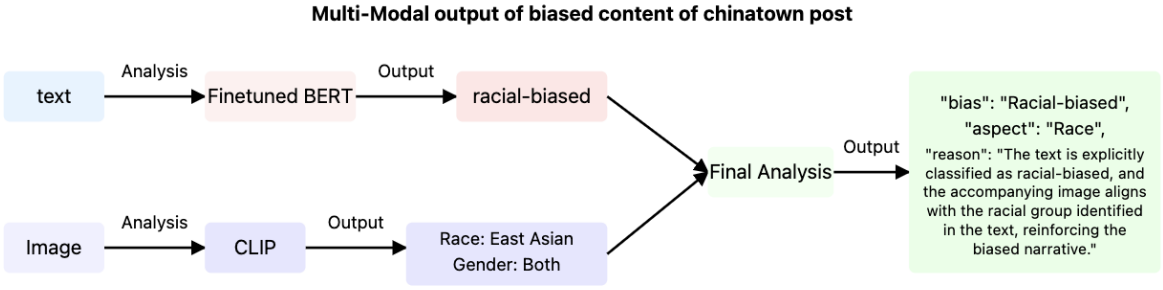

Figure 5: Multi-Modal output of biased chinatown post

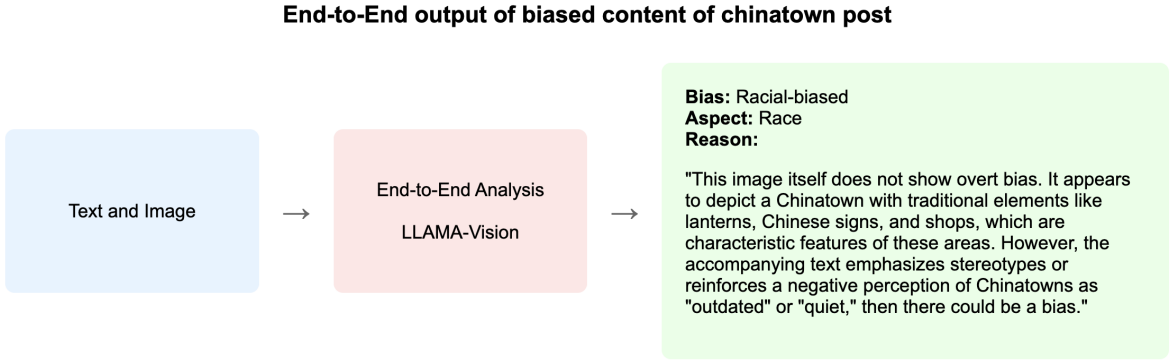

Figure 6: End-to-End output of biased chinatown post

demonstrated an additional capability by providing a more detailed explanation of the specific reasons behind the biased nature of the content.

In the second set of image and text example, we used a more challenging content where even humans needed to carefully analyze and reflect to determine whether bias was present. Additionally, the biased content was closely tied to the context of the image. The same text, when paired with different images, could result in varying levels or types of perceived bias, further demonstrating the complicated relationship between visual and textual elements. The image of the post depicts a group of people participating in a protest or demonstration.

Here lies the unbiased post along with the image, stating a historical observation, pointing out the actual struggle for justice and equality for Black Americans. Its tone is neutral, as it does not place blame, criticize specific groups, or use emotionally charged language. It simply states an issue, making it objective and free of bias.

