Key Female Characters in Film Have More to Talk About Besides Men: Automating the Bechdel Test

Apoorv Agarwal
Computer Science Department
Columbia University
NY, USA
apoorv@cs.columbia.edu

Jiehan Zheng[†]
Trinity College of Arts and Sciences
Duke University
NC, USA
jiehan.zheng@duke.edu

Shruti Vasanth Kamath[†]
Columbia University
NY, USA
svk2113@columbia.edu

Sriram Balasubramanian[†]
Facebook
CA, USA
grambler@fb.com

Shirin Ann Dey
Columbia University
NY, USA
sad2166@columbia.edu

Abstract

The *Bechdel test* is a sequence of three questions designed to assess the presence of women in movies. Many believe that because women are seldom represented in film as strong leaders and thinkers, viewers associate weaker stereotypes with women. In this paper, we present a computational approach to automate the task of finding whether a movie passes or fails the Bechdel test. This allows us to study the key differences in language use and in the importance of roles of women in movies that pass the test versus the movies that fail the test. Our experiments confirm that in movies that fail the test, women are in fact portrayed as less-central and less-important characters.

1 Introduction

The Bechdel test is a series of three questions, which originated from Alison Bechdel's comic "Dykes to Watch Out For" (Bechdel, 1986). The three questions (or tests) are as follows: (T1) are there at least two named women in the movie? (T2) do these women talk to each other? and (T3) do these women talk to each other about something besides a man? If after watching a movie, the viewers answer "yes" to all three questions, that movie is said to pass the Bechdel test.

[†]These authors contributed equally. The work was done while Jiehan Zheng was at Peddie School and Sriram Balasubramanian was at Columbia University.

The test was designed to assess the presence of women in movies. Some researchers have embraced the test as an effective primary detector for male bias (Scheiner-Fisher and Russell III, 2012). Due to its generality, the Bechdel test has also been used to assess the presence of women in dialogues held on social media platforms such as MySpace and Twitter (Garcia et al., 2014). Several researchers have noted that gender inequality roots itself in both the subconscious of individuals and the culture of society as a whole (Žižek, 1989; Michel et al., 2011; García and Tanase, 2013). Therefore, combining the Bechdel test with computational analysis can allow for the exposure of gender inequality over a large body of films and literature, thus having the potential to alert society of the necessity to challenge the status quo of male dominance.

In this paper, we investigate the task of automating the Bechdel test. In doing so, we aim to study the effectiveness of various linguistic and social network analysis features developed for conducting this task. Our results show that the features based on social network analysis metrics (such as betweenness centrality) are most effective. More specifically, in movies that fail the test, women are significantly less centrally connected as compared to movies that pass the test. This finding provides support for the long held belief that women are seldom portrayed as strong leaders and thinkers in popular media. Our results also show that word unigrams, topic modeling features, and features that capture mentions of men in conversations are less effective. This may look like a rather surprising result since the question,

Human Language Technologies: The 2015 Annual Conference of the North American Chapter of the ACL, pages 830–840,
Denver, Colorado, May 31 – June 5, 2015.

(T3) do these women talk to each other about something besides a man? seems to be one that linguistic features should be able to answer. A closer analysis suggests why this may be the case. Consider the screenplay excerpt in Figure 1 (on the next page). This excerpt is from the movie *Hannah and Her Sisters*, which passes the Bechdel test. Even though the conversation between named women *Mickey* and *Gail* mentions a man (*He*), the conversation is not *about* a man. The conversation is about *Mickey's* brain tumor. Now consider the following (contrived) conversation between the same characters:

> *Mickey*: Ssssss, if i'm in love, I don't know what I'm gonna do.
> *Gail*: You're not in love. Didn't he tell you that it was over.
> *Mickey*: No, naturally

This conversation is clearly about a man (or being in love with a man). Much like the original conversation, this conversation mentions a man only once. The linguistic phenomena that allows us to infer that this contrived conversation is about a man is quite complex; it requires a deeper semantic analysis and world knowledge. First, we need to infer that *it being over* refers to a relationship. Relationships typically have two participants. In order to identify the participants, we need to use world knowledge that relationships can end and that the person ending the relationship was once part of the relationship, and so on. Eventually, we are able to conclude that one of the main participants of the conversation or the event being discussed is a man.

