Hollywood Identity Bias Dataset: A Context Oriented Bias Analysis of Movie Dialogues

Sandhya Singh*, Prapti Roy*, Nihar Sahoo*, Niteesh Mallela*, Himanshu Gupta*, Pushpak Bhattacharyya*, Milind Savagaonkar†, Nidhi†, Roshni Ramnani†, Anutosh Maitra†, Shubhashis Sengupta†

*Indian Institute of Technology Bombay, India; †Accenture Labs, India

Warning: This poster has contents that may be upsetting, however, this cannot be avoided owing to the nature of the work.

Motivation

- Movies reflect society and hold the power to transform opinions at a larger scale.
- An AI assistant identifying the social biases can help the production houses avoid the inconvenience of stalled release, lawsuit and commercial losses.

Introduction

- We introduce a new dataset as Hollywood Identity Bias Dataset (HIBD) consisting of 35 movie scripts annotated for multiple identity biases.
- The dataset contains annotated scripts for *Sensitivity, Stereotype and Social Bias labels as Gender, Race, Religion, Age, Occupation, LGBTQ, and Other,* that has biases like *body shaming, personality bias, etc.*
- Each annotated bias is further labeled *Implicit or Explicit* to convey the nature of bias along with their corresponding *target group and the rationales* behind it.
- We are annotating *sentiment* as positive or negative and its associated *emotion and intensity* based on plutchik's emotion wheel for each bias.

Problem Statement

Given a Hollywood movie script, identify the biased/ sensitive dialogues in it and detect the category, target of the bias. In our work, we focus on six major types of social biases, *i.e., Gender bias, Race bias, Religion bias, Occupation bias, Age bias, LGBTQ bias.*

Dataset - HIBD

Labels	Sentence Level	Dialogue level
Bias	1181 (2.40%)	976 (3.42%)
Neutral	47936	27558
Total	49117	28534

Table 1:Distribution of Biased sentences and dialogues.

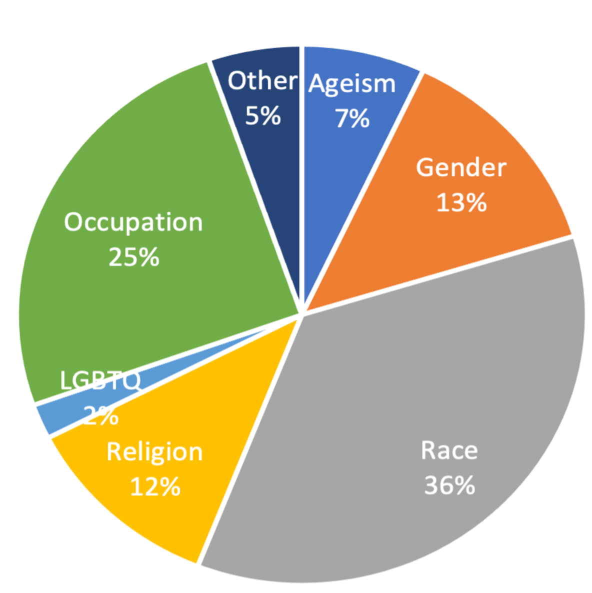

Figure 1:Distribution of social biases across 7 categories. We show percentages of each category annotated in the dataset.

Terminologies

Stereotype: It is an overgeneralized belief about a particular community. For example, *"Some Asians are good at maths."* is a fact statement; but *"All Asians are good at maths."* is a stereotype.

Sensitivity: The property of a statement targeted towards an individual or a group belonging to a section that is vulnerable due to identity such as *race, religion, occupation, etc.* It always bears a negative sentiment. For example, *"The church is a racket. I know how they operate."* is a sensitive statement against the Christian community.

Bias: Bias refers to prejudice towards or against an individual or community based on their identity such as *gender, race, religion, occupation etc.* Bias can be defined as a quintuple $< S, L, C, T, R >$ where,

- S is the communicator (speaker, author)
- L is the communicatee (audience, reader)
- C is the category of bias.
- T is the target of the bias.
- R is the reason for bias.

