Quantifying Gender Stereotypes in Bollywood Movies: A Study of Plot and Poster Bias and Proposal for Debiased Story Generation

Abstract

This research paper focuses on the prevalent gender stereotypes and biases in the Hindi movie industry (Bollywood). The paper presents an algorithm designed to detect and remove these stereotypes from movie text. The study involves analyzing movie plots and posters for all movies released since 1970. Gender bias is detected through semantic modeling of plots at both the sentence and intra-sentence levels. Various features such as occupation, introductions, associated actions, and descriptions are used to highlight the pervasiveness of gender bias in movies. Additionally, the paper examines the centrality of each character in the derived semantic graph to observe similar biases. Interestingly, the research finds that gender bias is not as evident in movie posters, where females receive equal importance despite their limited impact on the plot.

1. Introduction

In this research paper, I explore the idea that movies serve as a reflection of our society, creatively depicting its issues, thoughts, and perceptions. With the freedom of artistic expression, they can also be an avenue for portraying gender bias and stereotypes prevalent in our culture. As such, I believe that movies can act as a valuable proxy to understand the extent of such biases in any given society. To address this topic, I leverage Natural Language Processing (NLP) and image understanding techniques for a quantitative analysis of gender bias.

To illustrate the significance of the issue, I refer to a specific plot excerpt from the Bollywood movie "Kaho Na Pyaar Hai." The excerpt portrays Rohit, an aspiring singer and a car showroom salesman under Malik (Dalip Tahil). During the plot, he meets Sonia Saxena (Ameesha Patel), who is introduced as the daughter of Mr. Saxena (Anupam Kher) when he delivers a car as her birthday present.

Analyzing this excerpt, it becomes evident that the male character (Rohit) is portrayed with both a profession and a personal aspiration, while another male character (Malik) is depicted as a business owner. On the other hand, the female character (Sonia) is introduced without any mention of a profession or aspiration, her identity seemingly reliant on her relation to another male character ("daughter of").

With this research, my goal is to quantify and analyze gender-based stereotypes by thoroughly examining the allocation of roles to males and females in movie plots, considering both intra-sentence and inter-sentence contexts along with cast information. By using a comprehensive dataset of Hindi movies released between 1970 and the present, sourced from Wikipedia, I aim to shed light on the characteristics assigned to male and female roles. Additionally, I employ deep image analytics to investigate gender bias in movie posters and previews.

1.1. Analysis Tasks

In this study of gender bias in Bollywood, the following tasks are central to my investigation

1. Occupations and Gender Stereotypes:
I will examine how males and females are depicted in their respective occupations. By comparing their portrayals, I aim to identify any disparities in how their professions are represented. This analysis will help me understand the levels of gender bias and stereotypes perpetuated in the movie industry.

2. Appearance and Description:
I will explore how males and females are described based on their appearance within movie contexts. My goal is to uncover any variations in the descriptions given to each gender, and how these differences contribute to gender stereotyping in Bollywood films.

3. Centrality of Male and Female Characters:
I intend to investigate the roles of male and female characters in movie plots to understand their centrality within the narratives. By assessing the prevalence of central male characters compared to central female characters, I hope to reveal any gender bias in movie storytelling.

4. Mentions (Image vs Plot):

To gain insight into how characters are represented, I will count the number of male and female faces featured on promotional posters and correlate this data with their mentions in the movie's plot. This combined analysis will help me understand if there is a connection between on-screen representation and significance in the storyline.

5. Dialogues:
   By comparing the number of dialogues spoken by male and female cast members in the official movie script, I will analyze whether there is any imbalance in the distribution of dialogue opportunities between genders.

6. Singers:
   I will investigate whether gender bias extends to movie songs by examining the distribution of male and female singers over time and across different movies.

7. Female-centric Movies:
   In this task, I will analyze whether the portrayal of female characters and the focus on female-centric stories have evolved in recent times. I will look for evidence of movies that predominantly revolve around female characters.

