Stereotypical Bias Analysis in Large Language Models

Priya Patel, Neha Rana, Hardi Rakholiya, Shreya Sejani, Aneri Thakkar
Undergraduate students Nirma University, Ahmedabad, Gujarat, India 382481
e-mails:
21bce238@nirmauni.ac.in
21bce246@nirmauni.ac.in
21bce241@nirmauni.ac.in
21bce266@nirmauni.ac.in
21bce299@nirmauni.ac.in

***Abstract*—Stereotypical bias, an extensive cognitive phenomena, continues to impact perceptions, decisions, and behaviors across a wide range of human interactions.Because pre-trained language models are quite prominent and are trained on large real-world datasets, there are concerns that these models could incorporate and reinforce stereotypical biases. The necessity to measure the biases inherent in these models is discussed in this work. Previous research usually assesses pre-trained language models using a small number of artificially constructed bias-assessing sentences. In order to offer a comprehensive evaluation, we present a novel dataset that combines four distinct datasets to analyze stereotyped biases in ten distinct domains: race/color, socioeconomic, gender, disability, nationality, sexual orientation, physical-appearance, religion, age, profession. We analyze well-known models like BERT and RoBERTa systematically, exposing their tendency to show considerable stereotypical biases in a variety of domains. Our results highlight the pervasive nature of the typical biases present in these models are and emphasize how critical it is to eliminate bias in applications involving natural language processing.**

***Index Terms*—large language models, natural language processing, stereotypes, bias.**

I. Introduction

Large Language Models (LLMs), have proven highly effective at a variety of tasks [1]. Their versatility has led to their widespread integration into human decision-making processes, from quickly summarizing documents to handling queries regarding mathematics and even delivering chat-based support. Still, despite their abilities, LLMs are not without flaws. Their challenges stem from algorithmic biases, which call for the application of stringent assessment and mitigation techniques [2]–[5]. Furthermore, these models may display cognitive biases similar to those found in human decision-making processes, in addition to social prejudices, which might influence users' opinions [6].

Both in human cognition and human-machine interaction, cognitive bias is a phenomena that is typified by consistent departures from norms of reasonable judgment [7], [8]. It is crucial to ensure comprehensive model audits, especially in situations when LLMs support consequential decision-making, such evaluating individuals [9], in order to reduce the influence of cognitive biases on decision outcomes.

Stereotypical bias, often referred to as stereotyping, is a cognitive process where people categorize individuals based on characteristics such as race, gender, age, or other attributes, and then make assumptions about them. Stereotypical bias presents serious challenges in a range of different fields since it is characterised by excessively generalized ideas about particular groups of people. Examples of biases include assumptions that women are naturally more emotional than men or that older people are not proficient in technology. These ideas - often referred to as biases - can negatively impact the targeted groups, restricting their opportunities and sustaining inequality. To thoroughly examine biases, we introduce a new dataset by merging four separate datasets. This dataset allows us to study biases across ten different areas listed below:

- Race/Color : When someone judges by race or skin color, he or she is simply making assumptions about an individual or group because of their ethnic background which can be called racial discrimination. This bias can manifest in various forms, such as racial profiling, discrimination in hiring, and unequal treatment in the criminal justice system.

- Socioeconomic : Socioeconomic status bias refers to biases that are based on a person's social or financial status. Stereotypes about individuals from lower socioeconomic backgrounds, such as presumptions about their morality, work ethic, or IQ, might result from this bias.

- Gender: Gender discrimination encircles prejudicial views coupled by gender identity or gender expression. This bias may lead to unfair treatment in contexts like work, school, and leadership roles. It can also contribute to the perpetuation of gender stereotypes and discrimination.

- Disability : Discriminatory attitudes or actions that marginalize or exclude people with mental, cognitive, or physical health impairments constitute bias against people with disabilities. This bias has the potential to cause discrimination, exclusion, and the encounter of obstacles in several of aspects of life, such as work, school, and interpersonal relationships.

- Nationality : Prejudice or discrimination against individuals or groups on the basis of their nationality or citizenship status is known as bias based on nationality or citizenship. This prejudice can cause fear or dislike of people from other countries, unfair ideas about them, and treating them unfairly.

- Sexual Orientation : Bias related to this category involves stereotypes or discrimination based on a person's sexual orientation or identity. This bias can lead to harassment, discrimination, and unequal treatment of LGBTQ+ individuals in various kinds of fields, including employment, healthcare, and housing.

- Physical Appearance : Prejudices or preconceptions about people based on their physical attributes, such as their height, weight, facial features, or shape, are examples of bias based on physical appearance. In contexts like work, relationships, and social interactions, this bias can result in body shaming, stereotyping, and discrimination.

- Religion : Religious biases are prejudices or preconceptions formed based on a person's religious beliefs or behaviors. Due to this prejudice, people or groups may be excluded because of their lack of faith or religious discrimination, intolerance, or both.

- Age : Age bias is when someone are stereotyped or discriminated against because of their age or generational cohort. This prejudice can take the form of ageism, which results in unfair treatment, discriminatory attitudes, and barriers to opportunities for older or younger people in contexts like employment, healthcare, and social interactions.

- Profession : Professional bias refers to preconceptions or stereotypes formed because of a person's occupation or career choice. This prejudice can result in unfair treatment in areas like hiring, promotions, and social status, as well as the devaluation of particular professions and the development of stereotypes about people based on their industry or job title.

