Team 54 Final Report: Analyzing Representation and Stereotyping in Comedies

Pia Rodriguez pvrodrig@usc.edu, Frederick Zhang fyzhang@usc.edu, Christian Becker beckerch@usc.edu,
Rudra Singh rpsingh@usc.edu, Ethan Feng eyfeng@usc.edu

Abstract

This study investigates the prevalence of implicit bias within comedic scripts through the development and evaluation of machine learning models. We explore the limitations of current datasets and models. The project implements and evaluates four distinct classification models, including SVM, Perceptron, BiLSTM, and BERT, on various comedy scripts dating back to the early 2000s. The findings demonstrate the effectiveness of using deep learning models like BERT to uncover bias within modern comedic narratives.

1 Introduction

Comedy, a cornerstone of entertainment, holds immense power to shape cultural perceptions. However, humor can be a double-edged sword, sometimes perpetuating stereotypes and biases without our conscious awareness. This project tackles the question of implicit bias within comedic scripts. We leverage Natural Language Processing to investigate the prevalence of these biases and their potential impact. By analyzing comedy scripts through this lens, we aim to foster a more nuanced understanding of humor and its role in reinforcing or challenging stereotypes.

2 Related Work

2.1 Detection of Bias in Textual Media

Nadeem and Raz (2022) provide a comprehensive methodology for employing NLP algorithms to detect biases in written content. This study offers a structured approach to understanding bias in media texts, which has significantly influenced the field. Building on the methodologies outlined, our project adapts these approaches to the specific context of comedic scripts. Their work primarily focuses on news articles, we extend this research to explore how biases manifest in humor, a less formally studied area in bias detection.

2.2 Advanced NLP Models in Media Analysis

The application of BERT for various NLP tasks has been well-documented, as seen in the work by Sun et al. (2019), which outlines methods for fine-tuning BERT for specific text classification tasks. Our project extends this work by adapting BERT to the nuanced task of detecting implicit bias within comedy. This involves training the model on specialized datasets and evaluating the performance of comedic scripts and whether it can still identify implicit bias.

3 Problem Description

This project tackles the challenge of identifying and analyzing implicit bias within comedic scripts. The key challenge lies in developing models capable of accurately identifying implicit bias within the specific context of comedic language. Humor often relies on exaggeration, irony, and wordplay, which can make it difficult to distinguish between genuine bias and comedic devices. Furthermore, the lack of large, well-annotated datasets specifically focused on implicit bias in humor presents an additional hurdle.

4 Methods

4.1 Materials

In pursuit of identifying implicit bias, we utilized three distinct datasets to both train and evaluate our models: Crowdsourced Stereotype Pairs Benchmark (CrowS-Pairs), Stereoset, and NewsMediaBias.

The CrowS-Pairs [1] dataset annotates each sentence with various bias labels encompassing race, socioeconomic, gender, nationality, sexual orientation, religion, and others. Similarly, Stereoset [2] follows a comparable framework to CrowS-Pairs, categorizing bias into four main domains: gender, profession, race, and religion. Distinctively, NewsMediaBias[3]differentiates from the structure of the former datasets by being comprised of tweets rather than sentences. We exclusively utilized the inputs from this dataset that contained the "neutral" label to supplement our training data with instances lacking bias, a critical label that was absent from the other datasets. We were unable to locate another dataset akin to CrowS-Pairs and Stereoset, offering complete sentences annotated with the "no-bias" label. This label is crucial in our model's training and evaluation, ensuring it doesn't excessively detect bias in the script data.

In the second phase of our project, focusing on scripts, we selected twenty-one comedy scripts from the early 2000s to evaluate our model's bias detection capabilities.

4.2 Procedures

To preprocess the bias data for both training and evaluation, we initially transformed each input sentence or tweet into a word list. This involved cleaning each word by removing non-alphabetical characters, expanding contractions, and converting all words to lowercase. Each word is then tokenized based on its position in the data. If a word is present in the GloVe embeddings, it retained its original form; otherwise, it was replaced with "<unk>" and assigned the appropriate index. Each input was then labeled according to one of the nine categories used in our evaluation process: race, gender, socioeconomic, nationality, religion, sexual orientation, profession, other, and none. All three preprocessed datasets were combined into a unified dataset, which we then divided into training and testing subsets in an 80:20 ratio.

