Automating the Bechdel Test

Serina Chang
Computer Science Department
Columbia University
sc3003@columbia.edu

Navraj Narula
Computer Science Department
Columbia University
nnn2112@columbia.edu

Abstract

Conversations surrounding gender inequality in Hollywood films saturate online forums and media outlets. The question as to whether or not women are given enough screen time as well as how complex their characters are in a movie can be answered by a series of three questions known as the Bechdel test. In this paper, we present methods and results to automatically predict whether a given movie passes each Bechdel requirement. Our findings improve previous work on all three classification tasks, and we additionally contribute new fine-grained analyses on the relationship between Bechdel scores, power, agency, and gender.

1 Introduction

The Bechdel test was introduced by cartoonist Alison Bechdel in a comic strip, "Dykes to Watch Out For." A movie passes the test if it meets three requirements: (T1) two named women appear in the movie, (T2) who speak to one another, (T3) about something besides a man (Bechdel, 1986).

The simplistic nature of the Bechdel test hints at the lack of female representation in film. According to a report released by the University of Southern California, viewers will see 2.18 male characters for every 1 female character on screen. This statistic follows from the analysis of 4,583 speaking characters across the U.S. Box Offices top 100 movies in 2016. Moreover, the prevalence of female speech on screen has only increased by 1.5% between 2007 and 2016 (Smith et al., 2017). As evidenced by these statistics, the degree of gender bias in film becomes especially apparent when studied at a macro-level. This motivates our objective to automate the Bechdel test, so that female representation in film can be evaluated efficiently at scale.

2 Related Work

Quantifying gender bias in fiction has been a long-standing topic of interest in popular culture, social science, and recently, natural language processing communities. Picture books and movie scripts have drawn particular interest, because they bring together visuals, dialogue, and storylines to shape the ways that young people understand gender roles as they grow up. Weitzman et al. (1972) analyzed hundreds of prize-winning picture books by hand, and found that women were underrepresented in central roles and often portrayed stereotypically. Smith et al. (2013) brought this work into the realm of movies by analyzing 500 top-grossing films and 1,200 characters. Again, females were grossly underrepresented and more likely than their male counterparts to be hypersexualized. Like Weitzman et al., their methodology was all manual annotation by trained coders, which produced high quality insights but required intensive time and often expenses.

Our approach aims to uncover insights like these about gender representation in fiction but to do in a computational manner. Computational techniques save time and money over manual effort, and allow evaluations to scale up to much larger corpora. Already there is some progress at the intersection of natural language processing and social science to automatically quantify gender bias in fiction, particularly film.

Ramakrishna et al. (2015) expanded a lexicon of "gender ladenness" (perceived association with femininity or masculinity) and quantitatively studied gender ladenness in movie dialogue, with respect to variables such as the genders of the characters and the genders of the screenplay writers. Schofield and Mehr (2016) also studied language and gender in movies, examining differences in conversations between different gender pairs (female-female, male-male, and female-male). Sap et al. (2017) also contributed important work in this realm, collecting verbs from movie dialogues to analyze connotation frames of agency and power. Utilizing crowdsourcing platform Amazon Mechanical Turk, they built one corpus of 2,000 verbs labeled for their agency frames from low to high agency, and a second corpus of 1,700 verbs labeled for their power frames indicating the power dynamic between the agent and theme of the verb.

Compared to previous work, the Bechdel test uniquely combines the study of numerical representation of women in film and analyses of their conversations. Despite the popularity and usefulness of the test, little work has been done to automate it. Agarwal et al. (2015) have done the most extensive work to date, developing classifiers for each of the three requirements of the Bechdel test and designing features for the classification tasks. Our work is an extension of their research, as we recreated their approaches and improved them with variations and additional features. Our changes boost performance on all three classification tasks, most notably by 11 points on T2. We also utilized Sap et al.'s connotation frames of agency and power as novel features for Bechdel prediction, and furthermore conducted fine-grained analyses of how agency and power in female-female, male-male, and female-male exchanges relate to Bechdel scores.[1]

3 Dataset

We acquired movie screenplays from Agarwal et al. and Gorinski and Lapata (2015). After aligning these movies to Bechdel ratings from bechdeltest.com, we were able to build a corpus of 727 Bechdel-rated screenplays, where 53 failed all three Bechdel requirements, 222 passed only T1, 103 passed only T1 and T2, and 349 passed all three requirements and thus passed the Bechdel test. Previously, Agarwal et al. had 457 Bechdel-rated screenplays in their corpus, so we believed our augmented dataset would contribute to better learning capability and generalization in our classification experiments.