We aim to understand biases' fundamental causes, expressions, and effects on people as individuals, social groupings, and society by looking at biases in each of these aspects. Furthermore, we investigate the use of sophisticated natural language processing methods in identifying and reducing stereotypical biases in textual data, including pretrained language models like BERT and RoBERTa. By utilizing an interdisciplinary approach that incorporates perspectives from computer science, sociology, and psychology, we hope to advance our knowledge of stereotyped biases and provide solutions for advancing inclusivity, equity, and fairness in social interactions and AI technology.

II. Literature Review

In recent studies as seen in table I, several researchers have delved into the critical issue of bias assessment in pre-trained LLMs. A study by Nadeem et al. [3] sought to quantify and evaluate stereotyped biases in four important areas: profession, race, gender, and religion. They analyzed models like BERT, RoBERTA, XLNet, and GPT2 using the Context Association Test (CAT) and Idealized CAT (ICAT) score methods. Their methodical approach to measurement, the dataset's public accessibility, and the creation of a leaderboard to track progress all helped to increase our understanding of biases. Nevertheless, issues were identified, including the wide range of demographics represented in the dataset construction, the possibility of false preconceptions, and the complexity of locating sentences that defy stereotypes. Nonetheless, their findings revealed GPT2 as the best model, outperforming others by 27.0 ICAT points.

Similarly, Nangia et al. [10] highlighted stereotypes associated with nine forms of social prejudice in the US while concentrating on examining and quantifying social biases in pretrained masked language models (MLMs). They employed CrowS-Pairs, a dataset of cases intended to draw attention to stereotyped characteristics, as part of their technique. The paper acknowledged many limitations, including the dataset's specificity to US biases and its potential impact on downstream tasks, but it also gave a clear measure of these biases and possible avenues for debiasing efforts. Significant bias was found in various models, including ALBERT, RoBERTa, and BERT. BERT had the lowest bias score but performed poorly on later tasks.

The goal of Manela et al. [11]'s research was to identify and address gender bias in contextual language models. Their novel measures—skew and stereotype metrics, for example—addressed gender bias in models such as BERT, RoBERTa, and ALBERT while also investigating tactics like data augmentation and online skewness mitigation. Despite their successful use of data augmentation to reduce bias, questions were raised about the potential overlap with other factors and the capture of professional biases. Their results, it's interesting to note, showed that while RoBERTa and ALBERT-xxlarge had less skew but more stereotype, models like DistilBERT and BERT showed high skew but less stereotype.

A thorough analysis of bias in models like ChatGPT and GPT-4 was conducted by Shrawgi et al. [13], with an emphasis on nationality, gender, race, and religion. The LLM Stereotype Index (LSI), a tool they developed, made it possible to evaluate prejudice in a variety of jobs. Limitations included significant processing costs and the possibility of small but detrimental effects of biases in complicated tasks, even though the approach to bias identification was flexible. Notably, prejudice was present in ChatGPT and GPT-4 in a variety of social contexts. However, GPT-4's capacity to hide biases was surpassed by its propensity to highlight major

TABLE I: Summary of Bias Assessment Studies

Author	Year	Objective	Methodology	Pros	Cons	Performance Analysis
Nadeem et al. [3]	2020	to measure and assess stereotypical biases present in pretrained language models within four key domains: gender, race, religion, and profession	CAT and ICAT score to analyze models BERT, RoBERTA, XLNet, GPT2	Implemented a systematic measurement approach, made the dataset publicly available, and established a leaderboard to monitor advancements	Diversity of demographics represented in the dataset creation process, the potential for inaccurate or misleading stereotypes, and difficulty in finding anti-stereotype sentences	the best model (GPT2) behind an idealistic language model by 27.0 ICAT points.
Nangia et al. [10]	2020	to analyze and measure social biases in pre-trained masked language models (MLMs) through a benchmark dataset that highlights stereotypes related to nine types of social bias in the United States	CrowS-Pairs, a dataset containing 1,508 examples. Each example will consist of two sentences - one that is more stereotypical and one that is less stereotypical.	Provided a clear measure of stereotyping tendencies in language models and can help guide future debiasing efforts	Dataset is limited to US-specific biases and may not encompass all cultural contexts. There could be a potential decrease in performance on natural text when models are debiased	BERT, ALBERT, and RoBERTa all exhibit significant bias. BERT has the lowest bias score but does not perform as well on downstream tasks. Training the models on more diverse data sources may introduce additional societal bias.
Manela et al. [11]	2021	to examine and lessen gender bias in contextual language models by introducing innovative measurements while considering the balance between various biases in the WinoBias pronoun resolution task	Created skew and stereotype metrics to measure gender bias in models such as BERT, RoBERTa, and ALBERT, as well as studied two methods to address bias: Online Skewness Mitigation and Data Augmentation	Explored different strategies to diminish bias, demonstrating some success with Data Augmentation in decreasing both skew and stereotype and offered insights into model behavior	Datasets may not adequately capture professional biases and there is a risk of biases overlapping with other factors	DistilBERT and BERT demonstrate significant skew but minimal stereotype, whereas RoBERTa and ALBERT-xxlarge show less skew but increased stereotype.
Barikeri et al. [12]	2021	to come up with a complete set of standards that can be used to determine and minimize prejudice in social interactions involving machine learning algorithms	REDDITBIAS, which includes human conversations from Reddit as an example dataset for detecting bias in conversational language models such as DialoGPT	Offers a real-world dataset for multi-dimensional bias analysis; presents an assessment framework that gauges dialogue task performance as well as bias	Pretraining biases still exist despite debiasing attempts, REDDITBIAS may not cover all forms of societal bias	DialoGPT exhibited significant religious bias, REDDITBIAS catalyzes further research in bias mitigation in conversational AI and ethical dialog systems.
Liang et al. [4]	2021	to improve the equity of large-scale pretrained language models (LMs) by the detection and reduction of representational biases in text creation.	AI-NLP, a way to find tokens that are bias-sensitive. To evaluate bias prevention while maintaining the quality of text production, use empirical data and human review	Provides a fresh approach to lessen representational biases while preserving text generation's quality and context	Performance may be impacted by debiasing, and the range of biases addressed may be constrained	improves fairness from 393 to 400 while maintaining the former's near to 5-scoring generated text
Shrawgi et al. [13]	2024	to examine and measure bias in models such as ChatGPT and GPT-4, specifically looking at factors like nationality, gender, race, and religion.	Developed the LLM Stereotype Index (LSI), a tool based on the Social Progress Index, which allows us to assess bias across various tasks of varying difficulty	Offered a comprehensive and adaptable way to detect bias in AI models, showcasing a consistent pattern of unfairness across different versions of the models	The study's scope and reproducibility are limited by high computational costs, and biases in complex tasks could lead to subtle but harmful effects	ChatGPT and GPT-4 both show bias across various social aspects. GPT-4 is better at concealing biases but may reveal more significant ones when they surface. LLaMA2-7B, the lone open-source model assessed, struggled on these tasks, suggesting a widespread problem with Language Models' biases in diverse situations.

