An Analysis of Gender Stereotypes in Movies Over Time

Jared Belsky and **Ryan Bennett**
University of Pittsburgh
[jab540, rmb176]@pitt.edu

Abstract

Large textual datasets such as a corpus of movies are used to train conversational AI models and other AI models. Evaluating these datasets for potential biases is crucial to avoid potential biases in models that utilize them. In this paper, we analyze the CMU Movie Summary Corpus using sentence level analysis of the plot summaries. We find that the tendency for movies to reinforce gender stereotypes has slightly increased over time; however, our methods are too limited for these findings to be reliable, meaning these results are largely inconclusive. Ultimately, we believe more research needs to be done to more robustly identify the biases present in this data set and others.

1 Introduction

Large textual datasets are frequently used to train AI models today; movies in particular serve as an excellent source to emulate conversation. What happens when these datasets contain some sort of bias? The model being trained with said data can very easily reflect these biases. This is dangerous, as biased models can lead to unfair and unethical consequences. For example, a conversational AI trained on biased data could respond with an offensive comment to a child.

This motivated us to take a popular source for model training, movies, and perform analysis for gender stereotypes and bias. In this work, we analyze the movie summaries of over 20,000 movies for gender biases with the aim to discover any potential biases that may have consequences on models trained on datasets like this. In addition to this, we were also particularly interested in analyzing how these biases changed over the course of time. We were inspired to adopt this particular lens on our project after reading through "The Transformation of Gender in English-Language Fiction" by Underwood et al. (2018). This paper found that the gender stereotypes in fiction overall declined over time. After reading this, we became interested to see if this same trend was present in movies. As such, we began our research with the initial hypothesis that gender biases and stereotypes would decrease in movies over time, with older movies supporting these biases and newer movies challenging them.

2 Dataset

For our analysis, we took a look at the CMU Movie Summary Corpus created by Bamman et al. (2013). This corpus contains three separate datasets: the movie metadata, character metadata, and the plot summaries. For our analysis, we needed components from all three datasets. From the movie metadata, we needed the Wikipedia ID, movie name and release date. From the character metadata, we needed each character name and their associated gender, and the Wikipedia ID of the movie the character was in. The plot summaries data consisted of only the Wikipedia ID and it's summary. Due to the Wikipedia ID being a constant between all the datasets, we used it as a merge point to get all the data together.

To read in all the data, we used pandas[1]. The dataset comes in 3 different pieces for character metadata, movie metadata, and plot summaries, so we first needed to get the appropriate data organized. After some data cleaning, we were able to extract all information necessary for our analysis and combine them into one pandas dataframe that consisted of: Wikipedia ID, Movie Name, Movie Release Date, Plot Summary, and Character List (including genders). Without any modifications, the CMU Movie Summary Corpus contains information about 81741 movies in total. However, we were not able to use all of these movies, as many of them were missing either a plot summary or character metadata. After removing all of the

[1] https://pandas.pydata.org/

movies missing the information we needed, the final dataframe had 22510 movies. With all of these components collected in one dataframe, we could now run our methods to calculate bias score.

3 Methods

Since our main intention is to analyze how gender biases in movies change over time, our methods are fairly straightforward; for each movie, we simply want to run some analysis which allows us to assign it a bias score between 0 and 1. In order to do this, we took pointers from the paper by Fast et al. (2016) which inspired our research. In their research of gender bias in online amateur fiction, they calculate bias by extracting subject/verb and subject/adjective pairs and looking for stereotyped words.

For our research, we used the NLP tools provided by the spacy python library[2]. To calculate bias scores, we ran the same method on every movie in the dataframe. First, we ran through the plot summary sentence by sentence and extracted all of the subject/verb and subject/adjective pairs. For every pair we found, we identified the gender of the subject as well as the gender stereotype of the adjective or verb.

3.1 Subject Gender

To identify the gender of the subject, we used the list of characters and their associated genders as well as pronouns. For example, if the subject was "She", it would be marked female, and if the subject was "He", it would be marked male. Our analysis was limited to only male and female, and did not include any non-binary genders. If the subject was a name that was missing from the character list or anything else that didn't have an easily identifiable gender, we simply marked it as "N" for neutral.