We performed a 70-30 split on our dataset to create a development set and test set. We constructed individual classifiers for each Bechdel subtest and tuned them on the development set (Table 1). For each classifier, we evaluated its ability to predict whether a movie passed that subtest by performing 5-fold cross-validation over all movies that passed the previous subtest. For example, we evaluated the T2 classifier on all of the movies in the development set that at least passed T1.

	Fail	**Pass**
T1	41	467
T2	152	315
T3	68	247

Table 1: Distribution of movies in the development set.

4 Preprocessing

In order to work meaningfully with the screenplays, we had to execute a number of preprocessing steps.

(1) We normalized character names, such that we could map all variants of a name in the screenplay to the same root. This was necessary because one character could appear in many ways in a single script. For example, if a character AMY had a voice-over, the screenplay might read AMY'S VOICE or AMY (V.O.). Downstream tasks included steps such as determining how many characters there were of each gender, aggregating dialogue for each character, and constructing social networks, so it was important to normalize variants so that we did not count each variant as a separate character.

Once we identified a line as a character name, we normalized it by first stripping leading and trailing punctuation from each token in the line. Then, we removed tokens in parentheses and the token 'THE' if it appeared. If the token 'VOICE' appeared following a possessive, we kept the portion that preceded the possessive (for instance, AMY'S VOICE

[1] "We acknowledge that gender lies on a spectrum, and reducing it to a male-female binary is simplistic, however our data limits a more complex understanding of gender" (Sap et al., 2017).

would become AMY). Finally, we lowercased each token.

(2) We aligned screenplays with IMDb ids in order to match the screenplays to Bechdel ratings, since the bechdeltest.com database identifies ratings uniquely by IMDb id. Finding IMDb ids was also important so that we could identify duplicates between the Agarwal and Gorinski datasets. In fact, once we aligned ids, we found that the Agarwal and Gorinski sets overlapped by 245 movies.

We aligned screenplays to ids by first parsing the title from the screenplay and querying the title in Python's IMDbPy library. We then chose the search result that had the highest overlap in character names with the screenplay (Ramakrisha et al., 2015). As a second check, we calculated the Levenshtein distance between our screenplay title and the title of the movie we chose. If the distance exceeded 10, then we would further query the movie for its production year as well as the writer. If these contents were present in our script, we concluded that the id did match the movie.

(3) A final preprocessing step we took was scene and dialogue extraction. This was important for later tasks such as identifying whether two women were speaking to one another (T2). Screenplays provided by Agarwal et al. were already parsed line-by-line, with markers denoting character names, dialogues, scene descriptions, scene changes, and metadata. The "S" marker for scene changes made the process of scene extraction simple for the Agarwal dataset. Gorinski screenplays were not parsed, but there were still indicators we were able to rely on for scene extraction given the specific structure of screenplays. We divided content from screenplays into scenes for this dataset by relying on these indicators: ":SC:,", which denotes a specific scene change, "INT.," which denotes a scene taking place indoors, and "EXT.," which denotes a scene taking place outdoors.

5 Test 1: Are there at least two named women?

5.1 Name-to-Gender

We developed a name-to-gender algorithm so that we could determine whether there were at least two named women in the movie. First, we retrieved name lists from the Social Security Administration (SSA) of the United States. The name lists span years 1888-2016, and each list provides the names of babies born in that year and how many babies of each gender were born with that name.

Given the character names that we normalized during preprocessing, our name-to-gender algorithm first checked whether there were gendered prefixes in the name, such as Miss or Sir. If no prefixes were found, the algorithm checked one token at a time to see whether it appeared in the name lists from the decade leading up to (and including) the year of the movie release. We limited our search to the preceding decade because name usage can change significantly over time. If that token appeared in the decade, a "gender score" was returned, which was the number of times that name was given to a female baby divided by the overall number of times that name was given to a baby of either gender. Thus, a gender score closer to 1 indicated that the name was likelier to be assigned to a girl, and a gender score closer to 0 indicated that it was likelier to be assigned to a boy.

Through this scoring method, we were able to provide degrees of confidence along with our name-to-gender assignments. We utilized these degrees of confidence to build two versions of a rule-based classifier for T1. In the "hard" version (RB-H), the classifier would only predict Pass if there were at least two characters with scores of over .9. In the "soft" version (RB-S), we softened the criteria such that the classifier would predict Pass as long as at least two characters had scores of over .5.

