GROUP NAME: Superheroes Beyond Stereotypes

GROUP MEMBERS: Lavanya Vaddavalli, Sharadruthi Muppidi, Shweta Parihar, Simon Sazian

GROUP LEADER: Sharadruthi Muppidi

PROBLEM AND GOAL:

What do you want to solve:

Existing superheroes in Marvel and DC comics exhibit gender and power disparity. So we intend to generate a new set of superheroes with improved diversity. We plan to fine-tune Large Language Models (LLMs), particularly GPT-2, to automate content generation with a focus on reducing gender bias and promoting creative fairness.

Why is it important:

Addressing gender bias in content generation is of paramount importance in today's world. Biased content can perpetuate stereotypes, reinforce societal inequalities, and undermine the principles of diversity and inclusivity. As creators and developers, it is our ethical responsibility to rectify such biases and contribute to a more equitable representation of gender in media and storytelling. By focusing on superheroes, who serve as iconic symbols of inspiration and influence, we aim to set an example for fair content creation across diverse domains.

What results do you expect:

We aim to generate a new set of superheroes that are free from bias and characterized by fairness-aware attributes, effectively mitigating gender bias within their portrayals, thus enriching the storytelling universe for generations to come.

FORMALIZATION INTO ML/DATA MINING TASK:

Which data type?

It’s a structured dataset containing textual, categorical and numerical features.

Which function?

This task involves a combination of text generation (where transformer based models like GPT-2 and BERT are used), natural language processing, and fairness-aware NLP techniques.

Our approach involves converting structured data into text format and passing it through a pre-trained BERT model to obtain contextual embeddings, which enhance context awareness. We chose GPT-2 for our task since it is a text generation model that can be used to generate entirely new text. Additionally,

our dataset is relatively small with just over 1400 rows, hence we are using a smaller LLM. We then leverage these embeddings to train/fine-tune GPT-2. This process enables us to perform prompt-based generation to create a new set of superheroes and then analyze biases in the new set and determine the fairness of our results.

To mitigate the bias in the subsequent stages, we adjust the training data weights for the GPT-2 model, apply bias correction algorithms based on predefined fairness criteria, and meticulously define and validate the loss function. We also incorporate sampling strategies to finely balance creativity and fairness adherence.

DATA PLAN:

What kind of data?

It's a structured dataset containing 1400+ superheroes history and powers description. It contains textual, categorical and numerical features.

Where and how do you get the data?

Data is obtained from Kaggle website.
Link: https://www.kaggle.com/datasets/jonathanbesomi/superheroes-nlp-dataset/data

SCHEDULE:

Detailed plan of the project:

Timeline	**Planned work**
Week 1 to 2	Analyze dataset, perform EDA, preprocessing, augment dataset if required. The dataset we intend to utilize for training the LLM currently consists of 1400 rows. If necessary, we plan to implement data augmentation techniques on this dataset.
Week 3 to 4	Train dataset on BERT to generate embeddings if required.
Week 5 to 7	Fine tune existing LLM on dataset and observe the results
Week 8 to 10	Compare results, fine tune hyperparameters, modify loss function, integrate fairness metrics into loss function