Revisiting The Classics:
A Study on Identifying and Rectifying Gender Stereotypes in Rhymes and Poems

Aditya Narayan Sankaran*, Vigneshwaran Shankaran*
Sampath Lonka , Rajesh Sharma

GESIS Leibniz Institute
for the Social Sciences

Introduction

- Gender roles are socially constructed positions or behaviours that are learned and performed by individuals in accordance with their gender identity and the prevailing cultural norms (Bigler and Liben, 2007).
- This learning starts when toddlers are exposed to stories and rhymes, which are fundamental learning practices for language acquisition, and also facilitate their understanding of how society functions.
- The predominant mode of knowledge acquisition stems from educational materials, such as textbooks and oral instruction.

Example of two rhymes

Beans, beans, the magical fruit.
The more you eat, the more you toot,
The more you toot, the better you feel.
So let’s have beans for every meal!

A straightforward rhyme that is taught to children, clearly conveying the benefits of eating healthy foods like beans.

Peter, Peter pumpkin eater
Had a wife but couldn’t keep her;
He put her in a pumpkin shell
And there he kept her very well.

A seemingly humorous poem about a troubled marriage.
It perpetuates patriarchal values by depicting the husband’s control over his wife

Motivation

- While these poems and rhymes offer a window into the past, the world we live in today is vastly different.
- Some concepts within these works may perpetuate stereotypes that are no longer acceptable.
- These stereotypes are often internalized by individuals from a young age and can shape their beliefs, attitudes, and behaviours towards themselves and others (Haines et al., 2016)
- Crucial for educators to be mindful of these outdated ideas and need to evaluate their content meaningfully.

This study tries to fill the gap by using various machine-learning techniques to reduce the amount of human intervention to rectify such stereotypes.

Dataset Collection

- The selection process for the creation of a comprehensive dataset of children’s rhymes and poems was designed to ensure diversity in terms of style, content, and cultural background.
- Collected rhymes and poems from a variety of publicly available published sources, after extensive consultation with educators in the field of Literature and Education.
- These sources encompassed a broad range of content, including works by renowned poets such as Shakespeare and Frost, as well as popular collections such as Mother Goose.
- In addition to rhymes & poems originally written in English, we used 20 publicly available translated poems from 11 different languages
- Total = 339 Rhymes and 322 Poems from various sources.

Disagreement Analysis

Phase 1

- The initial phase involves establishing annotation guidelines utilizing a subset of the dataset.
- Annotators conducted an annotation procedure in which they were not aware of the identity of the poems or rhymes they were annotating.
- This was done to prevent any unconscious bias from influencing their annotations and ensured annotators' annotations were as objective as possible.
- Between iterations, annotators met to discuss and adjudicate any disagreements. The disagreements primarily revolved around the choice of words and the interpretation of the certain lines.

Disagreement Analysis

For instance, for lines:

1. "One for my **master**"
2. "Wilt thou be **mine**?"

Particular attention was paid to the terms **master** and **mine,**

They hold distinct connotations in terms of ownership.

- **“mine”** could be implied as the possession of the opposite gender,
- **“master”** does not connote ownership specific to a particular gender.

Words relating to aestheticism like **pretty** also had disagreements when their usage was tied to a particular gender but were decided to be non-stereotypical due to the subjective nature of beauty

Disagreement Analysis

Phase 2

- After four iterations of disagreement analysis, a Krippendorff's α of 0.96 was attained.
- Guidelines for annotations was established by means of deliberating and evaluating the discussions that transpired between the iterations.
- With the help of the guidelines, one of the two annotators labelled the rest of the remaining data.

Data Augmentation

- Due to the class imbalance present in the dataset, Data augmentation was performed by synthesizing synonym versions of the poems and rhymes in the training set using GPT-3.5 (OpenAI, 2022).
- The following prompt was used:

"*Replace* [*nouns or subject/objects from the poem or rhyme*] *with synonyms. Keep the poem rhyme scheme and sentence formation intact forcefully*"

- This prompt specifically targeting nouns and synonyms in order to augment the text without affecting the bias present.

Augmentation Example

OG: Jack and Jill went up the hill, To fetch a pail of water.

AG: Jack and Jill went up the mountain, To obtain a bucket of water.