biases when they did appear.

A thorough set of guidelines was put forth by Barikeri et al. [12] with the goal of identifying and reducing bias in social interactions employing machine learning algorithms. Multi-dimensional bias analysis was made easier by their use of the REDDITBIAS dataset, albeit there were some issues with pretraining biases and the dataset's coverage of societal bias. In particular, DialoGPT showed notable religious bias, highlighting the need for more study on ethical dialog systems and conversational AI's ability to mitigate bias.

Lastly, Liang et al. [4] reduced and identified representational biases in text generation in order to address the equity of large-scale pretrained language models. Their method, called A-INLP, was centered on empirical data-driven evaluation and bias-sensitive token recognition. Although their approach provided a novel viewpoint on bias reduction while preserving text quality, issues with debiasing's effect on performance and the limited scope of biases addressed were noted. However, their results showed gains in equity while preserving high-caliber text production, highlighting the significance of continuous work in bias reduction in language models.

III. Research Methodology

The steps towards analyzing and mitigating stereotypical biases in large language models, with data processing and model-based bias detection being part of it, are discussed in the following research flow diagram 1.

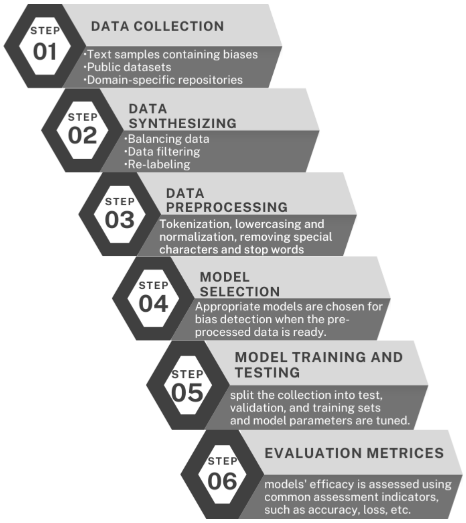

STEP
01
DATA COLLECTION
·Text samples containing biases
·Public datasets
·Domain-specific repositories
STEP
02
DATA SYNTHESIZING
·Balancing data
·Data filtering
·Re-labeling
STEP
03
DATA PREPROCESSING
Tokenization, lowercasing and normalization, removing special characters and stop words
STEP
04
MODEL SELECTION
Appropriate models are chosen for bias detection when the pre-processed data is ready.
STEP
05
MODEL TRAINING AND TESTING
split the collection into test, validation, and training sets and model parameters are tuned.
STEP
06
EVALUATION METRICES
models' efficacy is assessed using common assessment indicators, such as accuracy, loss, etc.

Fig. 1: Research flow

From data collection to a user interface that enables bias assessment on individual input words, the flow directs the whole process.

- Data Collection: Gathering a range of datasets including text samples with possible stereotyped biases is the first stage. These datasets come from news articles, social media, public resources, and specialist repositories. The goal is to compile a wide range of texts that illustrate various situations and contain examples of recognized biases. Here in this paper, we have used the datasets from the paper [14], [3], [15] and [10]. The gathered data was then converted into the form in which we would like to use. The gathered dataset has a total of 10 types of biases such as race color, socioeconomic, gender, disability, nationality, sexual orientation, physical appearance, religion, age, and profession.
- Data synthesizing: To increase the quality and diversity of data, data synthesizing entails either improving the existing datasets or producing new samples from them. To increase the dataset and reduce innate biases, this step is essential. It comprises several crucial procedures:
  1) Data augmentation: By rewriting, paraphrasing, or changing the original text, this method produces more samples. It improves generalization during model training by expanding the amount and heterogeneity of the sample.
  2) Data Balancing: Data balancing resolves any inequalities in the dataset to guarantee a balanced representation of various populations.
  3) Eliminating Bias from the Dataset: This procedure uses a variety of techniques to reduce the dataset's innate biases. It could involve re-labeling and human screening.