3.2 Adjective/Verb Gender

Identifying the gender stereotype of the adjective or verb in question is a bit more difficult. For our purposes, we started with a set of six stereotype words for each gender:

Gender	Stereotype Words
Male	Attack, Anger, Swear, Sexual, Independent, Achieve
Female	Submissive, Home, Money, Family, Social, Friends

Table 1: Starting set of stereotype words for each gender

However, looking for only these words would mean we would find an unusually low frequency of bias. To work around this, we used the word vector tools provided by spacy to identify similarity between words and expand our scope of stereotype words. For each adjective or verb we came across, if it had a similarity score over a certain threshold with any of the words in the male or female word list, it would be marked with the appropriate gender. If the word was not similar enough to any word in either list, it would be marked as "N" for neutral. In our research, we used a threshold of 0.4.

After marking the gender of the subject and the gender of the verb/adjective in each pair, we added one to a bias counter if the genders matched and added one to a total counter for every pair (If both words had a neutral gender, we did not increment the bias counter). After running through every pair in the plot summary, we divided the bias count by the total count to give the movie a final bias score.

4 Analysis

After obtaining a bias score for each movie, we were able to generate the graphs seen in Figure 1 and Figure 2 by averaging the bias score for every movie in each year. This gives a single year it's own bias score which we then plot as seen in a line graph in Figure 1 and a Scatter Plot with a line of best fit in Figure 2.

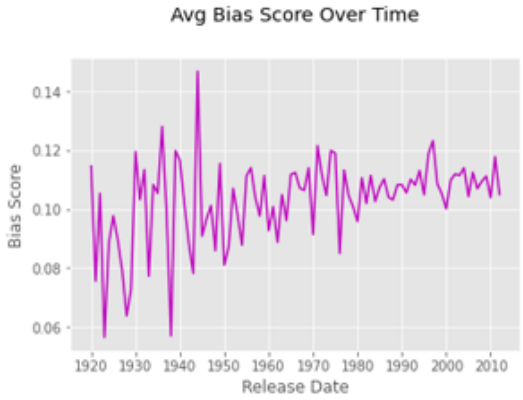

Figure 1: Line Graph

Although results look varying, the y-axis for bias score only has a range between 0.06 and 0.15. Some years had a very low sample size, resulting in bias scores lying far outside of this range, making the results quite messy. To fix this, we began our analysis from 1920, since the bias scores before

[2]https://spacy.io/

then were highly variable, indicating a lack of sufficient movies. In our analysis, all of the movies scored very low in terms of bias.

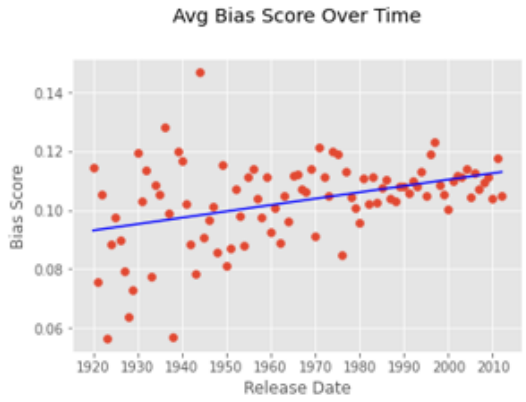

Figure 2: Scatter Plot

Figure 2 gives a better idea of how these bias scores change over time relative to each other, The blue line represents a line of best fit for the graph. This line of best fit has a slight positive slope, signifying an increase in gender bias in movies over time, challenging our initial hypothesis that gender bias would decrease over time. Of course, while these results per our methods are interesting, there are reasons to question the legitimacy of these results, as discussed in the next section. However, based purely on the methods we discussed, we were more than happy to see some sort of change over time, and see this as a jumping-off point for further, more in-depth research.

5 Limitations

There are a number of limitations with our methodology that weaken the reliability of our results, as discussed in this section.