As a first attempt to automate the test, we only experiment with basic linguistic features. However, we believe that the task itself offers an opportunity for the development of— and subsequent evaluation of— rich linguistic features that may be better equipped for determining the *aboutness* of conversations.

The rest of the paper is structured as follows. Section 2 reviews the related literature. Section 3 introduces the terminology regarding movie screenplays that we use throughout the paper. Section 4 describes the data and gold standard used for the purposes of automating the test. Sections 5, 6, and 7 present our approach, evaluation and results for the three Bechdel tests, respectively. We conclude and present future direction for research in Section 8.

2 Related

There has been much work in the computational sciences community on studying gender differences in the way language is used by men versus women (Peersman et al., 2011; Mohammad and Yang, 2011; Bamman et al., 2012; Schwartz et al., 2013; Bamman et al., 2014; Prabhakaran et al., 2014). In fact, researchers have proposed linguistic features for supervised classifiers that predict the gender of authors given their written text (Koppel et al., 2002; Corney et al., 2002; Cheng et al., 2011). There has also been a growth in research that utilizes computational techniques and big data for quantifying existing gender biases in society (Sugimoto et al., 2013; Garcia et al., 2014; Wagner et al., 2015).

More closely related to our application is the ongoing work in the social sciences community regarding the study of gender biases in movie scripts and books (Weitzman et al., 1972; Clark et al., 2003; Gooden and Gooden, 2001; McCabe et al., 2011; Chick and Corle, 2012; Smith et al., 2013). This work has largely depended on manual effort. McCabe et al. (2011) analyzed the presence of male and female characters in titles, and their centralities, in 5,618 children's books. The authors employed multiple human coders for obtaining the relevant annotations. Smith et al. (2013) employed 71 research assistants to evaluate 600 films to study gender prevalence in their scripts. Our work offers computational techniques that may help reduce the manual effort involved in carrying out similar social science studies.

Recently, Garcia et al. (2014) used 213 movie screenplays for evaluating the correlation of two novel scores with whether or not movies passed the Bechdel test. However, the main focus of their work was not to automate the test. The focus of their work was to study gender biases in MySpace and Twitter (using these scores). Nonetheless, we experiment with these scores and in fact they provide a strong baseline for automating the task. Furthermore, we use our previous work (Agarwal et al., 2014b) to *clean* noisy screenplays found on the web and carry out the study on a larger data-set of 457 screenplays.

```
M|                                                              CUT TO:
 |
S|              INT. MICKEY'S OFFICE - NIGHT
 |
N|              Gail, wearing her glasses, stands behind a crowded but well-
N|              ordered desk.  Two assistants, a man and a woman, stand around
N|              her.
 |
C|                                    MICKEY
M|                               (turning to Gail,
M|                               gesturing nervously)
D|                         Sssss, if I have a brain tumor, I
D|                         don't know what I'm gonna do.
M|                               (sighing)
 |
C|                                    GAIL
D|                         You don't have a brain tumor.  He
D|                         didn't say you had a brain tumor.
 |
C|                                    MICKEY
M|                               (sighing)
D|                         No, naturally
```

Figure 1: A scene from the movie *Hannah and Her Sisters*. The scene shows *one* conversation between two *named women Mickey* and *Gail*. Tag S denotes scene boundary, C denotes character mention, D denotes dialogue, N denotes scene description, and M denotes meta-data.

Researchers in the Natural Language Processing (NLP) community have used movie screenplays for a number of different applications. Ye and Baldwin (2008) used movie screenplays for evaluating word sense disambiguation in an effort to automatically generate animated storyboards. Danescu-Niculescu-Mizil and Lee (2011) utilized movie screenplays for studying the coordination of linguistic styles in dialogues. Bamman et al. (2013) used movie plot summaries for finding personas of film characters. Agarwal et al. (2014c) used screenplays for automatically creating the *xkcd* movie narrative charts. In this paper, we use movie screenplays for yet another novel NLP task: automating the Bechdel test.