  Although we tried to create the most balanced data, data for some of the biases such as disability is quite less than other types. After applying the techniques for data synthesizing on the gathered data from various resources, the final dataset is created which will be further utilized for the bias analysis. The final dataset is publicly available [16].
- Data preparation guarantees that the dataset is clear, consistent, and suitable for use with big language models. It is an essential step in getting the dataset ready for training. It entails several changes that enhance the text's quality and better prepare it for machine learning algorithms. Tokenization, lowercasing and normalizing, and the elimination of stop words and unusual characters are the main elements of data preparation.
  - The process of dividing text into smaller pieces, or tokens, is called tokenization. Depending on the tokenization approach, tokens can represent single letters, subwords, or whole words. When text is divided into individual words to aid in understanding context and word connections by models is called word tokenization. Character tokenization: Break text up into individual characters; helpful for some jobs requiring in-depth text structure research. Big language model's needs and the dataset's characteristics determine which tokenization approach should be used.
  - To minimize variability and guarantee consistency throughout the dataset, lowercasing and normalizing, are used to normalize the text. Lowercasing is used to prevent disparities brought about by case differences, and convert all text to lowercase. This

stage aids in the model's concentration on the real material instead of case variations. Normalization is the process of standardizing text by getting rid of irregularities like excess spaces, punctuation, or special characters. This procedure guarantees that the text in the dataset is represented consistently. Normalization includes lemmatization or stemming. Lemmatization: To increase consistency, condense words to their most basic or root form. Stemming: This process is comparable to lemmatization, except it eliminates affixes from words to get them at their base.
  - Eliminating stop words and special characters makes the text easier to read and concentrates on important information. Punctuation, stop words, and special characters can introduce noise into the data and lower the effectiveness of the model. Getting rid of any non-alphanumeric characters, punctuation, and other extraneous items that don't add anything meaningful to the text is necessary. Deleting popular terms that don't have much significance in numerous instances, such as ”the,” ”and,” ”is,” etc is essential. By concentrating on more pertinent phrases, stop words can be removed to enhance model performance.

  In general, pre-processed data guarantees that the data is uniform, clean, and appropriate for training big language models, which enhances the models' dependability and accuracy in identifying and evaluating biases.

- In AI-related projects, choosing a model is a crucial first step, particularly for complicated tasks like bias detection and text categorization. With the correct model, insightful insights, economical resource use, and precise forecasts may all be achieved. However, making a bad decision might lead to inaccuracies, inefficiencies, and resource waste. The reasons why selecting a model becomes so important:
  - Performance and Accuracy
  - Resource Optimization
  - Task Suitability
  - Flexibility and scalability
  - Project requirements and constraints

  RoBERTa and BERT were chosen as the main models for bias detection and text categorization in this investigation. These models were chosen for usage because of their reliable performance, capacity to comprehend intricate linguistic patterns, and adaptability when it comes to fine-tuning. Because both models have already been trained, they offer a strong basis for additional training using our particular datasets. By using these models, we want to develop a strong framework that can effectively detect and classify stereotyped biases in huge datasets and classify text. These models' adaptability enables adaptation and optimization to satisfy the particular objectives of the research.

- Model training and testing: For LLMs to be effective in detecting and categorizing biases, the training and testing procedure is essential. This section outlines the procedures for training the models using preprocessed data and assessing their generalization abilities using unseen data. Using our processed dataset, we fine-tune the pre-trained models that have been chosen. The dataset is split up into three different subsets: 80% for training, 10% validation, and 10% for testing, to guarantee effective training and the resilience of the models.
  1) Training set: By using it to train the models, a range of instances are provided for them to learn from. The models modify internal parameters during training to maximize their performance for the classification.
  2) Validation set: The validation set provides intermediate evaluation during training, allowing for the best results to be obtained by adjusting model parameters like learning rate, batch size, and number of epochs.
  3) Testing set: This set is essential for assessing the trained models' performance for generalization. It enables us to evaluate the models' performance on inference data and their precision in identifying and categorizing biases.

  Our objective in using this approach for training and testing models is to create models that are not only good at identifying biases but also have good cross-dataset and cross-context generalization. The testing phase's outcomes direct subsequent enhancements and adjustments, guaranteeing the models' suitability for real-world uses in bias detection and text categorization.

- Evaluation metrics: Metrics for evaluation are crucial in determining how well our models work. They offer measurable metrics for assessing how successfully a model categorizes or predicts results. The models' efficacy is evaluated using the following metrics in the context of bias detection and text classification:
  - Accuracy: It is computed by taking the total number of predictions (false positives, false negatives, true positives, and true negatives) and dividing it by the number of right predictions (true positives and true negatives). Accuracy gives a broad impression of correctness, but as it doesn't differentiate between different kinds of mistakes, it might not always be the optimal metric when there is a class imbalance.
  - Precision: The ratio of actual positive predictions to all positive predictions the model returned is known as precision. It aids in assessing how accurate the model's favorable predictions are. When bias identification is involved, precision is especially helpful since false positives can be costly and result in misleading results if non-biased content is mistakenly classified as biased.
  - Recall: The ratio of true positive predictions to all real positives in the dataset is called recall, which is often referred to as sensitivity or true positive rate. A high recall score indicates that the intended results may be detected by the model with effectiveness. A high recall score indicates that the intended results may be detected by the model with effectiveness.