5.1 Summaries

For one, movie summaries inherently obscure the finer details about the movie itself. Bias often shows itself in very subtle ways, and a lot of circumstances where bias will show itself won't necessarily make it into the movie summary. There are lots of details about how a character dresses, how a character interacts with others, how a character is treated by others, minor mannerisms and more that can all be potential carriers of bias and stereotypes.

5.2 Sentence Level Analysis

By using sentence level analysis, we run into a similar problem as working with summaries, but in the other direction. Instead of missing the finer details, all we have is two details: a subject and a verb or a subject and an adjective. Regardless, we still lose the information that we need. If we can't see outside the scope of a pair of words, we can't identify how a character has changed throughout the story, what the person they're interacting with is like, or anything else; all we can know is whether or not the character is described in a stereotypical way or performing a stereotypical action.

While these two problems have very similar effects on the results, we believe it is much more straightforward to fix the summaries problem in future research. By using scripts instead of summaries, there is plenty of information in character dialogue and stage directions that would reveal biases not present in a summary. On the other hand, training a model to understand the nuances of plot and characters as they progress throughout a story is much more difficult.

5.3 Word Similarity

Using word vectors to detect similarity with stereotyped words greatly reduced our engineering effort and allowed us to finish this research, but it also opened the door for potential mistakes in the similarity score. If a word has a similarity score with a stereotype above the selected threshold, it is not necessarily synonymous or interchangeable with that word. For example, the words "attack" and "defend" have a very high similarity score (0.60), high enough that our model would consider "defend" a stereotypical action for a male subject. However, these two words don't have the same meaning, and they certainly don't carry the same stereotype. To overcome this problem in future research, we recommend creating a much larger list of stereotype words and using a very high threshold for similarity score if it is necessary to further expand that list. Even still, it would be better to avoid using similarity scores, as they make the program much more computationally expensive, making it more difficult to scale up to longer text datasets such as movie scripts.

5.4 Minor Limitations

In addition to these three major limitations, there are also a number of smaller (but still significant) problems that we ran into during our research

5.4.1 Character Names

In plot summaries, characters are often referred to using their last name. While this is usually not much of an issue (character last name is typically included in the character's metadata), it can become an problem when two character's have the same last name. In these situations, it is very difficult to tell which character it is actually referring to, which makes it hard to assign a gender to them. We were not able to implement a proper fix for this.

5.4.2 Direct Objects

While we included subject/verb pairs in our analysis, we did not include the direct object. This means we are missing more situations where a character is potentially having actions done to them that reinforce stereotypes instead of doing actions themselves. This is particularly important because female characters are often denied a certain level of agency, meaning we won't find instances of bias if the character is never allowed to do anything for themselves in the first place.

6 Conclusion

Ultimately, our research found that gender bias in movies has slightly increased in the past 100 years. However, our methodology is very rudimentary and has a lot of room for error, so we do not recommend taking these findings with much weight.

Nevertheless, we believe this problem is important to study. As pointed out by Paullada et al. (2021), datasets are the foundation of machine learning, and if we don't make sure we have good datasets, we can end up with flawed models. When we use film datasets to train conversational AI models, we want to make sure that potential biases don't sneak their way into the finished product. For example, if women are often spoken to in dismissive or demeaning ways in movie scripts, we don't want a conversational model to learn these patterns. While our approach was not able to identify bias as effectively as we would have hoped, we strongly encourage further research to be done that can fill in the gaps in our evaluation and provide a better understanding of this dataset and others.

References

David Bamman, Brendan O'Connor, and Noah A. Smith. 2013. Learning latent personas of film characters. *ACL 2013*.

Ethan Fast, Tina Vachovsky, and Michael S. Bernstein. 2016. Shirtless and dangerous: Quantifying linguistic signals of gender bias in an online fiction writing community. *Stanford University*.

Amandalynne Paullada, Inioluwa Deborah Raji, Emily M. Bender, Emily Denton, and Alex Hanna. 2021. Data and its (dis)contents: A survey of dataset development and use in machine learning research. *Patterns*, 2.

Ted Underwood, David Bamman, and Sabrina Lee. 2018. The transformation of gender in english-language fiction. *Cultural Analytics*.