Bridging the Gap: An NLP-based Tool for Studying and Visualizing Bias in Film

Christopher Ang, Alpha Diallo, Joshua Andrews, Dan Levere

1 Introduction

Research on racial and gender bias in literature is impeded by limitations in acquiring traditional human ratings. Our project provides an analytic and visual aid to overcome these limitations and aid interdisciplinary research. Drawing from a collection of previously validated text analytics, and using large databases of film data, we will create an application that displays various metrics of racial and gender bias as they appear within the spoken lines of narrative characters within films. Recently, big data and NLP approaches have shown promise in aiding research on this topic, specifically by scraping films and book narratives from online sources.

2 Survey of Literature

Work done in this topic area has ranged from manual coding and data aggregation, to text analysis and visual analysis.

Manual Analysis

We first examine some of the manual intensive research done in this area. [6] provides an example of manual coding and inter-rater reliability calculations on gender role portrayal within 9 Disney princess films. While [5] is a two part study examining physical attractiveness stereotypes across 40 Disney films by having humans rate the attractiveness and friendliness/goodness of characters. In [7], three trained raters reviewed and rated racial and gender stereotypes for characters portrayed by top-billed actors across 44 films. The study provides a rich literature review and an example of traditional rating techniques. Finally, [3] examines 14 sentiment analysis methods for the English language in literature, including Stanford Recursive Deep Model, Opinion Lexicon, Opinion Finder, and Happiness Index. The purpose of the study was to show that current manual methods can be used with machine translation systems to provide similar, if not better, results.

Overall, while it is possible with manual approaches to take a deep dive into particular films, it is challenging with such approaches to discern longer term trends. We will attempt using modern analytics approaches to glean insight into the longer term trends.

Automated Analysis

Automated approaches have focused on a variety of aspects of the films. Some focus on characters, such as [4], in which the authors introduce the concept of latent character profiles (personas) for narrative characters. [13] examined the psycho-linguistic properties of characters across 1547 movies by parsing scripts and using Linguistic Inquiry and Word Counts (LIWC). Social network extraction from film text was the focus of [1]. While [10] used text analysis in conjunction with character names to develop a network graph on film characters and a machine-learning classifier to determine the female level of representation in film. [2] also focuses on gender bias via automation of the Bechdel test-a popular test for measuring gender stereotyping-via automating the test by way of parsing film scripts. We aim to use similar techniques to construct character networks, but having the ability to discern more than just gender bias as a result.

Other techniques focus on discerning information from the narratives. [11] showcases methods for data collection and preprocessing using 978 move scripts. The results suggest NLP is able to identify patterns in the narrative structure of films. While the chapter in [12] provides an in-depth review of corpus methodology (analyzing large corpora of text). In [8] the authors use various google word2vec models to analyze gender and ethnic biases in various sources of literature (e.g. news). [16] outlines the use of text-to-animation frameworks (eg. ScriptViz), which are used to create animated scenes from script text. [14] analyzed 772 movie scripts to test and validate the power/agency connotation framework for text analysis. Findings suggest imbalances in portrayal of power/agency between male and female characters. [15] conducted a sentence by sentence analysis of movie scripts in order to visually chart the level of happiness male and female characters experience throughout the narrative.

These studies showcase what kinds of information can be gleaned purely from the text, and we will leverage similar techniques in our approach. Our advances will be in the automation, combination of these different techniques and metrics, and in the visualization.

Finally, in [9] the authors used computer vision (CV) techniques to analyze the degree of bias present in cinematic visual representations of women across 40 films. We will use DeepFace in this project to analyze the race/gender of character actors.

3 Proposed Method

3.1 Intuition

Gender and racial biases in film have been historically observed by many. However, having data-driven visualizations based on analyses to quantify the extent of these biases does not exist in the mainstream. Having the ability to visualize not only that a bias exists, but also the magnitude of the bias will paint a much clearer picture for those interested in the subject.

3.2 Data

We first scraped movie scripts from IMSDB. Movie script data is generally unstructured text. This presented a challenge, as some of the analyses we wished to do required having the actual dialogue lines. We thus wrote a program that would:

(1) Find characters in the script
(2) Extract the lines for said characters
(3) Find scenes within the script
(4) Associate scenes to the lines of dialogue and the characters

Due to challenges with processing unstructured text, along with some differences in the way IMSDB handles their pages for scripts, not all movies on IMSDB were able to be processed. We still ended up with over 1000 scripts, and so we have a sizeable sample for our analysis.