3 Terminology Related to Screenplays

Turetsky and Dimitrova (2004) described the structure of a movie screenplay as follows: a screenplay describes a story, characters, action, setting and dialogue of a film. The content of a screenplay follows a (semi) regular format. Figure 1 shows a snippet of a screenplay from the film *Hannah and Her Sisters*. A scene (tag "S") starts with what is called the *slug line* (or scene boundary). The slug line indicates whether the scene is to take place inside or outside (INT, EXT), the name of the location ("MICKEY'S OFFICE"), and can potentially specify the time of day (e.g. DAY or NIGHT). Following the scene boundary, is a scene description (tag "N"). A scene description is followed by a character name (tag "C"), which is followed by dialogues (tag "D"). Screenplays also have directions for the camera, such as "CUT TO:, DISSOLVE TO:". For lack of a better name, we refer to these as meta-data (tag "M").

Screenplays are expected to conform to a strict grammar – scene boundaries should be capitalized and start with markers such as INT./EXT., character names should be capitalized with an optional (V.O.) for "Voice Over" or (O.S.) for "Off-screen.", dialogues and scene descriptions should be *indented*[1] at a unique level (i.e. nothing else in the screenplay is indented at this level). However, screenplays found on the web have *anomalies* in their structures (Gil et al., 2011). In order to parse screenplays found on the web, we presented a supervised machine learning approach in Agarwal et al. (2014b). By parsing we mean assigning each line of the screenplay one of the following five tags: {S, N, C, D, M}. We showed that a rule based system, often used in

[1] By level of indentation we mean the number of spaces from the start of the line to the first non-space character.

the literature (Turetsky and Dimitrova, 2004; Weng et al., 2009; Gil et al., 2011), is not well equipped to handle anomalies in the structure of screenplays. Our supervised models outperformed the regular expressions based baseline by a large and significant margin (0.69 versus 0.96 macro-F1 measure for the five classes). We use these parsed screenplays for the purposes of this paper.

Many of our features designed to automate the Bechdel test rely on the definition of a scene and a conversation. We define them here:

Scene: A scene is the span of screenplay that lies between two scene boundaries (tag "S").

Conversation: A conversation between two or more characters is defined as their dialogue exchange in *one* scene.

4 Data

The website `bechdeltest.com` has reviewed movies from as long ago as 1892 and as recent as 2015. Over the years, thousands of people have visited the website and assigned ratings to thousands of movies: movies that fail the first test are assigned a rating of 0, movies that pass the first test but fail the second test are assigned a rating of 1, movies that pass the second test but fail the third test are assigned a rating of 2, and movies that pass all three tests are assigned a rating of 3. Any visitor who adds a new movie to the list gets the opportunity to rate the movie. Subsequent visitors who disagree with the rating may leave comments stating the reason for their disagreement. The website has a *webmaster* with admin rights to update the visitor ratings. If the webmaster is unsure or the visitor comments are inconclusive, she sets a flag (called the "dubious" flag) to *true*. For example, *niel (webmaster)* updated the rating for the movie *3 Days to Kill* from 1 to 3.[2] The dubious flag does not show up on the website interface but is available as a meta-data field. Over the course of the project, we noticed that the dubious flag for the movie *Up in the Air* changed from false to true.[3] This provided evidence that the website is actively maintained and moderated by its owners.

[2] `http://bechdeltest.com/view/5192/3_days_to_kill/`

[3] `http://bechdeltest.com/view/578/up_in_the_air/`

	Train & Dev. Set		Test Set	
	Fail	Pass	Fail	Pass
B. Test 1	26	341	5	85
B. Test 2	128	213	32	53
B. Test 3	60	153	15	38
Overall	214	153	52	38

Table 1: Distribution of movies for the three tests over the training/development and test sets. B. stands for Bechdel.

We crawled a total of 964 movie screenplays from the Internet Movie Script Database (IMSDB). Out of these, only 457 were assigned labels on `bechdeltest.com`. We decided to use 367 movies for training and development and 90 movies (about 20%) for testing. Table 1 presents the distribution of movies that pass/fail the three tests in our training and test sets. The distribution shows that a majority of movies fail the test. In our collection, 266 fail while only 191 pass the Bechdel test.

5 Test 1: are there at least two named women in the movie?

A movie passes the first test if there are two or more *named women* in the movie. We experiment with several name-to-gender resources for finding the character's gender. If, after analyzing all the characters in a movie, we find there are two or more *named women*, we say the movie passes the first test, otherwise it does not.