Following the pre-processing stage, the input sentences were transformed into TF-IDF embeddings if fed into the SVM or Perceptron models, or BERT embeddings if utilized by the transformer model. The processed data was then inputted to train and evaluate our four classification models. These models consist of SVM, Perceptron, BiLSTM, and BERT-Transformer. Once the models were fine-tuned to produce satisfactory metrics, we proceeded to input the scripts.

The scripts underwent a pre-processing procedure similar to that of the training dataset. However, the scripts are divided into blocks consisting of consecutive sentences separated by blank lines, unlike the training data, where we evaluate individual sentences. These processed scripts are then fed into each of the trained models, which assigns a bias classification to each block from the nine predefined classes mentioned earlier. It is important to note that the scripts are exclusively employed during the evaluation phase of our model, as we lack accurate implicit bias label assignments for each block.

4.3 Models

We concentrated on employing the models we had studied in our coursework and assignments to accomplish the project goals. Given the project's emphasis on classification tasks, we employed both straightforward models like the Support Vector Machine (SVM) and the Perceptron model. Additionally, we incorporated more sophisticated models such as the Bidirectional Long Short-Term Memory (BiLSTM) and the Bidirectional Encoder Representations from Transformers (BERT).

The SVM model was constructed with the linear kernel and the following parameters: penalty set to "l1", random state set to 42, maximum iteration set to 1000, and dual set to False.

We didn't make any modifications to the Perceptron model or provide any hyperparameters for it. We opted to use a single-layer perceptron for our classification.

In the case of the BiLSTM, we employed the following hyperparameters: embedding dimension set to 100, hidden dimension set to 256, number

[1] https://huggingface.co/datasets/crows_pairs

[2] https://huggingface.co/datasets/McGill-NLP/stereoset

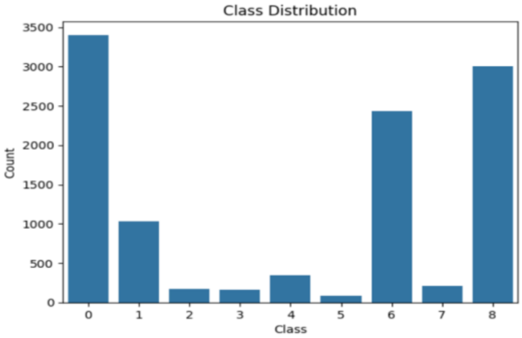
[3] https://huggingface.co/datasets/newsmediabias/news-bias-full-data/blob/main/train.csv

of classes set to 9, dropout rate set to 0.33, output feature set to 128, learning rate set to 0.1, batch size set 14, number of epochs set to 4, and the embedding matrix created from our GloVe embeddings.

As for BERT, we utilized the “bert_base_uncased” pre-trained model from the Hugging Face Library and the BertTokenizer to tokenize our dataset. Given our task of classifying data into nine distinct classes, we opted for the multi-label classification variant for our model. During the training phase of this model, we used the following hyperparameters: batch size of 18, weight decay of 0.01, learning rate of 2e-5, and the number of epochs set to 3.

5 Experimental Results

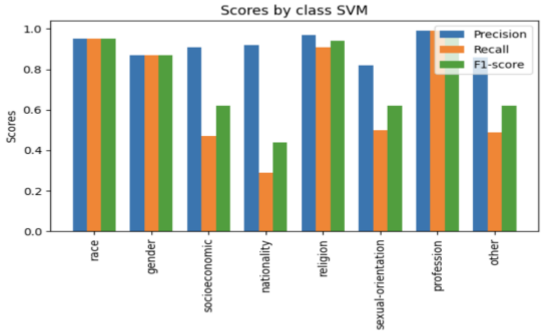
After preprocessing our datasets, we had 10,825 rows of data available for our model. As previously mentioned, words not found in the GloVe embeddings were assigned the tag “<unk>”. Even with this conversion, the percentage of unknown words in our dataset remained relatively low, at 3.25%. Below, you will find the distribution of our classes. Notably, there is a higher frequency of data labeled with race, profession, and none. This disparity stems from the uneven distribution in the datasets we utilized. Consequently, our model may exhibit a bias towards predicting race and none more accurately than the other bias labels.