We also used our name-to-gender scores to build a machine learning classifier for Test 1. We provided two features to this classifier: first, the number of female characters who spoke more than once, and second, the number of female characters who spoke exactly once. Our intuition was that providing a count of female characters would be more informative than a binary indicator of whether at least two female characters appeared, and female characters who had more lines were likelier to be noticed by annotators on bechdeltest.com. We experimented with a number of machine learning architectures (Random Forest, kNN, Decision Tree, and SVM), and found that a linear SVM with class weights biased toward the minority class, Fail, performed best.

Method	Fail F1	Pass F1	Macro-F1
RB-H	.269	.903	.586
RB-S	.263	.907	.585
SVM	**.317**	.886	**.602**
Agarwal (SSA)	.24	**.94**	.59

Table 2: Results for T1, cross-validated on development set.

5.2 Results & Discussion

Table 2 shows the results for T1. RB-S predicted Pass on more movies than RB-H, since the requirement for passing was softened when the gender score cutoff was brought down from .9 to .5. This change contributed to a higher Fail F1 for RB-H and a higher Pass F1 for RB-S, but these differences balanced each other out in the macro-F1. Since there was no significant difference between the two, we defaulted to using the "soft" cutoff for the rest of our experiments: the machine learning classifier for T1, and all classifiers for T2 and T3.

Our SVM outperformed the rule-based classifiers and Agarwal et al. (where their only external gender resource was SSA). In particular, our SVM achieved a much higher Fail F1, due to the class weights and inclusion of non-binary features that distinguished between more present and less present female characters. Still, it is worth noting that Agarwal et al.'s best T1 classifier achieved macro-F1 of .75 by using IMDb and Stanford's named entity coreference resolution system.

6 Test 2: Do these women speak to one another?

6.1 Types of Interaction

To answer this test, we needed to consider what would indicate two characters speaking to one another. We based our design on two approaches introduced in Agarwal et al., since they were simple, intuitive, and proved useful in their work. We call the first approach OVERLAP, where two women speak to one another if they overlap in a scene. The second approach we call CONSECUTIVE, where two women speak to one another if they speak consecutively in a scene. Using the scene extractions we developed during preprocessing, we built rule-based classifiers for these methods (RB-O and RB-C).

However, even passing CONSECUTIVE might not guarantee that two women are actually speaking to one another. For example, a scene could have groups of people in different parts of a room, where two characters speaking consecutively might be in different groups and not interacting with each other. To address this, we added a third approach CONSECUTIVE-STRONG, where the two women must speak consecutively for more than four lines. If there was an exchange longer than four lines between only these women, it was very likely that they had spoken to one another.

Raising the criteria for passing (for instance, upgrading CONSECUTIVE to CONSECUTIVE-STRONG) would result in a trade-off between higher precision when predicting Pass cases but also lower recall, mistaking true Pass cases for Fail. To avoid this trade-off, we designed features for our machine learning classifier that could capture all levels of criteria. We extracted features by rating each scene in the movie by the highest criteria they could fulfill and counted the number of scenes that fell into each rating. A scene could receive a rating of:

3 if it contained a CONSECUTIVE-STRONG,

2 if it contained a normal CONSECUTIVE,

1 if it contained an OVERLAP,

0 if it contained none of the above.

We provided these features for machine learning classification and experimented with a number of architectures (Random Forest, kNN, Decision Tree, and SVM). Again, we found that a linear SVM with class weights biased toward the minority class, Fail, performed best.

Method	Fail F1	Pass F1	Macro-F1
RB-O	.367	.772	.574
RB-C	.507	**.78**	.644
SVM	**.627**	.758	**.693**
Agarwal	.39	.77	.58

Table 3: Results for T2, cross-validated on development set.

6.2 Results & Discussion

RB-C did significantly better than RB-O, improving Fail F1 drastically and Pass F1 slightly. This mirrored Agarwal et al.'s findings, and their best T2 classifier (Table 3) also used CONSECUTIVE.

Our SVM outperformed the rule-based classifiers and Agarwal et al. by 11 points. We wanted to see whether the boost in performance was because

we included screenplays from Gorinski in our corpus and perhaps they were easier to predict than the Agarwal set. However, our SVM was able to achieve macro-F1 .671 on just the Agarwal set, which was still 9 points above their benchmark. We believe the inclusion of different levels of criteria contributed to this increase in performance.