8. Screen Time:
   I aim to assess which gender, if any, receives greater screen time in movie trailers to gauge whether there are imbalances in on-screen representation.

9. Emotions of Males and Females:
   By examining the emotions most commonly displayed by male and female characters in movie trailers, I intend to determine if these emotional portrayals align with existing gender stereotypes in society.

Through these analysis tasks, I hope to gain a comprehensive understanding of the gender biases and stereotypes present in Bollywood movies and shed light on the evolving representation of gender roles in the industry.

2. <u>Data and Experimental Study</u>

2.1. Data Selection

In our analysis of Bollywood movies, we utilize three different types of data, which are as follows:

2.1.1. Movies Data:

Our dataset comprises information from all Hindi movie pages sourced from Wikipedia. This extensive dataset covers movies released between 1970 and 2017, totaling 4000 movies. From each movie page, we extract essential details such as the movie title, cast information, plot, soundtrack details, and associated images.

For the cast information, we gather data on 5058 female cast members and 9380 male cast members. To ascertain the gender of each cast member, we traverse their respective Wikipedia pages and extract the relevant gender information.

As official movie scripts were not readily available, we employ the Wikipedia plot as a suitable proxy for understanding the storyline of each movie. We firmly believe that the Wikipedia plot provides an accurate representation of the movie's narrative. If an actor holds a significant role in the film, it is highly unlikely that their presence will be omitted from the wiki plot.

By utilizing this comprehensive dataset, we aim to conduct a thorough analysis of gender bias and stereotypes prevalent in Bollywood movies.

2.1.2 Movies Scripts Data

For our analysis tasks, we will also include data from movie scripts to gain deeper insights into gender representation in Bollywood. We were able to acquire PDF scripts of 13 Bollywood movies, which are publicly available online.

2.1.3. Movie Preview Data

The average duration of these trailers is 146 seconds, with a standard deviation of 35 seconds. The videos have a frame rate of 25 FPS and a resolution of 480p. To perform my analysis, I will extract every 25th frame from each video and apply face classification techniques for gender and emotion detection.

2.2. Task and Approach

In this section, I will discuss the tasks we perform on the movie data extracted from Wikipedia and the scripts. Additionally, I will outline our approach for each task to draw meaningful inferences. Our analysis is broadly divided into four groups:

a) At intra-sentence level:
I will conduct this analysis at the sentence level, treating each sentence independently without considering its context. This task allows us to examine gender representation and stereotypes within individual sentences.

b) At inter-sentence level:
In this analysis, I will extend our examination to multiple sentences, taking the context from one sentence to inform the analysis of others. By doing so, I can better understand the broader picture of gender portrayal in movie narratives.

c) Image and Plot Mentions:
For this task, I will explore the correlation between the presence of genders in movie posters and their mentions in the plot. By analyzing both promotional materials and narrative content, I can investigate how gender representation is depicted visually and contextually.

d) At Video level:
In this task, I will perform gender and emotion detection on the frames of each movie trailer. I will apply the method introduced by Octavio Arriaga in 2017 to gain insights into how genders are represented visually and emotionally in movie previews.

I define specific tasks corresponding to each level of analysis to capture the diverse aspects of gender bias and stereotypes present in Bollywood movies. These tasks will enable me to draw meaningful conclusions about the portrayal of male and female characters and their roles in the cinematic landscape.

2.2.1. Tasks at Intra-Sentence Level

For our intra-sentence analysis, we took the necessary steps to prepare the movie plots for analysis. We utilized OpenIE (Fader et al., 2011) for co-reference resolution on the movie plot text. This co-referenced plot is then used for all subsequent analyses.

The following intra-sentence analysis tasks were performed:

1. Cast Mentions in Movie Plot:
I will use co-reference resolution on movie plot text to extract mentions of male and female cast members in the co-referred plot. This will allow me to determine how many times male and female characters are referenced in the plot. I anticipate finding a significant disparity, with male characters mentioned around 30 times, while female characters are mentioned only around 15 times. I expect this consistent ratio to persist

from 1970 to 2017, spanning nearly 50 years, highlighting the enduring gender bias in movie plots.