- F1-score: The F1-Score is a balanced statistic that takes into account both precision and recall. It is calculated as the harmonic mean of these two variables. It can provide an overall model efficacy in a single number and is helpful when recall and precision need to be matched. When both false positives and false negatives are significant factors, a high F1-Score suggests that the model has successfully struck a solid balance between accuracy and recall, making it a trustworthy metric.

IV. Proposed Framework

A thorough strategy to overcome the aforementioned prejudices is provided by a suggested framework for assessing and detecting stereotyped biases in LLMs. This section contains a thorough description of a suggested framework. The algorithm 1 shows all the stages of the implementation. The proposed framework is shown using figure 2. The description of each layer is explained in this section.

A. Data pre-processing layer

Preparing data for further investigation and model training involves several fundamental tasks, all of which are included in data preparation. The goal of the several subtasks involved in this process is to guarantee that the data is consistent, well-organized, and appropriate for the planned analysis. These subtasks include data collecting, dataset building, and data cleaning.

1) Dataset description: The dataset used in this paper [16] is created through various publically available datasets. The process of dataset creation is mentioned in the research methodology section. Here, we have described what the data looks like. The dataset contains a total of 28,000 rows each row contains a sentence as well as the bias present in that sentence. The dataset has the encoded class labels that are represented by the table II. The bar plot3 shows the distribution of the data.

TABLE II: Bias Types and Labels

Bias Type	**Label**
Race Color	0
Socioeconomic	1
Gender	2
Disability	3
Nationality	4
Sexual Orientation	5
Physical Appearance	6
Religion	7
Age	8
Profession	9

B. Classification layer

Text classification challenges in NLP have been revolutionized by huge language models such as BERT and RoBERTa. The last layer in these systems is the classification layer when the model produces predictions using the features it has learned from earlier levels. In this section, we examined BERT and RoBERTa's architectures, showing how they assist classification jobs and going over the fine-tuning procedure.

1) BERT: Based on the Transformer architecture, BERT processes text using attention methods. Here's a summary of the essential elements: Transformer Encoder: Depending on the model type (BERT-base or BERT-large), BERT consists of many Transformer encoder layers, usually 12 or 24. A feedforward neural network and a multi-head self-attention mechanism are features of each encoder layer. Bidirectional Context: BERT is taught to concurrently comprehend context from the left to the right and the left to the right. Because it is bidirectional, BERT may gather extensive contextual data. Masked Language Model (MLM): Using a masked language model technique, which involves masking some words and teaching the model to anticipate the terms that are missing, BERT is pre-trained. The model gains an understanding of linkages and contextual signals from this pre-training exercise. Usually placed on top of the last Transformer encoder layer, the classification layer in BERT is thick. The dense layer receives the output from the previous encoder layer and uses it to generate the output for classification tasks. This thick layer is trained alongside the rest of the model during fine-tuning so that it can adapt to certain tasks, such as text categorization.

2) RoBERTa: With a few significant changes, RoBERTa is an enhanced BERT variant. No Next-Sentence Prediction: RoBERTa just trains on the masked language model, without the next-sentence prediction task that BERT includes in its pre-training phase. RoBERTa can now learn from the pre-training data more effectively thanks to this modification. Extended Pre-training on More Data: RoBERTa is pre-trained for a greater number of epochs and on a larger dataset, which produces a more resilient model. An improved comprehension of linguistic patterns is provided by this prolonged pre-training. RoBERTa employs dynamic masking, in which the masked tokens are switched during pre-training on each epoch. By using this method, the model is exposed to more diverse scenarios. RoBERTa employs dynamic masking, in which the masked tokens are switched during pre-training on each epoch. By using this method, the model is exposed to more diverse scenarios. A thick classification layer is placed on top of the final encoder layer in RoBERTa, the same as in BERT. This thick layer receives its input from the output of the previous Transformer layer.

3) Fine tuning: The process of developing pre-trained models, such as BERT and RoBERTa, for particular tasks is called fine-tuning. It entails further training on a smaller dataset associated with the desired application. The model is trained on the labeled training dataset during the fine-tuning process. The rest of the model is trained in tandem with the dense classification layer, which is modified to fit the particular task. To attain the best results, fine-tune variables like as learning rate, batch size, and number of epochs which is known as hyperparameter tuning. Overfitting is avoided and progress is tracked by validation sets. he model's performance in the particular job is evaluated by testing it on validation and test sets after its fine-tuning. As a result, the classification layer serves as a link between the final task-specific output and the Transformer model that has already been trained.