Using script titles, we searched the TMDB database for associated information about the film. Since IMSDB does not provide TMDB IDs, we were limited to this approach. By using the search API we were able to collect film metadata such as budget, revenue, genre, and release date. We then used the film IDs from the TMDB search to pull character information (e.g. character name, character order) as well as the celebrity information (e.g. celeb name, gender, picture URL).

For the movie metadata scraped from TMDB, gender for many actors and actresses, as well as race information was not present. A modified version of the DeepFace prediction package was used to predict the race and gender of each character based on pictures, whose URLs were scraped from TMDB. DeepFace is a face recognition framework that incorporates the use of VGGFace and weights to predict race and gender.

Finally, in support of some of the analysis, we needed to match up the characters from TMDB with the characters we parsed from the IMSDB scripts. In order to achieve this, FuzzyMatching techniques were leveraged.

The steps performed were as follows:

(1) Clean the names of non-ASCII characters, and sort the tokens alphabetically.
(2) Find any names in this movie that match exactly between the data sets, and remove them.
(3) Attempt to match the top 15 characters according to a modified longest common subsequence algorithm. (PyLCS was leveraged here.)

(4) Using the same, attempt to match the "most talkative" characters from the script.
(5) Attempt to run though any remaining characters and match, using a minimum threshold of 40%.

Additionally, there were cases where the character names from TMDB used full names, while the names from the scripts were nicknames, such as Jo for Joanne. To help, in the last two steps, a map of names to alternatives was used to create possible name representations of the names in TMDB that one might find in the film scripts.

Overall, while the results of this aren't perfect, we believe they're adequate enough to support our analysis.

3.3 Description of Approaches

By parsing the film scripts we were able to extract dialogue and scene information to provide the below analyses. These features will be visualized using Tableau views embedded in a web page.

Character Network Analysis

Any two or more characters that share dialogue within a scene are considered to interact. Scenes are determined based on film script stop words such as FADE TO, EXTERIOR, INTERIOR, etc. Any dialogue between characters within these terms are determined to be sharing a scene and thus interacting with each other. Thus, a directed graph can be employed to represent character connectivity, and overall importance of any one character.

Scene Frequency

Using the parsed scene data, we identify the total number of character lines and scene appearances through the movie. These metrics provide a high-level view of a character's narrative importance.

Word Frequency Analysis

For the listed film characters, each line was parsed (tokenized) into individual words. A list of stopwords from the `nltk` corpus were removed, and the remaining words were counted for frequency to generate the top *n* words spoken by each film character.

Sentiment Analysis

Each film line was analyzed using `nltk VADER` polarity scores. `VADER` is a pre-trained sentiment analysis model, which uses a defined lexicon and rule based algorithm to define sentiment intensity on a positive and negative spectrum. We used the data from this analysis to chart character sentiment across the narrative.

Power Agency Analysis

Each film line was parsed by replacing punctuation (excluding contractions) and tokenizing by whitespace. Upon initial analysis, the lexicon for power/agency [13] appeared best suited for third person narrative and not spoken lines. This is essentially an issue caused by different verb conjugations (e.g. look vs. looks). Therefore, we stemmed the verbs in the lexicon in addition to the sentence tokens using `nltk PorterStemmer`.

Bechdel Test

The Bechdel test is a heuristic device that examines whether women have a significant and active presence within a narrative. The test has three conditions:

(1) The film includes a scene with at least 2 conversing female characters
(2) These female characters speak to each other.
(3) The content of the dialogue must be about something other than a man.

For the first condition, we identified the characters in each scene and connected them to the gender of the actor/actress. For the second condition, we tokenized each sentence using `spacy`. We then searched for each token in the `nltk` corpus names dictionary, which contains names with the most commonly associated gender. Using this data, we can identify scenes that pass the test.

3.4 User Interface

We deployed the visual elements of our analysis onto the internet using Github pages. You can find that at movievis.com. We leveraged Angular for the UI code, along with Tableau for the Visualization elements. Our UI is split up into three sections.