2. Cast Appearance in Movie Plot:
Using the Stanford Dependency Parser, I will analyze verbs and adjectives associated with male and female cast members. I expect to observe distinct patterns, where male characters are often linked to action-oriented verbs like "kills" and "shoots," while female characters are associated with emotional verbs like "marries" and "loves." Additionally, I expect adjectives used to describe males to frequently portray them as "rich" and "wealthy," whereas females will be described as "beautiful" and "attractive."

3. Cast Introductions in Movie Plot:
By employing OpenIE for capturing cast introductions, I will analyze how male and female characters are introduced in movie plots. I anticipate finding that males are usually introduced with a profession, like "a famous singer" or "an honest police officer," suggesting success and independence. On the other hand, I expect females to be introduced based on physical appearance or in relation to another male character (e.g., "daughter" or "sister of"), portraying them as dependent and not independent.

4. Occupation as a Stereotype:
I will analyze the distribution of occupations for male and female cast members using the Stanford Dependency Parser and an occupation list compiled from various sources. I expect to find that males are often given higher-level occupations, perpetuating the stereotype of certain professions being more suitable for men. I anticipate observing that occupations like "teacher" or "student" have higher female representation, while professions like "lawyer" and "doctor" are predominantly male-dominated.

5. Singers and Gender Distribution in Soundtracks:
I will examine the gender-wise distribution of singers in movie soundtracks over recent years (2010-2017). I expect to find a consistent gender gap, with female singers being underrepresented compared to their male counterparts. In the future, I plan to incorporate audio-based gender detection to further quantify this trend.

6. Cast Dialogues and Gender Gap in Movie Scripts:
By analyzing 13 movie scripts, I will perform a sentence-level analysis to study the distribution of dialogues between male and female characters. I expect to find some movies exhibiting a balanced dialogue distribution, while others show significant bias, with minimal or no female dialogues, emphasizing gender disparity in movie dialogues.

Through these intra-sentence level analyses, I hope to gain valuable insights into the gender bias and stereotypes present in Bollywood movies. I anticipate uncovering patterns in how male and female characters are portrayed, introduced, and engaged in movie narratives, shedding light on the industry's progress toward achieving gender balance and inclusivity in the future.

2.2.2. Tasks at Inter-Sentence level

At the inter-sentence level, I will conduct various analysis tasks on the Wikipedia movie data by leveraging plot information. To generate context flow, I will use a word graph technique, constructing a word graph for each sentence where each word is treated as a node. The nodes will be connected based on grammatical dependencies extracted using the Stanford Dependency Parser (De Marneffe et al., 2006). Using the word graph for each sentence, I will derive a knowledge graph for each cast member. The root node of the knowledge graph will be [CastGender, CastName], and the relations will represent the dependencies extracted using the dependency parser across all sentences in the movie plot.

After obtaining the knowledge graphs, I will perform the following analysis tasks on the data:

1. Centrality of each cast node:

I will measure the centrality of each cast member as a measure of how much they have been focused on in the plot. For this task, I will calculate the between-ness centrality for each cast node, which represents the number of shortest paths that pass through the node. By finding the between-ness centrality for male and female cast nodes, I will analyze the results. I will show the trend of male and female centrality across different movies over the years, anticipating that there will be a significant gap in centrality between male and female cast members.

2. Study of bias using word embeddings:

For this analysis, I will perform joint modeling of verbs, adjectives, and relations. I will generate word vectors using Google word2vec (Mikolov et al., 2013) of length 200 trained on Bollywood Movie data scraped from Wikipedia. I will use the CBOW model for training Word2vec. The knowledge graph constructed for male and female cast members for each movie will contain nodes connected to them, extracted using the dependency parser. I will assign a context vector to each cast member node, consisting of the average of word vectors of its connected nodes. The aim is to analyze the differences between contexts for male and female cast members.