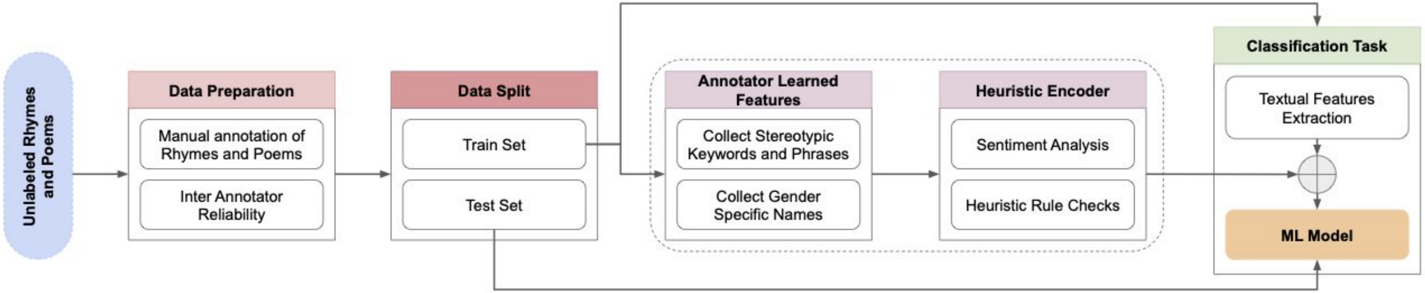
OG: And when I bake, I’ll give you a cake,

AG: And when I fry, I’ll give you a pie,

Dataset Statistic

Label	Verses	Lines
Stereotypical Rhymes	65	151
Non-Stereotypical Rhymes	274	5157
Stereotypical Poems	80	359
Non-Stereotypical Poems	242	7647
Augmented Stereotypes	290	1347
Total (Without Augmentation)	661	13314
Total (With Augmentation)	951	14661

Heuristic Encoder


Unlabeled Rhymes and Poems
Data Preparation
Manual annotation of Rhymes and Poems
Inter Annotator Reliability
Data Split
Train Set
Test Set
Annotator Learned Features
Collect Stereotypic Keywords and Phrases
Collect Gender Specific Names
Heuristic Encoder
Sentiment Analysis
Heuristic Rule Checks
Classification Task
Textual Features Extraction
ML Model

- Heuristic Encoder uses annotator-learned features to complement the input features to enhance the model's prediction ability, instead of using an external knowledge base.
- Annotators compiled a comprehensive list of words, phrases and gender-specific names that they consider stereotypical from the list of poems and rhymes that were annotated.
- Feature list, is therefore, limited to the dataset for better contextual understanding.
- This list acts as an additional source other than the labels of the text.

Heuristic Encoder

A binary valued feature vector is generated for the given text, using the annotator learned features collected as part of the annotation process.

The structure of each vector is defined as follows:

[Male Names, Female Names, Stereotypes, Negative Sentiment, Positive Sentiment]

For example, the line “**had a wife but couldn’t keep her**”

- **wife** signifying that it has female representation.
- The phrase “**couldn’t keep her”** is a stereotype signifying female ownership.
- Entire line has a negative sentiment.

Resultant vector is [0, 1, 1, 1, 0]

Identification Task

- Analyzed data using 4 different categorization schemes: monostichs (L), couplets (2L), tercets (3L), and full text (F).
- Our objective is to rectify stereotypical poems and rhymes
- Therefore a poem/rhyme is stereotypical even if only one line contains a stereotype.
- A poem/rhyme becomes a candidate for rectification if it is classified as stereotypical by the selected model.

Identification Task

- XGBoost utilised as baseline. The efficacy of the Heuristic Encoder has been evaluated in conjunction with this particular machine-learning model.
- BERT family of models have shown impressive performance in a variety of downstream NLP tasks. Therefore, $\text{BERT}_{\text{base}}$ is utilized for fine-tuning our objective of stereotype classification.
- BERT_{SS} is a BERT variant trained on StereoSet (Nadeem et al., 2021), dataset designed to quantify stereotypical bias in language models.

Identification Task (Training)

- XGBoost Training:
  - The first approach utilizes the Word Frequencies of the vocabulary present in the dataset,
  - The second approach, a binary vector was concatenated to the Word Frequencies.
- BERT Training
  - Learning Rate : 2e-5.
  - 5 Epochs.
  - Batch size : 16.
  - Token lengths for the BERT-based models were changed according to the input lengths of the approach used.

Results and Inferences

- BERT_{SS} (1L) is the best performing model in terms of all the metrics.
- 97% accuracy and a macro recall of 0.81 signifying lesser false positives.
- This intuitively makes sense since the model has an ingrained understanding of stereotypes and bias of a broader environment and here it adapts to the task of poems and rhymes classification.
- The proposed Heuristic Encoder is able to improve the model's performance by 5% with longer text input, since the avenue for checking the heuristics is more
- In shorter contexts, the addition of the Heuristic Encoder improves important metrics like precision, recall, and F1-Score by 1-2%.

Why Rectification?

