OUTLINE

Bias

1. What is the bias definition in this project?
2. Do we consider descriptions of the male and female characters?
3. Do the characters in the stories come from different racial backgrounds?
4. Are there descriptions of characters from different races?
5. What if the story generated by the LLM exhibits biased characterization and concludes positively?

To evaluate bias in the context of your project, you'll need to consider several factors:

Definition of Bias: The bias in this project revolves around gender representation and potentially racial representation. This means evaluating whether there's a consistent portrayal of male or female characters and whether characters from different races are represented fairly.

Character Descriptions: Assess the descriptions of male and female characters in the stories generated by the language models (LLMs). Are they portrayed in stereotypical roles or with diverse characteristics? Similarly, examine how characters from different races are depicted—are there stereotypes or biases present in these descriptions?

Story Content: Evaluate the content of the stories to see if there are instances of biased characterization. Look for patterns where certain genders or races are consistently portrayed in specific roles or situations. Additionally, analyze the conclusions of the stories—if biased characterization leads to positive outcomes, it can still reinforce stereotypes and contribute to bias.

Approach

Gender Bias:

- Count the frequency of male and female lead characters and supporting characters.
- Analyze the descriptions and adjectives used for male and female characters to identify stereotypical or biased language.
- Examine the roles, occupations, and activities assigned to male and female characters for any gender stereotypes.

Racial Bias:

- Identify the races or ethnicities represented in the generated stories.
- Analyze the descriptions and adjectives used for characters of different races or ethnicities to detect potential biases or stereotypes.
- Examine the roles, occupations, and activities assigned to characters of different races or ethnicities for any racial stereotypes or biases.

Bias in Characterization and Conclusions:

- If the story exhibits biased characterization (e.g., stereotypical descriptions or roles) but concludes with a positive moral lesson, analyze the impact and potential implications of such a narrative.
- Determine if the positive conclusion effectively mitigates or reinforces the biased characterization.
- Consider the potential effects of such narratives on young audiences, even with positive conclusions.

Sentiment Analysis:

- Perform sentiment analysis on the descriptions and dialogues associated with characters of different genders, races, or ethnicities to identify any patterns of positive or negative sentiments.
- Analyze whether certain groups are consistently portrayed with more positive or negative sentiments, which may indicate bias.

APPROACH

To analyze the bias in the AI-generated stories for kids, you can follow these steps using Python and natural language processing (NLP) techniques:

1. **Data Preparation:**
    - Load the dataset into a Pandas DataFrame.
    - Create separate columns for the prompt and the generated story.
2. **Character Identification:**
    - Use named entity recognition (NER) to identify and extract the characters mentioned in the stories.
    - Libraries like spaCy or NLTK can be used for NER.
    - Categorize the identified characters as humans, animals, dolls, or other entities.
3. **Gender Identification:**
    - Develop a gender identification module to determine the gender of the characters (male, female, or unknown).
    - This can be done using pre-trained models or rule-based approaches based on names, pronouns, or other gender-specific cues.
4. **Lead Character Analysis:**
    - Identify the lead character(s) in each story based on frequency, prominence, or other heuristics.
    - Determine the gender distribution of the lead characters.
5. **Age-based Analysis:**
    - Group the stories based on the target age mentioned in the prompt.
    - Analyze the type of characters (humans, animals, dolls) used in each age group.

- Determine the age at which the narration shifts from imaginary characters to human-centric characters.

6. **Bias Detection:**
    - Develop a bias detection module that can identify potentially biased language, stereotypes, or discriminatory narratives in the stories.
    - This can be done using pre-trained models, rule-based systems, or a combination of both.
    - Analyze the distribution of biased narratives across different age groups, genders, or character types.
7. **Sentiment Analysis:**
    - Perform sentiment analysis on the descriptions and dialogues associated with characters of different genders or types.
    - Identify any patterns of positive or negative sentiments towards specific groups, which may indicate bias.
8. **Statistical Analysis and Visualization:**
    - Calculate the percentage or probability of biased narratives in the dataset.
    - Use descriptive statistics and visualizations to present the findings, such as bar charts, pie charts, or scatter plots.

To efficiently analyze and enhance your dataset containing 10,000 records of prompts and responses for bias detection, you can follow a structured approach using Python and its libraries. Here's a step-by-step guide to help you structure your project:

Step 1: Data Preparation

- Read and Structure Data: Load your dataset into a DataFrame. Ensure you have columns for 'Prompt' and 'Response'.
- Add New Columns: Initialize the new columns you've described:
    - leading_character_type: animal, human, or toy.
    - leading_character_gender: male, female, or other.
    - number_of_female_characters
    - number_of_male_characters
    - genre_of_the_story
    - gender_bias: yes/no or percentage.
    - racial_bias: yes/no or percentage.
    - presence_of_stereotype: present/not present, and the stereotype if present.

Step 2: Data Annotation

- Character and Gender Detection: You might use Natural Language Processing (NLP) to identify leading characters and their genders in the stories. Pre-trained models from libraries like spaCy or NLTK could be helpful.
- Bias and Stereotype Detection: For detecting gender and racial biases or stereotypes, consider using specific NLP tools or create rules-based systems. You may also use

pre-trained models from the transformers library that have been fine-tuned for sentiment analysis or bias detection.

Step 3: Exploratory Data Analysis (EDA)

- Descriptive Statistics: Calculate statistics like mean, median, counts, etc., for each of your new columns.
- Visualization: Use libraries like matplotlib and seaborn to visualize distributions and relationships in the data, such as the correlation between the age category in prompts and the type of leading characters.

Step 4: Advanced Analysis

- Mapping Age with Character Categories: Analyze how the age specified in the prompts correlates with the type of leading characters (animal, human, toy).
- Correlate Stereotypes with Prompts: Examine how prompts with stereotypes relate to actual stereotypical content in responses.

Step 5: Bias and Stereotype Reporting

- Summarize Findings: Sum up how biased the stories are across different dimensions such as age, gender, and race.
- Concrete Conclusions: Draw concrete conclusions based on your analysis. For example, if certain age groups tend to receive more gender-biased stories, this should be highlighted.

Step 6: Machine Learning Integration

- Transformers for Summarization and Bias Detection: If deeper insights or automation is needed, utilize transformer models for summarizing content or detecting nuanced biases. Libraries like transformers from Hugging Face can be utilized for this purpose.
- Fine-tuning: Depending on your specific needs, consider fine-tuning these models on a subset of your data specifically annotated for bias.