Figure 7: Protesting image

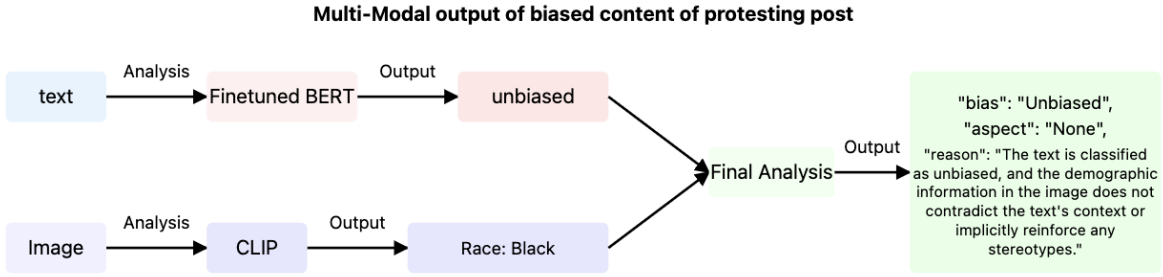

Figure 8: Multi-Modal output of biased protesting post

`More than a century after the Civil War and 16 years after the Supreme Court's school-desegregation ruling, the American black has not achieved justice or equality.`

And here is a given biased post because it implies a negative view and the black women are standing in the opposition from the government. By framing the protests and the black American group involved in this way, the post introduces a prejudiced perspective that undermines the legitimacy of their cause.

`The U.S. government is not pleased with the frequent demonstrations and is prepared to take action against these groups.`

However, our BERT model produced completely opposite results in bias detection for these two text posts. This is likely because BERT tends to classify sentences containing race-related terms as racially biased, without fully considering the underlying emotional tone. Additionally, BERT struggles to differentiate between objective descriptions of established facts and expressions of personal sentiment, which impacts its final analysis. As a result, the Multi-modal output becomes not well-performed, and, in this case, even contradictory. This also highlights a critical area for improvement in our model.

6 Conclusion

This project presents two approaches for detecting bias in news content: a pipeline-based approach integrating BERT (82.53% accuracy), CLIP (94.70% gender, 71.49% race detection accuracy), and LLM models, and an end-to-end approach using LLAMA-Vision. While both methods effectively identified explicit and implicit biases, each showed distinct strengths - the pipeline approach excelled in categorical classification, while the end-to-end method provided richer contextual analysis. Despite the models' strong performance, we observed limits in handling context-dependent bias detection, demonstrating how the intrinsic difficulty of defining bias affects model performance. These findings contribute to the field of automated bias detection while emphasizing the need for continued improvements in context understanding and multimodal integration.

7 Code repository

GitHub Link: https://github.com/Carrie1013/DS542-Bias-Detection

References

[1] Akaash Kambath, "A COMPAS That's Pointing in the Wrong Direction," UC Berkeley School of Information Blog, July 9, 2021. Available at: https://blogs.ischool.berkeley.edu/w231/2021/07/09/a-compas-thats-pointing-in-the-wrong-direction.

[2] Ming Liu, Alessandro Moschitti, and Barbara Plank, "Context in Informational Bias Detection," In *Proceedings of the 28th International Conference on Computational Linguistics (COLING 2020)*, December 2020, pp. 6359-6365. Available at: https://aclanthology.org/2020.coling-main.556.pdf.

[3] Shaina Raza, Oluwanifemi Bamgbose, Veronica Chatrath, Shardul Ghuge, Yan Sidyakin, and Abdullah Y. Muaad, "Unlocking Bias Detection: Leveraging Transformer-Based Models for Content Analysis," arXiv preprint, arXiv:2310.00347, 2023. Available at: https://arxiv.org/pdf/2310.00347.

[4] Maarten Sap, Saadia Gabriel, Lianhui Qin, Dan Jurafsky, Noah A. Smith, and Yejin Choi, "Social Bias Frames: Reasoning about Social and Power Implications of Language," ACL, 2020. Available at: https://maartensap.com/pdfs/sap2020socialbiasframes.pdf.

[5] Kimmo Karkkainen and Jungseock Joo, "FairFace: Face Attribute Dataset for Balanced Race, Gender, and Age for Bias Measurement and Mitigation," in *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, 2021, pp. 1548–1558. Available at: https://openaccess.thecvf.com/content/WACV2021/papers/Karkkainen_FairFace_Face_Attribute_Dataset_for_Balanced_Race_Gender_and_Age_WACV_2021_paper.pdf.