7 Test 3: Do these women talk to each other about something besides a man?

This final subtest was the most complex and required extensive feature engineering. To start, we still developed a rule-based classifier, which gathered all female-female conversations in a screenplay and predicted Pass if there was at least one conversation in which no male pronouns or male characters were mentioned. This classifier achieved macro-F1 .501, demonstrating the need for better features and machine learning techniques.

For our machine learning classifier, we first tried two unigram baselines. One was a bag-of-words of all dialogue by female characters (UNI-F) and the other was a bag-of-words of all dialogue in female-female conversations (UNI-FF). Secondly, we tried incorporating features inspired by our rule-based classifier (RB*). The features were two-fold: the number of female-female conversations that did not mention a male pronoun or male character, and the number of female-female conversations overall. Thirdly, we experimented with a collection of social network analysis features (SNA), and lastly, we tried incorporating Sap et al.'s connotation frames of power and agency into our prediction (FRA).

7.1 Social Network Analysis

Agarwal et al. found that social network features were the most helpful for their T3 classifier, so we pursued this approach as well. We built social networks based on character interaction, and drawing from T2, we experimented with interaction defined as overlapping in a scene versus speaking consecutively in a scene (SNA-O and SNA-C). OVERLAP networks were more dense and, when measuring centrality, would favor characters who appeared in more scenes. CONSECUTIVE networks were less dense and, when measuring centrality, would favor characters who spoke more often, especially to different people. Figures 1 and 2 demonstrate this comparison for the movie "10 Things I Hate About You" (1999). The pink and blue nodes indicate female and male characters respectively, as determined by our name-to-gender algorithm, and the grey nodes are characters for whom the algorithm could not determine gender.

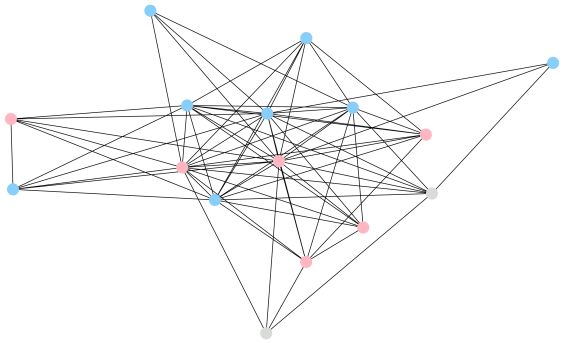

Figure 1: OVERLAP network of "10 Things I Hate About You" (1999)

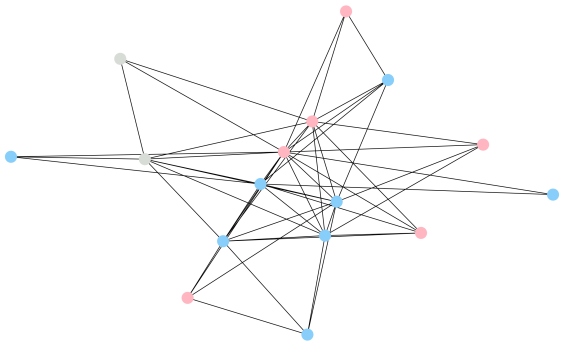

Figure 2: CONSECUTIVE network of "10 Things I Hate About You" (1999)

Based on the social networks we built, we considered the degree, betweenness, closeness, and eigenvector centralities of the characters. For each centrality type, we included the average and sum of the female characters' centralities as well as the average and sum of the male characters' centralities. Inspired by Agarwal et al., we also incorporated as features the number of men who were connected to a woman, the number of women who were connected to a woman, and the percentage of women who were in a 3-person clique with another woman and a man. We lastly tried counting the number of women who were in only-female cliques, which made in total 21 SNA features.

7.2 Power and Agency in Character Dialogue Exchanges

We acquired from Sap et al. their connotation frames of agency and power. The frames set includes 2,000 verbs annotated for how much agency they communicate and 1,700 verbs annotated for the power differentials indicated between agent and theme. The frames are as follows:

agency_pos: The agency attributed to the agent of a verb denotes power and decisiveness.

agency_neg: The agency attributed to the agent of the verb denotes passiveness and indeterminism.

agency_equal: The agency attributed to the agent of the verb neither denotes power nor inability.

power_agent: The agent in the dialogue has more power than the theme.

power_theme: The theme in the dialogue has more power than the agent.

power_equal: Neither the theme nor the agent has more power in the dialogue.