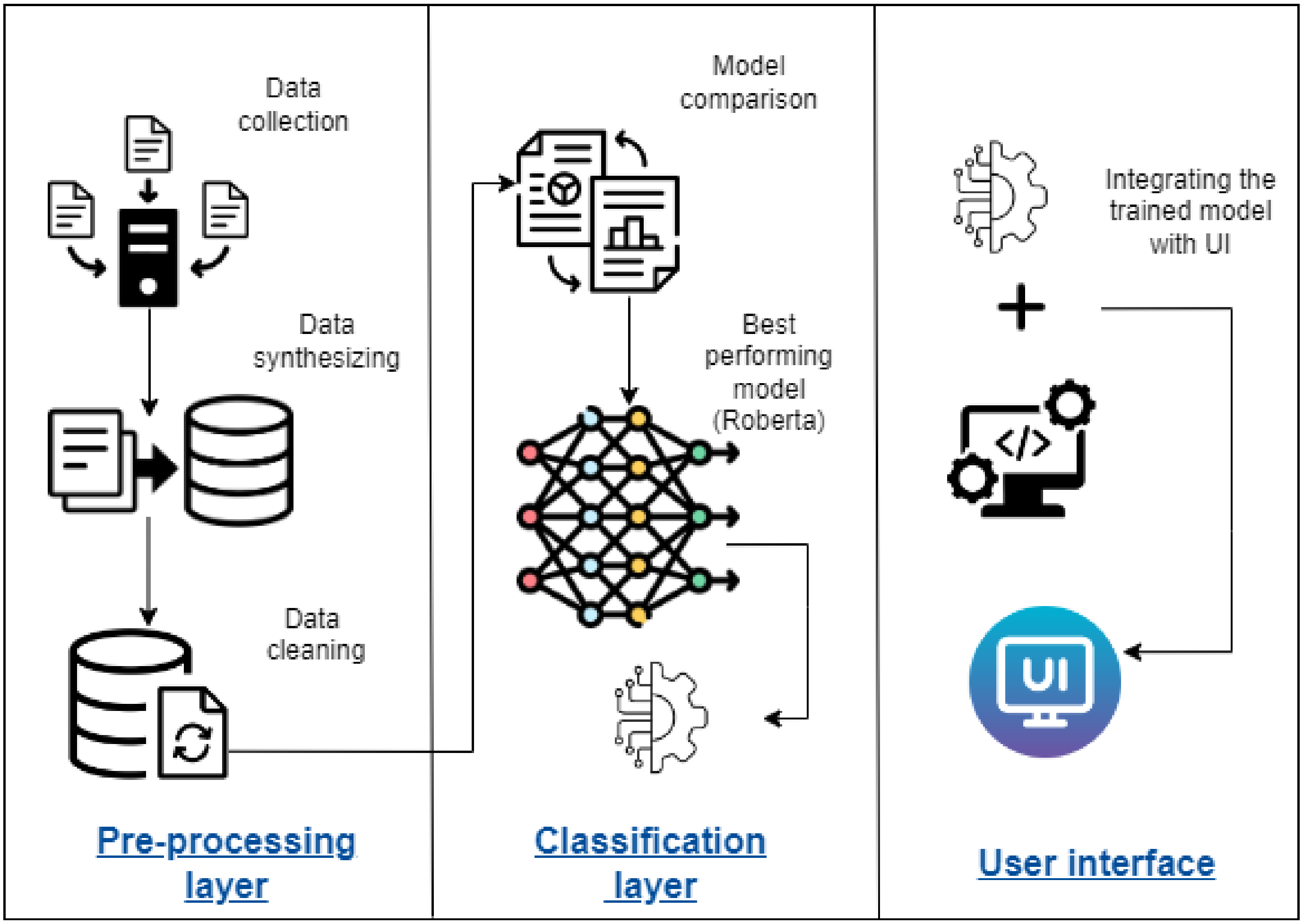


Fig. 2: Proposed model

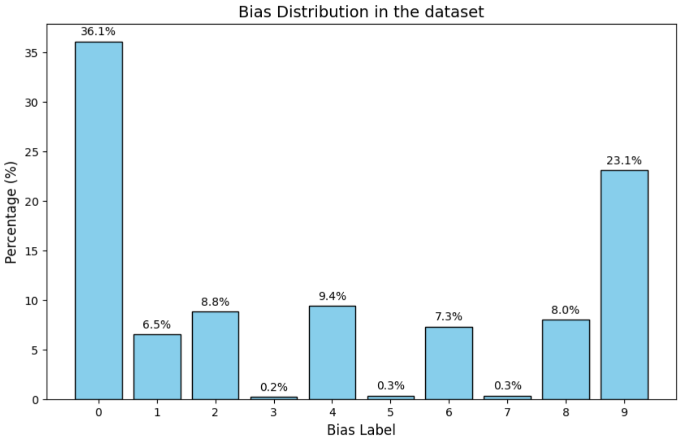


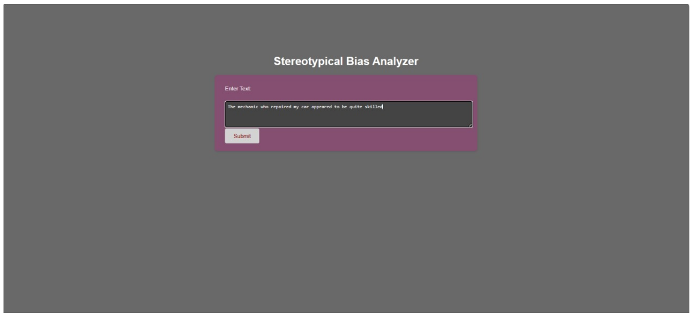
Fig. 3: Bias distribution in the dataset

C. User interface

The point of contact between end users and the trained model is the UI layer. In this part, we explain how to provide an interface for users to engage with the model for text categorization and bias detection. We used how Hugging Face is used to create a pipeline of pre-trained models which can be used further to connect it to the user interface for smooth interaction. By removing the complexity of communicating with the model, the Hugging Face pipeline offers a simple-to-use interface for analyzing text and producing predictions. To make sure the pipeline is ideal for the particular tasks of bias detection and text classification, it is adjusted using the pre-trained model. The model pipeline may be interacted with by users through the user interface layer. Through this connection, users may submit text and get feedback regarding possible biases or categorization outcomes. Numerous frameworks and technologies can be used to integrate the Hugging Face process with the user interface. However, in our work, we used the flask framework which also makes it possible to create web-based user interfaces. We used HTML, CSS, and JavaScript to build the front end. Users can submit text for analysis over this link, evaluate the results, and check various facets of text categorization and bias detection. Figure 4 and 5 shows the user interface that we built.