The General Analysis page highlights summaries across our entire data set. Users can filter by race and see statistics for films with actors/actresses of that race. Additionally, users can see charts highlighting the number of movies during different periods, along with gender breakdowns based on their filter.

The Film Analysis page is a deep dive into various elements for a chosen movie. Users can choose whether to see the visuals according to race or gender. If race is chosen, the colors will distinguish between white and non-white races. Users will be able to see character network graphs, power/agency charts highlighting each character's power/agency, word charts showing relative frequency of words highlighted by race/gender, scene appearance in the movie, and relative action in the film.

The Character Analysis page enables a deeper dive into particular characters for a movie. The primary driver here is the ability to view the total power/agency and sentiment analysis for the lines for the chosen character.

4 Design of Experiments

4.1 Description of Testbed

We hope to answer the following:

(1) Whether or not a scene, and thus movie, passes the Bechdel test.
(2) What the most frequently used words for each character are.
(3) How many lines are associated to characters by race and/or gender in specific films
(4) The general sentiment and power/agency for individual characters, and by race or gender
(5) Character importance, and how those characters are connected to others, again by race and gender.

4.2 Experiments and Observations

Bechdel Test

Automatic detection of the Bechdel test proved to be challenging. For one, determining what constitutes two characters speaking directly to one-another is a bit fuzzy. We decided that placement in the same scene was a good substitute, and that's how we determined whether a scene passed the first criterion. If at least one of those female characters spoke, and her line did not contain a male name, we considered criteria two and three satisfied. We encountered issues in implementing this test (see limitations), and instead we present manual verification of the Bechdel Test with data scraped from bechdeltest.com. Based on on this data, we found that 63 percent of our movies failed the Bechdel test. These findings are fairly consistent with reports from previous research, which suggests that slightly over 50 percent of movies fail the test.

Line by race/gender

An examination of spoken lines is another approach for visualizing the relative importance of characters in relation to their gender/race. Results indicate that women were less represented in films. This trend persists when filtered to racial groups. Our findings are consistent with past research suggesting women are underrepresented in film narratives. Coupled with results from the Bechdel test, we see that when women are represented their roles tend to be less significant to the narrative.

Sentiment Analysis

For our sentiment analysis, we used the pretrained `VADER` model available in `nltk`. One reason we decided to use the `VADER` model was because it was trained on social media data, which has similarities to film dialog. Notably, the text may be short but convey emotion through punctuation (e.g. "Oh, my!"). We believe VADER performed well on film data and was able to capture character emotion on a positive/negative scale fairly accurately. For example, the analysis for *Die Hard* shows that John McClane displays a persistent pattern of negative emotion throughout the film, consistent with the narrative.

Power/Agency Analysis

Our initial implementation of the lexicon yield sparse results, mainly arising from verb conjugation. We decided to use a stemming approach to overcome this issue, but this was an imperfect solution (see limitations). Based on our findings, we might suggest an update to the power/agency dictionary to include verb conjugations. To create a total power score, we summed the number of spoken positive power verbs and subtracted the number of negative power verbs for each character. Results from the power/agency analysis suggest variance in the word choice between male and female characters. For example, in the movie *Jurassic World: Fallen Kingdom*, the character Claire (female) has the third lowest power score, despite being a central plot character. Owen (male), the co-protagonist to Claire, speaks 6 power verbs to Claire's 1. In the movie *Indiana Jones and the Temple of Doom*, Indiana Jones (male) has a power score of 123 compared to his co-protagonist Willie's (female) score of 76. Results appear consistent with findings from previous research; even when in central roles, female characters tend to display less power than male characters.

Word Frequency (Word Clouds)

We used a word cloud visualization to display "match-words" from the power/agency lexicon broken down by gender and or race. Results suggest some noticeable distinctions in power/agency verbiage between males and females. For example, in the movie *10 Things I Hate About You*, female top words include "loves", "kisses", "happens" and "dates", while male top words include "thinks", "lets", "tells", "gets", and "says". Overall, results do suggest that our analysis captured trends that appear consistent with previous research assertions. Trends in the data show underlying stereotypical associations; women "kiss" and "love" while men "think" and "get".