Annotator Agreement

Labels	Cohen Kappa
Ageism	0.72
Gender	0.54
Race/Ethnicity	0.61
Religion	0.67
LGBTQ	1
Occupation	0.47
Other	0.49
AVERAGE (all categories)	**0.64**
Stereotype	0.44
Sensitivity	0.33
Bias	0.71

Method

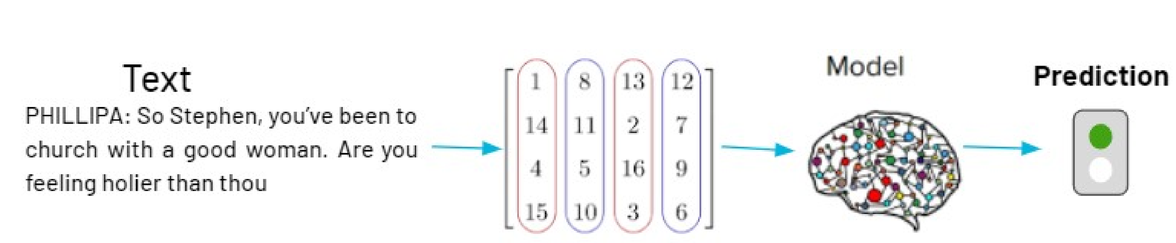

Figure 2:Model diagram for binary bias detection task.

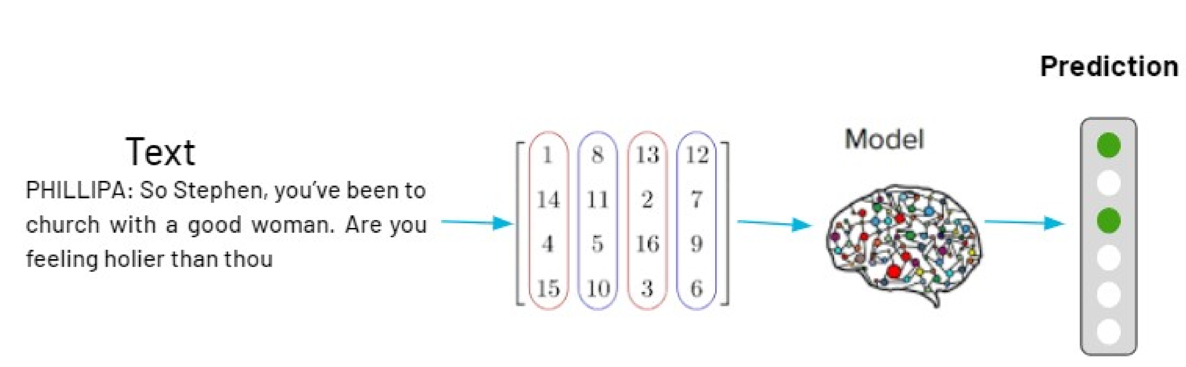

Figure 3:Model diagram for multi-label bias category detection.

Due to shallow presence of biased instances in our dataset, we use transfer learning for our experiments. First, we fine-tune the model on a curated dataset of a few related tasks before fine-tuning again on our dataset.

Results

Models	P	R	F1
LR	0.53	0.71	0.51
LR-contrl	0.52±0.008	0.70±0.011	0.49±0.021
SA	**0.55**	**0.81**	**0.58**
SA-contrl	0.55±0.007	0.80±0.012	0.57±0.017

Figure 4:Performance of binary classification[Bias vs. Neutral].

	LR			BART-large (SA)		
	P	R	F1	P	R	F1
Race/Ethnicity bias	0.500	0.410	0.450	0.77	0.89	**0.83**
Religion bias	0.226	0.259	0.241	0.86	0.67	**0.75**
Gender bias	0.302	0.432	0.355	0.73	0.73	**0.73**
Occupation bias	0.321	0.464	0.380	0.59	0.48	**0.53**
Ageism bias	0.171	0.462	0.250	0.62	0.62	**0.62**
LGBTQ bias	0.158	0.273	0.200	1.00	0.73	**0.84**

Observations

- The BART-large model substantially outperforms logistic regression for bias category detection task. This is mainly due to the predictive power of transformer based models.
- We have observed that the category detection model sometimes predicts some extra categories for the dialogue which are not available in the ground truth label.
- The bias detection model, generally, fails for implicit cases. Because to capture implicit biases, we need to model previous dialogues and speaker attributes.

Conclusion

- We release a dataset of 35 Hollywood movies annotated for identity social biases in movie scripts.
- The dataset is labeled for *Sensitivity, Stereotype, Identity Biases as Gender, Ageism, Race/Ethnicity, Religion, Occupation, LGBTQ, Other (body shaming, personality, etc.), Target of the bias, Sentiment, Emotion, Emotion Intensity, and reason* for bias.
- The dataset has been benchmarked for bias identification and categorization task using the BART-large model.

Dataset Code Repository

https://github.com/sahoonihar/HIBD_LREC_2022

भारतीय भाषा प्रौद्योगिकी केन्द्र