V. Results and Discussion

The test dataset yielded an impressive accuracy of 0.9832 for our bias identification model, demonstrating the model's effectiveness in accurately recognising instances of biases in a variety of scenarios. This high degree of accuracy highlights

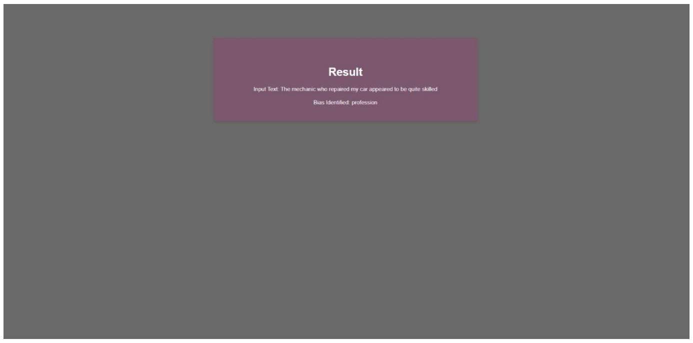


Stereotypical Bias Analyzer

Enter Text:

The mechanic who repaired my car appeared to be quite skilled

Fig. 4: User interface to take input from user



Result

Input Text: The mechanic who repaired my car appeared to be quite skilled

Bias Identified: profession

Fig. 5: Result shown to the user for the given text

the model's resilience and potency in identifying minute linguistic clues that point to biases in text written in natural language.

A. Model Performance and Identification of Bias

Our bias recognition model's exceptional accuracy is a testament to its capacity to detect subtle biases that are incorporated into natural language text. This capacity is crucial for tackling the widespread problems of prejudice and discrimination that may be found in news stories, social media posts, and online forums, among other communication mediums.

1) Precision and Recall Analysis: It is crucial to examine precision and recall numbers for every bias category in addition to overall accuracy. Recall is the percentage of correctly detected biases out of all actual occurrences of bias, whereas precision is the percentage of correctly identified biases out of all cases anticipated to be biased. This analysis shows the model's strengths and weaknesses and offers insights into how well it performs across various bias dimensions.

B. Dimension-specific Performance Evaluation

1) Strengths and Weaknesses Across Dimensions: The model's performance varied according to the many degrees of bias, as our analysis showed. For example, the model might be quite accurate at detecting biases based on gender or ethnicity, but it might be less accurate at recognising biases based on age or occupation. Comprehending these advantages and disadvantages is essential for formulating focused approaches to enhance model efficacy and tackle particular issues linked to distinct bias facets.

Algorithm 1 Text Classification and Bias Detection

```
Require: Raw dataset $D$
Ensure: Trained model and user interface for bias detection
Data Preprocessing Layer
$D_{combined} \leftarrow$ combine_datasets($D$)    ▷ Gather datasets from various sources
$D_{stratified} \leftarrow$ stratified_sampling($D_{combined}$)    ▷ Stratified sampling for balanced data
for all $sample \in D_{stratified}$ do    ▷ Preprocess each text sample
    $sample \leftarrow$ tokenize($sample$)    ▷ Break down text into tokens
    $sample \leftarrow$ lowercase($sample$)    ▷ Standardize to lowercase
    $sample \leftarrow$ normalize($sample$)    ▷ Standardize text format
    $sample \leftarrow$ remove_stopwords($sample$)    ▷ Remove common stop words
    $sample \leftarrow$ remove_special_chars($sample$)    ▷ Remove unwanted characters
end for
Model Training Layer
$D_{train}, D_{val}, D_{test} \leftarrow$ split_dataset($[0.8, 0.1, 0.1]$)    ▷ Split into training, validation, and test sets
$model \leftarrow$ initialize_model()    ▷ Initialize pre-trained RoBERTa or BERT
train(model, D$_{train}, D_{val}$)    ▷ Fine-tune the model on the training set with validation
User Interface Layer
$pipeline \leftarrow$ create_pipeline($model$)    ▷ Create pipeline with Hugging Face
integrate_with_UI(pipeline)    ▷ Integrate the pipeline with the user interface
user_input $\leftarrow$ get_user_input()    ▷ Get text input from the user
$result \leftarrow$ pipeline($user\_input$)    ▷ Get bias detection results from the pipeline
display_result($result$)    ▷ Display results on the user interface
```

2) Implications for Bias Mitigation Efforts: There are important ramifications for bias mitigation efforts across several domains from the dimension-specific performance evaluation. Stakeholders can successfully address issues of prejudice by prioritising resources and activities in areas where biases are more common or difficult to detect. Moreover, the analysis's conclusions might guide the creation of specialised interventions and educational programmes that try to raise people's knowledge of and sensitivity to biases in a variety of contexts.

C. Generalization and Real-world Applicability

1) Real-world Scenario Simulation: Evaluating the model's practical applicability requires assessing its capacity to generalize in simulated real-world circumstances. We may learn more about the model's adaptability to novel data distributions, linguistic styles, and cultural quirks that are frequently found in real-world communication settings by exposing it to a variety of realistic contexts.