5.1 Resources for Determining Gender

IMDB_GMAP: The Internet Movie Database (IMDB) provides a full list of the cast and crew for movies. This list specifies a one-to-one mapping from character names to the actors who perform that role. Actors are associated with their gender through a meta-data field. Using this information, we created an individual dictionary for each movie that mapped character names to their genders.

SSA_GMAP: The Social Security Administration (SSA) of the United States has created a publicly available list of first names given to babies born in a given year, with counts, separated by gender.[4] Sugimoto et al. (2013) used this resource for assigning genders to authors of scientific articles. Prabhakaran

[4] http://www.ssa.gov/oact/babynames/limits.html

	Fail Test 1			Pass Test 1			
Gender Resource	P	R	F1	P	R	F1	Macro-F1
IMDB_GMAP	0.35	0.63	0.45	0.97	0.91	0.94	0.71
SSA_GMAP	0.26	0.21	0.24	0.94	0.95	0.94	0.59
STAN_GMAP	0.22	0.96	0.36	0.996	0.74	0.85	0.71
STAN_GMAP+ IMDB_GMAP	0.52	0.55	**0.54**	0.97	0.96	**0.96**	**0.75**

Table 2: Results for **Test 1**: "are there at least two named women in the movie".

et al. (2014) used this resource for assigning gender to sender and recipients of emails in the Enron email corpus. The authors noted that a first name may appear several times with conflicting genders. For example, the first name *Aidyn* appears 15 times as a male and 15 times as a female. For our purposes, we removed such names from the original list. The resulting resource had 90,000 names, 33,000 with the gender male and 57,000 with the gender female.

STAN_GMAP: In our experiments, we found both IMDB_GMAP and SSA_GMAP to be insufficient. We therefore devised a simple technique for assigning genders to named entities using the context of their appearance. This technique is general (not specific to movie screenplays) and may be used for automatically assigning genders to named characters in literary texts. The technique is as follows: (1) run a named entity coreference resolution system on the text, (2) collect all third person pronouns (*she, her, herself, he, his, him, himself*) that are resolved to each entity, and (3) assign a gender based on the gender of the third person pronouns.

We used Stanford's named entity coreference resolution system (Lee et al., 2013) for finding coreferences. Note that the existing coreference systems are not equipped to resolve references within a conversation. For example, in the conversation between *Mickey* and *Gail* (see Figure 1) "He" refers to *Mickey's* doctor, *Dr. Wilkes*, who is mentioned by name in an earlier scene (almost 100 lines before this conversation). To avoid incorrect coreferences, we therefore ran the coreference resolution system only on the scene descriptions of screenplays.

5.2 Results and Discussion

Table 2 presents the results for using various name to gender mapping resources for the first test. Since it is important for us to perform well on both classes (fail and pass), we report the macro-F1 measure; macro-F1 measure weights the classes equally unlike micro-F1 measure (Yang, 1999).

The results show that SSA_GMAP performs significantly[5] worse than all the other name-to-gender resources. One reason is that movies have a number of named characters that have gender different from the common gender associated with their names. For example, the movie *Frozen* (released in 2010) has two named women: *Parker* and *Shannon*. According to SSA_GMAP, *Parker* is a male, which leads to an incorrect prediction (fail when the movie actually passes the first test).

The results show that a combination of STAN_GMAP and IMDB_GMAP outperforms all the individual resources by a significant margin. We combined the resources by taking their union. If a name appeared in both resources with conflicting genders, we retained the gender recorded in IMDB_GMAP. Note that the precision of IMDB_GMAP is significantly higher than the precision of STAN_GMAP for the class Fail (0.35 versus 0.22). This has to do with coverage: STAN_GMAP is not able to determine the gender of a number of characters and predicts fail when the movie actually passes the test. We expected this behavior as a result of being able to run the coreference resolution tool only on the scene descriptions. Not all characters are mentioned in scene descriptions.

Also note that the precision of IMDB_GMAP is significantly lower than the precision of STAN_GMAP for the class Pass (0.97 versus 0.996). Error analysis revealed two problems with IMDB_GMAP. First, it lists non-named characters (such as *Stewardess*) along with the named characters in the credit list. So while the movie *A*

[5]We use McNemars test with $p < 0.05$ to report significance throughout the paper.