To study bias using word embeddings, I will randomly divide the data into training and testing data. I will fit the training data using a K-Nearest Neighbor algorithm with varying values of K and study the accuracy results by varying the samples of train and test data.

Through the combined inter-sentence level analysis, the argument for the existence of gender bias will be strengthened, as the data consistently demonstrates the presence of bias in various linguistic aspects related to male and female characters. The results will highlight the need for a more balanced and equitable representation of genders in Bollywood movies.

2.2.3 Movie Poster and Plot Mentions

I will analyze images on Wikipedia movie pages for the presence of males and females on publicity posters for each movie. Using Dense CAP (Johnson et al., 2016), I will extract male and female occurrences from the posters by checking the top 5 responses with positive confidence scores.

Next, I will analyze the male and female mentions in the movie plot and co-relate them with the extracted poster data. The purpose of this analysis is to understand if there is a bias in how movies are publicized through posters, featuring females prominently while having a small or inconsequential role in the actual movie plot.

Surprisingly, despite 80% of movie plots having more male mentions than females, over 50% of movie posters feature actresses. For instance, movies like "GangaaJal 2," "Platform 3," and "Raees" have 100+ male mentions in the plot but 0 female mentions, yet their posters prominently showcase female characters. Moreover, I note that over time, such biases are decreasing, which is an encouraging trend.

In conclusion, while gender bias still persists in movie representations, the increasing number of female-centric movies and the decreasing bias over time indicate a positive direction towards more balanced and inclusive portrayals of women in Bollywood films.

2.3 Movie Preview Analysis

In the movie preview analysis, I will analyze all the frames extracted from the movie preview dataset to gather information about the presence or absence of male and female characters in each frame. If any person is present, I will determine the displayed emotion, which can be one of angry, disgust, fear, happy, neutral, sad, or surprise. It is important to note that a single frame may contain more than one person, and in such instances, emotions will be detected for each individual.

Based on the collected data, I will conduct the following analysis tasks:

a) Screen-On Time:
b) Portrayal through Emotions:
In this task, I will analyze the most commonly exhibited emotions by male and female characters in movie trailers.

By conducting these analyses, I aim to shed light on the portrayal of male and female characters in movie trailers, highlighting any potential biases and stereotypes in terms of screen time and displayed emotions. The results will contribute to a better understanding of how genders are represented in the promotional material for movies and can provide insights for creating more balanced and diverse movie trailers in the future.

3. Conclusion:

In conclusion, this research paper presents a comprehensive analysis of gender stereotypes and biases in Bollywood movies using a dataset containing 4000 movies from Wikipedia. The analysis is conducted at both the sentence and multi-sentence levels, utilizing word embeddings and context vectors to study biases in the data. Several key observations were made during the analysis:

1. Occupations and centrality: The analysis revealed that higher-level roles are often designated to male characters, while female characters are given lower-level roles. Additionally, female characters tend to be less central in the movie plots compared to their male counterparts.

2. Gender prediction and bias: By using context word vectors and very small training data, the research observed a high accuracy in predicting gender, indicating a substantial amount of bias present in the data.

4. References:

1. Hanah Anderson and Matt Daniels. https://pudding.cool/2017/03/film-dialogue/. 2017.
2. Molly Carnes, Patricia G Devine, Linda Baier Manwell, Angela Byars-Winston, Eve Fine, Cecilia E Ford, Patrick Forscher, Carol Isaac, Anna Kaatz, Wairimu Magua, et al. Effect of an intervention to break the gender bias habit for faculty at one institution: a cluster randomized, controlled trial. Academic medicine: journal of the Association of American Medical Colleges, 90(2):221, 2015.
3. Angela M Gooden and Mark A Gooden. Gender representation in notable children's picture books: 1995–1999. Sex roles, 45(1-2):89–101, 2001.
4. Justin Johnson, Andrej Karpathy, and Li Fei-Fei. Densecap: Fully convolutional localization networks for dense captioning. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pages 4565–4574, 2016.
5. Proceedings of Machine Learning Research 81:1–14, 2018