- (Prosic-Santovac, 2015) argues that many rhymes were created more than a hundred years ago when society cherished somewhat different values from those in the modern day.
- Care should be exercised when choosing the rhymes to be used in teaching modern-day children.
- By rewriting classic literature, writers can help to correct these biases and create a more accurate and inclusive representation of the past.
- **Note:** Rewriting classic literature is not about erasing the past. Rather, it is about re-imagining the past in a way that is more inclusive and representative of the diverse experiences of women and other marginalized people.

Rectification Task Setting and Survey

- Poems and Rhymes identified as stereotypical were selected by an educator with over 20 years of experience in Montessori and primary education.
- The educator then rectified the poems and rhymes to suit modern sentiments.
- LLM (ChatGPT) was also employed to rectify by means of the prompt:

Change the poem to remove gender stereotypes and make sure to keep sentence formation and rhyme scheme close to the original as much as possible

Rectification Examples

Original Text : Georgie Porgie, pudding and pie; Kissed the girls and made them cry

Human Rectification : Georgie Porgie, pudding and pie; Kissed the girls and got into a fight.

ChatGPT Rectification : Georgie Porgie, friendly and kind; Shared a smile, left worries behind.

Survey

- A specific subset of 5 rhymes and 5 poems incorporating gender stereotypes was selected, with the intention of encompassing a diverse range of linguistic structures and content lengths to ensure variability.
- A survey-based statistical analysis was undertaken to examine the rectification capacity of humans compared to ChatGPT in adapting poems and rhymes to align with contemporary sentiments.
- An evaluative survey was conducted in which the participants were unaware of the identity of the rectifiers (i.e., whether human or ChatGPT).
- The rectifications were randomly shuffled and presented to the participants as Version 1 and Version 2.

Hypothesis Testing

- Based on the questionnaire, we formulated our hypothesis.

  H_0: There is no significant difference between the two versions

  H_1: There is a significant difference between the two versions.

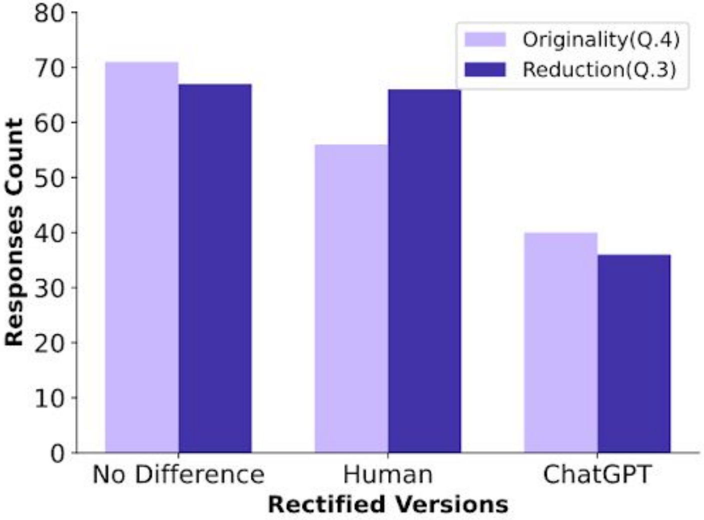
- Due to the small participant sample pool (17 participants), paired t-tests were conducted on two criteria:
  - Reduction
  - Creativity

Results

- Participants were asked to rate the level of gender stereotype reduction on a scale of 1 to 5 and to describe how creative they perceived the version to be reducing gender stereotypes and how creatively it has been done.
- Upon testing the hypothesis to compare the difference in gender stereotype reduction between human and ChatGPT rectification, it was determined that the p-value, exceeding the significance level of 0.05, led to the failure of rejecting H_0 suggesting a lack of evidence against the null hypothesis.
- Further studies with a larger sample size might be needed to detect a potential difference between the methods.

Results

- The plot shows the majority of participants found no difference in originality between human rectified and Chat GPT-rectified text.
- Only a small number believed that human rectification was more original.
- Highest number of participants felt that humans were effective in reducing gender stereotypes, an almost equal number of people believed there was no difference between humans and ChatGPT.
- Observed that ChatGPT is improving its capacity to correct rhymes and poems as well as humans.



Conclusion

- Investigated the presence of gender stereotypes in a diverse set of rhymes and poems from a variety of sources, creating an annotated dataset, which we hope, will be a valuable resource for future research and addressing gender bias in classical literature
- Gender stereotypes were rectified using large language models (LLM) and human educators and reveal the potential of LLMs in rectifying gender stereotypes by means of a survey based analysis.
- By raising awareness and promoting inclusivity in artistic expressions, this work contributes to the discourse on gender equality.