Network Analysis

Network graphs were a great way for us to display the inter-connectivity of narrative characters. The graph communicates character importance through both scene frequency (node size) and co-appearances (node links). Our results suggest that network graphs are particularly useful for examining character importance through the lens of gender and racial background. For example, the movie *Gladiator* shows a network analysis of roughly equal node counts, sizes, and links between characters portrayed by white actors and those portrayed by actors of color. In the movie *Nightmare on Elm Street* we see that most characters are male, but the main character, Nancy (female), has the highest scene frequency (largest node) and connectivity (most links). These results suggest an accurate assessment of narrative importance using network analysis.

Overall, our analyses shows consistent trends that align with previous research on the topic. We were generally successful at identifying racial and gender discrepancies across and within film narratives.

4.3 Discussion of Limitations

There exist many imperfections about our approaches, in large part to the challenges present in processing unstructured text. The following sections outline the most severe limitations in the provided analyses.

Gender and Race Predictions

During data collection, we were unable to predict the gender and race of roughly 6,700 characters due to missing picture URLs and/or pictures that were too low quality to from confident predictions. Our attempt to match celebrity metadata to character dialogues also had a degree of error due to FuzzyMatching being an imperfect approach.

Bechdel Test

While doing this analysis, we discovered a variety of issues that prevented us from properly detecting scenes that pass the Bechdel Test. As mentioned, there were issues surrounding the identification of gender for some characters. There is also a strong likelihood we missed scene transitions which would help isolate female character dialogues. When running our analysis, we found that less than 10% of films we scraped actually passed the Bechdel test. Given that manual verifications for our movies push this to 37%, we clearly had issues in this analysis. However, due to member attrition (see appendix), we were not able to do everything we wanted here.

Power/Agency Analysis

Our approach to analyzing power/agency was primarily limited by the fact that the lexicon we used was trained on third person narration and included mainly third person verb conjugations. To circumvent this issue, we stemmed the lexicon and sentences tokens for analysis. This creates new issues, such as verbs like mans becoming man, which then identified instances of the word "man" as a noun. Our suggestion for future research is to create a new lexicon that is trained on first person dialog.

Network Analysis

The limitations in network graphs for this project stem from anomalies in parsing the film scripts. In some cases, a character's name may have been mistranscribed. Scenes were automatically detected based on the script, and characters who are actually in the same scene may be separated due to a false detection of a scene transition. Finally, we were only able to detect characters with lines of dialogue as being present in a scene, so silent characters are not be represented.

5 Conclusion and Discussion

In its entirety, we believe this project was a successful application of text and NLP analysis to the task of analyzing film scripts. There exists areas for improvement, but results suggest promise for applying the listed techniques. In particular, we were largely successful in parsing a large volume of film script information and in obtaining information about actors/actresses for analysis. Collecting and cleaning data constitutes a large effort for any research project and our open-source scripts can be useful to future researchers. Furthermore, while there is room for improving upon the presented analyses, our project offers a valuable collection of analysis scripts that can act as a starting point for future endeavors. Finally, our visualization website makes exploration of results accessible to audiences that lack the technical knowledge to conduct these analyses themselves, offering a potentially valuable pedagogical and research tool to researchers outside of computer science.

6 Distribution of Effort

There were 5 development areas for this project, each with a team lead. Members worked across groups, but lead their respective task area. Resulting from team member attrition, each remaining member was required to perform more cross-area development, in particular, working towards the visualizations as a result of that team lead dropping the course.

Data Collection (Alpha): collecting film scripts from IMSDB, pulling actress/actor information from TMDB, and predicting character's ethnicity based using the `DeepFace` Framework.

Data Integration (Alpha & Christopher): Determine how to best combine the data for data analysis and model building.

Data Analysis and Models (Joshua): Here we study our data and build models to determine representation of gender and race based on the metrics discussed in the in lit. survey.

Visualization (Everyone): We will be leveraging tableau to create a variety of interactive visualizations to illustrate our analysis.

Deployment (Christopher): We will be leveraging a combination of Github pages and Tableau Public for deployment.

6.1 Acknowledgements

Alpha Diallo: Alpha was an emergent leader throughout the project and provided our team with the topic of studying gender and racial bias in films. His leadership was pivotal for our maintained focus as circumstances shifted throughout the semester.

Chris Ang: Chris provided leadership and technical skills for realizing our project goals. His effort and willingness to aid other team members was a cornerstone to our project's success.