In our social networks, we already captured simple relations between characters by checking whether they interacted. We used the connotation frames to go beyond that simple representation of relations, by studying at the dialogue level the agency and power dynamics expressed in deeper character relationships. Thus, in our analysis of frames we considered scenes in which only two characters appeared, and we focused on comparing female-female scenes (FF), male-male scenes (MM), and female-male scenes (FM) (Schofield and Mehr, 2016). In our corpora, 70 movies contained female-female scenes, 146 contained male-male scenes, and 145 contained female-male scenes.

For each movie's two-person scenes, we looked for the appearance of verbs from the frames set. To maximize the number of verbs we could match, we stemmed each verb in the frames set and the verbs in the scene dialogues with the Porter stemming algorithm. For each scene type, FF, MM, and FM, we derived the average probability of each of the six frames appearing by dividing the number of verbs found for each frame by the overall number of words in that scene type. For FM scenes, we further distinguished between frame probabilities for female characters in FM scenes versus male characters in FM scenes, so we could get a full understanding of gender, agency, and power in movie dialogue.

We included these frame probabilities per scene type as features to our T3 classifier, with 24 frame features in all (FF, MM, FM-F, FM-M * 6 frames). Separately, we analyzed the correlation between these probabilities and Bechdel scores, which we discuss in our penultimate section.

7.3 Results & Discussion

Method	Fail F1	Pass F1	Macro-F1
UNI-F	.292	.803	.548
UNI-FF	.185	.810	.498
SNA-O	.342	.802	.572
SNA-C	.419	**.826**	.622
FRA	.342	.802	.572
SNA-C + FRA	.416	.819	.618
SNA-C + RB*	.453	.815	**.634**
Agarwal	**.56**	.68	.62

Table 4: Results for T3, cross-validated on development set.

Table 4 displays a summary of our results for T3, with all experiments run on a linear SVM with class weights .72 and .28, for Fail and Pass respectively. In line with Agarwal et al., we found that SNA features were the strongest and the CONSECUTIVE version of SNA outperformed the OVERLAP version. We also found that betweenness was the most helpful centrality measure, and out of the other SNA features, all of them contributed besides the number of women in an all-female clique.

On the other hand, unigrams were the weakest features and worsened performance for other features when combined so we did not include them after the first few experiments. The frames worsened SNA performance as well, but they were a decent feature as they performed fairly well on their own. It is likely that the frames were less helpful because we only took 2-person scenes into account. Many movies were missing 2-person scenes, especially the FF type, which meant FRA features might have been sparse.

Importantly, we were able to improve SNA performance with the inclusion of RB*. This is significant because Agarwal et al. found that SNA was so strong that combining it with any other features only worsened performance. Our inclusion of RB* improved SNA performance by over 1 point, and when we tested SNA + RB* on only Agarwal movies

for the purest comparison, our classifier achieved macro-F1 .665, a 4.5 increase over their best performance.

7.4 Evaluation on Final Task

Here, we combined our optimal classifiers from each test to evaluate the final task: does this movie pass the Bechdel test? We first combined the classifiers as a pipeline (PIPELINE): if the T1 classifier predicted Pass on a movie, it would move onto T2; if the T2 classifier predicted Pass, it would move onto T3; if the T3 classifier predicted Pass, the movie would be predicted to pass the overall test. If any classifier predicted Fail, the movie would be predicted to fail the overall test. To prepare this pipeline, we trained each classifier on the movies in the development set that passed the previous subtest (Table 1).

Agarwal et al. used this pipeline on the final task, but we were unconvinced by the architecture. In the case of a false positive at the T1 or T2 stage, the mistake would be propagated down the pipeline. However, the later classifiers might not perform well on this mistake either since they were not trained to classify screenplays that did not pass the previous stage. In the case of a false negative at the T1 or T2 stage, the pipeline would predict Fail immediately, and later classifiers have no chance to correct that mistake. To address the drawbacks of a pipeline, we built a single classifier (SINGLE) that would predict the final task in one go, given the features from the best T1, T2, and T3 classifiers. We trained SINGLE on the entire development set.

Method	Fail F1	Pass F1	Macro-F1
PIPELINE	.779	.655	.717
SINGLE	.787	.693	.740
Agarwal	**.85**	**.73**	**.79**

Table 5: Results for Final Task, evaluated on held-out test set.