	Fail Test 2			Pass Test 2			
Network	P	R	F1	P	R	F1	Macro-F1
CLIQUE	0.55	0.20	0.29	0.65	0.92	0.76	0.57
CONSECUTIVE	0.63	0.28	**0.39**	0.67	0.90	**0.77**	**0.62**

Table 3: Results for **Test 2**: "do these women talk to each other?"

Space Odyssey actually fails the test (it has only one named woman, *Elena*), IMDB_GMAP incorrectly detects *Stewardess* as another named woman and makes an incorrect prediction. Second, certain characters are credited with a name different from the way they appear in the screenplay. Following is a user comment from `bechdeltest.com` on the movie *Up in the Air* that highlights this second limitation:

> Natalie refers to Karen Barnes as "Miss Barnes" when they first meet. She is also named later. Despite the fact that she's credited as "Terminated employee", she's definitely a named female character.

The methodology used for finding named women directly impacts the performance of our classifiers on the next two tests. For instance, if a methodology under-predicts the number of named women in a movie, its chances of failing the next two tests increase. In fact, we experimented with all combinations and found the combination STAN_GMAP+ IMDB_GMAP to outperform other gender resources for the next two tests. Due to space limitations, we do not present these results in the paper. We use the lists of named women and named men generated by STAN_GMAP+ IMDB_GMAP for the next two tests.

6 Test 2: Do these women talk to each other?

So far, we have parsed screenplays for identifying character mentions, scene boundaries, and other elements of a screenplay (see Figure 1). We have also identified the gender of named characters. For automating the second test (*do these women talk to each other?*) we experimented with two techniques for creating *interaction* networks of characters. Consider the following sequence of tagged lines in a screenplay: {S1, C1, C2, C3, S2, C1, ...}. S1 denotes the first scene boundary, C1 denotes the first speaking character in the first scene, C2 the second speaking character, and so on. One way of creating an *interaction* network is to connect all the characters that appear between two scene boundaries (Weng et al., 2009). Since the characters C1, C2, and C3 appear between two scene boundaries (S1 and S2), we connect all the three characters with pair-wise links. We call this the CLIQUE approach. Another way of connecting speaking characters is to connect only the ones that appear consecutively (C1 to C2 and C2 to C3, no link between C1 and C3). We call this the CONSECUTIVE approach.

Results presented in Table 3 show that the CONSEQUITIVE approach performs significantly better than the CLIQUE approach.

We investigated the reason for an overall low performance for this test. One reason was the over-prediction of named women by our gender resource (labeling *Stewardess* as a named woman). Another reason was the inconsistent use of scene descriptions in screenplays. Consider the sequence of scene boundaries, characters, and scene descriptions: {S1, N1, C1, C2, N2, C3, C4, S2, ...}. While for some screenplays N2 divided the scene between S1 and S2 into two scenes (S1-N2 and N2-S2), for other screenplays it did not. For the screenplays that it did, our CONSECUTIVE approach incorrectly connected the characters C2 and C3, which led to an over-prediction of characters that talk to each other. Both these reasons contributed to the low recall for the Fail class.

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

For the third Bechdel test, we experimented with machine learning models that utilized linguistic features as well as social network analysis features derived from the interaction network of characters.

7.1 Feature Set

We considered four broad categories of features: word unigrams (BOW), distribution of conversa-

tions over topics (TOPIC), linguistic features that captured mentions of *men* in dialogue (LING), and social network analysis features (SNA). We additionally experimented with the two scores proposed by Garcia et al. (2014).

For BOW, we collected all the words that appeared in conversations between pairs of women and normalized the binary vector by the number of pairs of named women and by the number of conversations they had in a screenplay. BOW was a fixed feature vector of length 18,889.

The feature set LING consisted of the following features: (1) the average length of conversations between each pair of named women (2) the number of conversations between each pair of named women, (3) a binary feature that recorded if *all* conversations between a particular pair of named women mentioned a man, and (4) a binary feature that recorded if *any* conversation between a particular pair of named women mentioned a man. Let us denote these feature vectors by $\{v_1, v_2, v_3, v_4\}$. Note that the length of these features vectors ($|v_i| \leq \binom{n}{2}$, where n is the number of named women in a movie) may vary from one movie to the other. We converted these variable length vectors into fixed length vectors of length four by using a function, GET_MIN_MAX_MEAN_STD(VECTOR), that returned the minimum, maximum, mean, and standard deviation for each vector. In all, we had $4 * 4 = 16$ LING features for each movie.