References

[1] Apoorv Agarwal, Sriramkumar Balasubramanian, Jiehan Zheng, and Sarthak Dash. 2014. Parsing screenplays for extracting social networks from movies. In *Proceedings of the 3rd Workshop on Computational Linguistics for Literature (CLFL)*. 50–58.

[2] Apoorv Agarwal, Jiehan Zheng, Shruti Kamath, Sriramkumar Balasubramanian, and Shirin Ann Dey. 2015. Key female characters in film have more to talk about besides men: Automating the bechdel test. In *Proceedings of the 2015 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*. 830–840.

[3] Matheus Araújo, Adriano Pereira, and Fabrício Benevenuto. 2020. A comparative study of machine translation for multilingual sentence-level sentiment analysis. *Information Sciences* 512 (2020), 1078–1102.

[4] David Bamman, Brendan O'Connor, and Noah A Smith. 2013. Learning latent personas of film characters. In *Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*. 352–361.

[5] Doris Bazzini, Lisa Curtin, Serena Joslin, Shilpa Regan, and Denise Martz. 2010. Do animated Disney characters portray and promote the beauty–goodness stereotype? *Journal of Applied Social Psychology* 40, 10 (2010), 2687–2709.

[6] Dawn Elizabeth England, Lara Descartes, and Melissa A Collier-Meek. 2011. Gender role portrayal and the Disney princesses. *Sex roles* 64, 7-8 (2011), 555–567.

[7] Sarah Eschholz, Jana Bufkin, and Jenny Long. 2002. Symbolic reality bites: Women and racial/ethnic minorities in modern film. *Sociological Spectrum* 22, 3 (2002), 299–334.

[8] Nikhil Garg, Londa Schiebinger, Dan Jurafsky, and James Zou. 2018. Word embeddings quantify 100 years of gender and ethnic stereotypes. *Proceedings of the National Academy of Sciences* 115, 16 (2018), E3635–E3644.

[9] Ji Yoon Jang, Sangyoon Lee, and Byungjoo Lee. 2019. Quantification of Gender Representation Bias in Commercial Films based on Image Analysis. *Proceedings of the ACM on Human-Computer Interaction* 3, CSCW (2019), 1–29.

[10] Dima Kagan, Thomas Chesney, and Michael Fire. 2019. Using Data Science to Understand the Film Industry's Gender Gap. *arXiv preprint arXiv:1903.06469* (2019).

[11] Seong-Ho Lee, Hye-Yeon Yu, and Yun-Gyung Cheong. 2017. Analyzing Movie Scripts as Unstructured Text. In *2017 IEEE Third International Conference on Big Data Computing Service and Applications (BigDataService)*. IEEE, 249–254.

[12] Effie Mouka, Ioannis E Saridakis, and Angeliki Fotopoulou. 2015. Racism goes to the movies: A corpus-driven study of cross-linguistic racist discourse annotation and translation analysis. *New directions in corpus-based translation studies* 1 (2015), 35.

[13] Maarten Sap, Marcella Cindy Prasettio, Ari Holtzman, Hannah Rashkin, and Yejin Choi. 2017. Connotation frames of power and agency in modern films. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*. 2329–2334.

[14] William R Shadish, Thomas D Cook, Donald Thomas Campbell, et al. 2002. *Experimental and quasi-experimental designs for generalized causal inference/William R. Shedish, Thomas D. Cook, Donald T. Campbell.* Boston: Houghton Mifflin,.

[15] Huimin Xu, Zhang Zhang, Lingfei Wu, and Cheng-Jun Wang. 2019. The Cinderella Complex: Word embeddings reveal gender stereotypes in movies and books. *PloS one* 14, 11 (2019).

[16] Yeyao Zhang, Eleftheria Tsipidi, Sasha Schriber, Mubbasir Kapadia, Markus Gross, and Ashutosh Modi. 2019. Generating animations from screenplays. *arXiv preprint arXiv:1904.05440* (2019).

A Team Member Attrition

The plan of activities was revised as a result of a group member's departure from the team (Matt). As he was the member with the most expertise with Tableau, and this was the visualization framework we decided to use, we had lot of challenges in creating the visualizations and displaying them in a nice readable format. Overall, we think we did a reasonable job given this loss, but it definitely impacted us and limited our ability to do more analysis, as we spent more time on the visualizations as a result.