As shown in Table 5, our SINGLE classifier outperformed PIPELINE on Fail F1 and Pass F1. This is likely because of the reasons discussed, since SINGLE classifier was not complicated by early false negatives or mistakes being propagated down the pipeline. However, Agarwal et al. outperformed both our models on the final task. Even though we outperformed them on T2 and T3, it is possible that their classifiers lent better to a pipeline approach, since they favored Pass F1 heavily over Fail F1 for T1 and T2, and their T3 classifier had an especially strong Fail F1, so it could catch propagated mistakes. Our classifiers had the opposite strengths which may have worked counter to the pipeline. Our SINGLE model comes closer to Agarwal et al., but as a single classifier for the entire Bechdel test, we would need to do more feature engineering to help it reach optimal performance.

8 Analysis of Bechdel and Connotation Frames of Agency and Power

Outside of classification, we conducted fine-grained analyses of agency and power in character relationships and how they related to Bechdel scores. Tables 6-8 demonstrate our complete results, with each table normalized to [0, 1] by subtracting the minimum in the table from each element and dividing by the table's range.

Frames	B.0	B.1	B.2	B.3
agency_pos	1	.663	.640	.504
agency_neg	0	.280	.280	.238
agency_equal	0	.191	.182	.163
power_agent	.663	.261	.485	.425
power_theme	.331	.331	.205	.196
power_equal	0	.140	.158	.158

Table 6: Results for female-female exchanges, average probability in relation to Bechdel score per movie.

Table 6 considers movies containing female-female scenes, and compares the movies' Bechdel scores to the the probabilities of each frame appearing in the female-female exchanges. Out of the 70 movies that contain these scenes, 1 movie received a score of 0, 4 movies received a score of 1, 12 movies received a score of 2, and 53 movies received a score of 3. As expected, movies that include female-female scenes are much likelier to pass the Bechdel test, and this distribution features a passing rate of over 75% compared to the overall corpus' passing rate of 48%.

We disregard the first column in the table, because only 1 movie received a score of 0. In the remaining three columns, the only frame with noticeable change is *power_agent*, which jumps by over 20% between Bechdel scores 1 and 2 and remains high for score 3. *power_agent* is concerned with the power demonstrated by the speaker, and so it follows

that movies with higher Bechdel scores might have female speakers expressing greater power.

Frames	B.0	B.1	B.2	B.3
agency_pos	1	.963	.945	.945
agency_neg	.2	.145	.181	.154
agency_equal	.127	.109	.172	.118
power_agent	.890	.681	.772	.745
power_theme	0	.054	.072	.036
power_equal	.172	.209	.190	.172

Table 7: Results for male-male exchanges, average probability in relation to Bechdel score per movie.

Table 7 considers frame probabilities and Bechdel scores for movies containing male-male scenes. Out of the 146 movies that contain male-male scenes, 6 movies received a score of 0, 45 movies received a score of 1, 27 movies received a score of 2, and 68 movies received a score of 3. There do not appear to be clear trends in any of the frames with respect to Bechdel scores. This is not surprising, because agency and power expressed by male characters when they are speaking to one another is not likely to impact the Bechdel score, since the score is concerned only with the representation of women in the movie.

Frames	B.0	B.1	B.2	B.3
agency_pos (F)	.617	.941	1	.794
agency_pos (M)	.794	1	.823	.676
agency_neg (F)	.058	.176	.117	.176
agency_neg (M)	.205	.117	.147	.088
agency_equal (F)	0	.147	.088	.058
agency_equal (M)	.088	.147	.117	.058
power_agent (F)	.382	.676	.647	.617
power_agent (M)	.647	.794	.617	.5
power_theme (F)	0	.058	.058	.029
power_theme (M)	0	0	.058	0
power_equal (F)	.058	.264	.176	.147
power_equal (M)	.088	.117	.147	.117

Table 8: Results for female-male exchanges, average probability in relation to Bechdel score per movie.