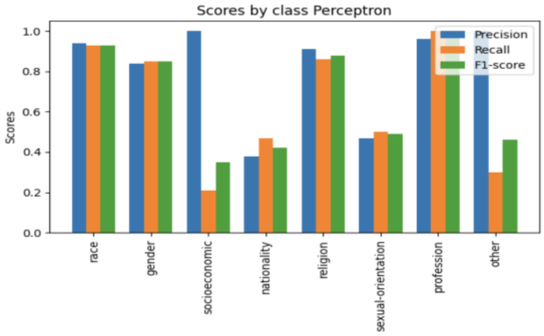


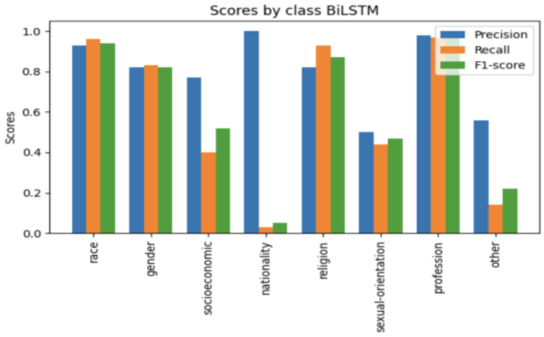
After training each of our four models, below are the scoring metrics of each model evaluated on the testing dataset. The precision, recall, and F1 metrics were assessed based on the weighted averages.

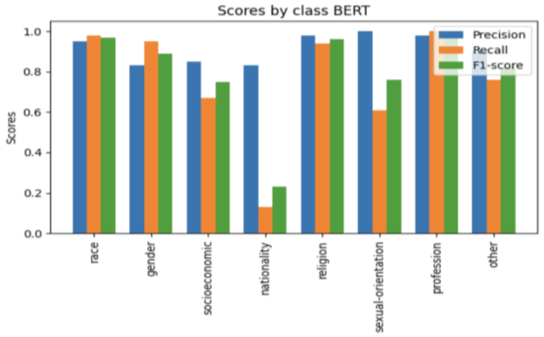
	Precision	Recall	F1	Accuracy
SVM	92.78%	92.70%	92.10%	92.70%
Perceptron	91.52%	91.03	90.38%	91.03%
BiLSTM	90.96%	91.22%	89.75%	91.22%
BERT	95.06%	95.01%	94.78%	95.01%

Given the uneven distribution of classes in the training and testing dataset, below are the scoring metrics by class for each of the models evaluated on the testing dataset.

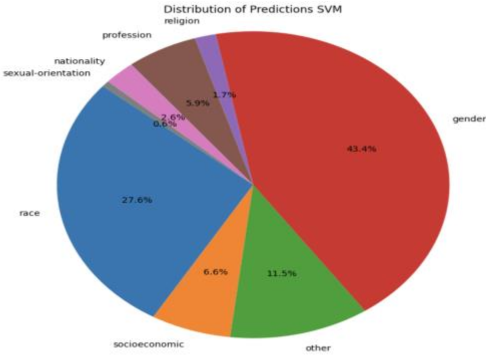


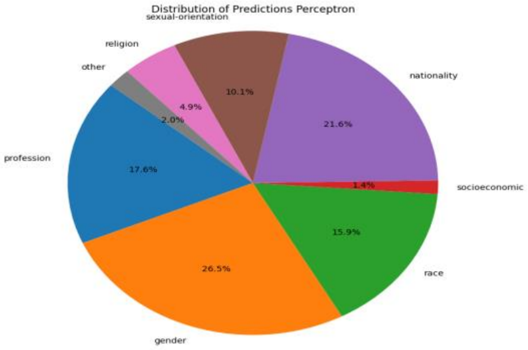


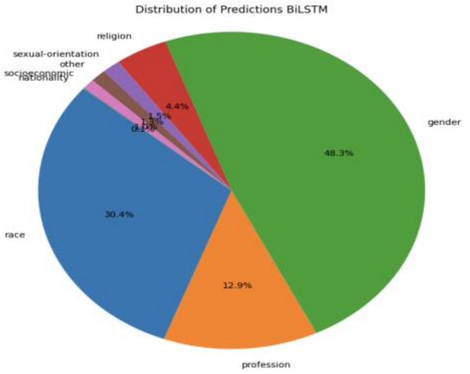


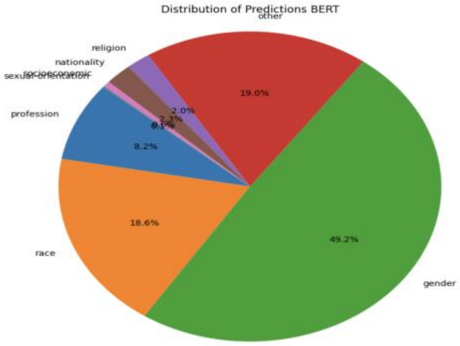


After training the models, we analyzed each script to determine the distribution of detected biases within them, as well as the overall bias detected across all scripts. Below is the aggregate bias identified from all the scripts for each model.