2) Deployment Considerations and Challenges: Although our approach shows encouraging outcomes in controlled experimental settings, there are several obstacles to overcome before implementing it in practical situations. These include concerns about interpretability, algorithmic bias, and data privacy. To assure the responsible and ethical deployment of bias detection tools, addressing these issues calls for a multidisciplinary approach that incorporates insights from computer science, ethics, law, and the social sciences.

VI. Limitations

While our study demonstrates promising results in bias identification, we acknowledge the inherent challenge posed by biased training data. Biases within the training set can stem from various sources, including societal prejudices, algorithmic biases, and sampling biases. These biases may result in skewed representations of certain demographic groups or perspectives, potentially leading to inaccurate predictions by our model, another limitation to consider is the number of bias classes represented in our model.

VII. Conclusion

In conclusion, our study provides valuable insights into the pervasive nature of stereotypical biases across various dimensions of identity and experience. We have illustrated the effects of prejudices against people based on race, gender, disability, nationality, sexual orientation, and other characteristics on individuals, communities, and society through an extensive analysis.Our analysis also demonstrates the potential of advanced methods in identifying stereotypical biases in textual data by looking at pretrained language models like BERT and RoBERTa. One key finding of our experiment is the higher accuracy of RoBERTa compared to BERT, suggesting the importance of model architecture and training strategies in addressing biases in NLP tasks. This underscores the need for continued research and innovation in developing AI technologies that are more inclusive, equitable, and fair.

References

[1] J. Albrecht, E. Kitanidis, and A. J. Fetterman, "Despite" super-human" performance, current llms are unsuited for decisions about ethics and safety," *arXiv preprint arXiv:2212.06295*, 2022.

[2] J. Zhao, T. Wang, M. Yatskar, V. Ordonez, and K.-W. Chang, "Gender bias in coreference resolution: Evaluation and debiasing methods," *arXiv preprint arXiv:1804.06876*, 2018.

[3] M. Nadeem, A. Bethke, and S. Reddy, "Stereoset: Measuring stereotypical bias in pretrained language models," *arXiv preprint arXiv:2004.09456*, 2020.

[4] P. P. Liang, C. Wu, L.-P. Morency, and R. Salakhutdinov, "Towards understanding and mitigating social biases in language models," in *International Conference on Machine Learning*, pp. 6565–6576, PMLR, 2021.

[5] Z. He, B. P. Majumder, and J. McAuley, "Detect and perturb: Neutral rewriting of biased and sensitive text via gradient-based decoding," *arXiv preprint arXiv:2109.11708*, 2021.

[6] P. Schramowski, C. Turan, N. Andersen, C. A. Rothkopf, and K. Kersting, "Large pre-trained language models contain human-like biases of what is right and wrong to do," *Nature Machine Intelligence*, vol. 4, no. 3, pp. 258–268, 2022.

[7] M. G. Haselton, D. Nettle, and P. W. Andrews, "The evolution of cognitive bias," *The handbook of evolutionary psychology*, pp. 724–746, 2015.

[8] A. Bertrand, R. Belloum, J. R. Eagan, and W. Maxwell, "How cognitive biases affect xai-assisted decision-making: A systematic review," in *Proceedings of the 2022 AAAI/ACM Conference on AI, Ethics, and Society*, pp. 78–91, 2022.

[9] C. Rastogi, M. Tulio Ribeiro, N. King, H. Nori, and S. Amershi, "Supporting human-ai collaboration in auditing llms with llms," in *Proceedings of the 2023 AAAI/ACM Conference on AI, Ethics, and Society*, pp. 913–926, 2023.

[10] N. Nangia, C. Vania, R. Bhalerao, and S. R. Bowman, "CrowS-pairs: A challenge dataset for measuring social biases in masked language models," in *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)* (B. Webber, T. Cohn, Y. He, and Y. Liu, eds.), (Online), pp. 1953–1967, Association for Computational Linguistics, Nov. 2020.

[11] D. de Vassimon Manela, D. Errington, T. Fisher, B. van Breugel, and P. Minervini, "Stereotype and skew: Quantifying gender bias in pre-trained and fine-tuned language models," in *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: Main Volume*, pp. 2232–2242, 2021.

[12] S. Barikeri, A. Lauscher, I. Vulić, and G. Glavaš, "Redditbias: A real-world resource for bias evaluation and debiasing of conversational language models," *arXiv preprint arXiv:2106.03521*, 2021.

[13] H. Shrawgi, P. Rath, T. Singhal, and S. Dandapat, "Uncovering stereotypes in large language models: A task complexity-based approach," in *Proceedings of the 18th Conference of the European Chapter of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 1841–1857, 2024.

[14] M. Kamruzzaman, M. M. I. Shovon, and G. L. Kim, "Investigating subtler biases in llms: Ageism, beauty, institutional, and nationality bias in generative models," *arXiv preprint arXiv:2309.08902*, 2023.

[15] W. Zekun, S. Bulathwela, and A. S. Koshiyama, "Towards auditing large language models: Improving text-based stereotype detection," *arXiv preprint arXiv:2311.14126*, 2023.

[16] "Priyapatel/bias_identification · datasets at hugging face." https://huggingface.co/datasets/PriyaPatel/Bias_identification. (Accessed on 04/28/2024).