Table 8 considers frame probabilities and Bechdel scores for movies containing female-male scenes, where the female characters' probabilities are separated from the male characters'. Out of the 145 movies that contain these exchanges, 5 movies received a score of 0, 39 movies received a score of 1, 28 movies received a score of 2, and 73 movies received a score of 3. This table contains insights about power with regards to Bechdel scores. The ratio between the *power_agent* scores of the female and male characters consistently increases as the Bechdel score increases. When the score is 0, female characters are only half as likely as male characters to use a verb indicating their power as an agent. The ratio reaches .85 when the Bechdel score is 1, and at scores 2 and 3, female characters are slightly more likely than male characters to express *power_agent*. Thus, for female-male exchanges, there does seem to be a relationship between power and Bechdel scores. Female characters are initially represented as subordinate to male characters when the Bechdel score is low, but this imbalanced dynamic disappears as the Bechdel score rises.

9 Conclusion & Future Work

In this paper, we extended and improved previous work on automating the Bechdel test.[2] We designed methods to preprocess screenplays, align character names to genders, identify whether characters were interacting, build their social networks, and analyze the agency and power dynamics in their relationships. The features and machine learning parameters that we tuned boosted classification performance on all three subtests of the Bechdel test. We also conducted nuanced analyses of the interaction between Bechdel scores, power, agency, and gender. Sap et al. provided preliminary studies of these interactions, but our analyses of dialogue divided into FF, MM, and FM scene types was novel and contributed new insights.

In the future, we hope to improve performance on the final task, especially with the single classifier architecture. We believe one key focus will be further developing our name-to-gender algorithm; for instance, by identifying who plays each character and aligning gender through metadata on the actor. We also hope to study linguistic cues that would help to determine whether characters are speaking to one another and whether they are speaking about men. Finally, we would like to explore techniques that leverage information in the screenplays besides dialogue and character names, namely scene descriptions and acting prompts.

[2]All of our code is available online at our public GitHub repository: https://github.com/serinachang5/bechdel.

References

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. *North American Chapter of the Association for Computational Linguistics: Human Language Technologies.*

Alison Bechdel. 1986. The Rule. Dykes to watch out for. *Firebrand Books.*

Philip J. Gorinski and Mirella Lapata. 2015. Movie Script Summarization as Graph-based Scene Extraction. *Proceedings of the 2015 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies.* Association for Computational Linguistics.

Maarten Sap, Marcella Cindy Prasettio, Ariel Holtzman, Hannah Rashkin, and Yejin Choi. 2017. Connotation Frames of Agency and Power in Modern Films. *Conference on Empirical Methods in Natural Language Processing.*

Alexandra Schofield and Leo Mehr. 2016. Gender-Distinguishing Features in Film Dialogue. *NAACL 2016 Workshop on Computational Linguistics for Literature.*

Stacy L. Smith, Marc Choueiti, Katherine Piper, Ariana Case, Kevin Yao, and Angel Choi. 2017. Inequality in 900 Popular Films: Examining Portrayals of Gender, Race/Ethnicity, LGBT, and Disability from 2007-2016. University of Southern California, Annenberg School for Communication and Journalism.

Stacy L. Smith, Marc Choueiti, and Katherine Pieper. 2013. Race/Ethnicity in 500 Popular Films: Is the Key to Diversifying Cinematic Content held in the Hand of the Black Director? *Media Diversity & Social Change Initiative.* University of Southern California.

Lenore J. Weitzman, Deborah Eifler, Elizabeth Hokada, and Catherine Ross. 1972. Sex-Role Socialization in Picture Books for Preschool Children, volume 77. *American Journal of Sociology.*

Individual Contributions

Serina designed the first set of scripts to align screenplays to ids and handled character name normalization in Preprocessing. In T1 and T2, she designed the name-to-gender algorithm, implemented the OVERLAP and CONSECUTIVE methods, constructed additional features, and ran experiments for rule-based and machine learning classifiers. For T3, she designed and implemented the rule-based classifier, and built and ran experiments for machine learning models that incorporated unigrams, rule-based features, social network features, and information from the connotation frames. To complete this step, she wrote scripts to create social networks, visualize them, and extract their features. For the final task, she built the pipeline of classifiers and the single classifier and compared their performance.

Navie implemented a sanity check measure for IMDb ids as well as filled in missing ones utilizing IMDbPy. She also worked to extract scene boundaries and dialogues from Agarwal and Gorinski's dataset of screenplays. After extracting dialogue exchanges from scenes that included two-person exchanges, she completed a fine-grained analysis of frames, agency, power, and Bechdel scores in female-female (FF), male-male (MM), and female-male (FM) exchanges. The frames she prepared were also used as a feature in the classification task for T3.

Together, they maintained a shared repository on GitHub and wrote the final paper.