After analyzing the distribution, it was observed that gender bias was frequently the most detected bias in all the scripts for each model. Given that the transformer model exhibited the highest metrics across all classes, we regarded its bias-detected output as the most accurate. It is important to note that our assessment of accuracy is solely based on the models' evaluation of the testing data, as we lack the correct labeled outputs of the scripts.

Analyzing the models' outputs sheds light on the prevalent biases presented in comedic scripts from the early 2000s. The outputs provide an opportunity for further investigation into the role and impact of biases in comedy, particularly regarding gender representation. One perspective is that biases in comedy may serve as a reflection of evolving societal attitudes toward gender during that era. On the other hand, there is a concern that biases in comedy may uphold harmful stereotypes and minimize important gender-related issues. This raises questions about the responsibility of comedy creators to navigate humor and reinforce harmful stereotypes. Further analysis and discussion can be done to understand better the complex dynamics at play and the potential implications for gender and other biases in comedic content.

6 Conclusion and Future Work

In this study, we developed a methodology to identify and analyze implicit bias within comedic scripts, a domain where humor's reliance on nuances such as exaggeration, irony, and wordplay adds complexity to textual analysis. Despite these challenges, our models demonstrated promising results in distinguishing between instances of implicit bias and comedic expressions. Among the models tested, BERT performed the best across all metrics, including precision, recall, F1 score, and accuracy, suggesting its suitability for tasks involving nuanced language understanding.

Our analysis also highlighted the impact of data imbalance on model performance, particularly in how it favors predictions related to the more frequently represented classes, such as race and profession. This underscores the importance of having a balanced dataset to reduce potential biases in model predictions.

As our next steps, we aim to improve the model's bias detection within scripts. This involves annotating each script block with human-verified bias labels to refine and evaluate our models further. Additionally, we are exploring methods to integrate humor detection into the models, which could enhance the identification of bias concealed within humorous language. Building upon existing research, transformer-based models offer a promising starting point for humor detection, and integrating this capability with bias detection could enhance the script evaluation processes.

7 Individual Contribution

7.1 Pia Rodriguez

For this project, my contributions have included searching and finding the datasets we are using to train the model, developing and tuning all the models for best accuracy, creating the metrics and graphs for each model, assisting in writing the proposal, status report, and the Methods and Experimental Results sections of the final paper, and created the final presentation.

7.2 Rudra Singh

In our project, I contributed by participating in brainstorming sessions, trying to search for relevant datasets, helping draft the project proposal, including the development of a detailed timeline to help manage the project's milestones, and writing the literature review and their corresponding references for the final proposal.

7.3 Christian Becker

My contribution to this project was primarily related to the preparation of the script data. Because there are almost no script-related datasets, substantial effort had to be put into procuring data that our model could be used on. Furthermore, once the actual scripts were obtained, they had to be formatted in a way that was appropriate for the model. Therefore, the scripts had to be modified to be more similar to the single, long sentences the models had been trained on, rather than short pieces of dialogue. This lead to approach of feeding in “blocks” of texts with maximum lengths, rather than every individual sentence.

7.4 Frederick Zhang

In our project, I focused on gathering the right datasets used for our training and testing. This meant finding suitable datasets and making sure they were imported properly in our code. I handled tasks like loading the datasets and preprocessing it to extract the features and labels we needed.

7.5 Ethan Feng

In our project, my primary responsibility was analyzing the efficiency of our models and concluding on their overall accuracy. I explored various ways to enhance the models' accuracy and reliability. This included tuning hyperparameters, experimenting with different architectural adjustments, and investigating the potential of ensemble methods to leverage the strengths of multiple models.

8 References

Chi Sun, Xipeng Qiu, Yige Xu, and Xuanjing Huang. 2020. *How to Fine-Tune BERT for Text Classification?* Chinese computational linguistics: 18th China national conference, CCL 2019. Springer International Publishing, 2019.

Dirk Hovy and Shannon L. Spruit. 2016. *The Social Impact of Natural Language Processing.* In Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers), (pp. 591–598). Association for Computational Linguistics.

Muhammad Umar Nadeem and Sarah Raza. 2022. *Detecting Bias in News Articles using NLP Models.* Stanford CS224N Custom Project.