We found multiple instances of conversations that were about men but did not explicitly mention a man. For example, *don't we all fall for those pricks?* and *which one did you fall in love with?*. We also found conversations that mentioned a man explicitly and were around the same topic (say, *love*). For example, *I'm not in love with him, okay!*. In an attempt to capture possible correlations between general topics and conversations in which women talked about men, we decided to experiment with features derived from topic models. We trained a topic model on all the conversations between named women (Blei et al., 2003; McCallum, 2002). Before training the topic model, we converted all the mentions of men to a fixed tag "MALE" and all the mentions of women to a fixed tag "FEMALE". For each conversation between every pair of women, we queried the topic model for its distribution over the k topics. Since the number of pairs of women and the number of conversations between them could vary from one movie to the other, we took the average of the k-length topic distributions. We experimented with $k = 2$, 20, and 50. Thus the length of the TOPIC feature vector was 72.

While the Bechdel test was originally designed to assess the presence of women, it has subsequently been used to comment on the importance of roles of women in movies. But does *talking about men* correlate with the importance of their roles? To study this correlation we designed the following set of SNA features. We created variable length feature vectors (length equal to number of women) for several social network analysis metrics (Wasserman and Faust, 1994), all appropriately normalized: (1) degree centrality, (2) closeness centrality, (3) betweenness centrality, (4) the number of men a woman was connected to, and (5) the number of other women a woman was connected to. We created two other variable length feature vectors (length equal to the number of pairs of women) that recorded (6) the number of men in common between two women and (7) the number of women in common between two women. We converted these variable length feature vectors to fixed length vectors of length four by using the GET_MIN_MAX_MEAN_STD(VECTOR) function described above. This constituted $7 * 4 = 28$ of our SNA features. We additionally experimented with the following features: (8) the ratio of the number of women to the number of men, (9) the ratio of the number of women to the total number of characters, (10) the percentage of women that formed a 3-clique with a man and another woman, (11, 12, 13) the percentage of women in the list of five *main* characters (main based on each of the three notions of centralities), (14, 15, 16) three boolean features recording whether the main character was a women, (17, 18, 19) three boolean features recording whether any woman connected another woman to the main man, and (20, 21, 23) the percentage of women that connected the main man to another woman. In all we had 28 + 15 = 43 SNA features.

As a baseline, we experimented with the features proposed by Garcia et al. (2014). The authors proposed two scores: B_F and B_M. B_F was the ratio of {dialogues between female characters that did not contain mentions of men} over {the total num-

			Fail Test 3			Pass Test 3			Macro
Row #	Kernel	Feature Set	P	R	F1	P	R	F1	F1
1	Linear	Garcia et al. (2014)	0.39	0.70	0.50	0.84	0.57	0.67	0.62
2	Linear	BOW	0.40	0.37	0.38	0.74	0.76	**0.75**	0.57
3	Linear	LING	0.39	0.37	0.37	0.75	0.76	**0.75**	0.57
4	Linear	TOPIC	0.28	0.29	0.27	0.71	0.70	0.70	0.50
5	RBF	SNA	0.42	0.84	**0.56**	0.90	0.55	0.68	**0.68**

Table 4: Results for **Test 3**: "do these women talk to each other about something besides a man?" Column two specifies the kernel used with the SVM classifier.

		Fail Test 3			Pass Test 3			Macro
Kernel	Feature	P	R	F1	P	R	F1	Macro-F1
Linear	Garcia et al. (2014)	0.72	0.93	0.81	0.81	0.47	0.60	0.73
RBF	SNA	0.80	0.91	**0.85**	0.83	0.66	**0.73**	**0.80**

Table 5: Results on the unseen test set on the end task: does a movie passes the Bechdel Test?

ber of dialogues in a movie$\}$. B_M was the ratio of $\{$dialogues between male characters that did not contain mentions of women$\}$ over $\{$the total number of dialogues in a movie$\}$.

7.2 Evaluation and Results

There were 60 movies that failed and 153 movies that passed the third test (see Table 1). We experimented with Logistic Regression and Support Vector Machines (SVM) with the linear and RBF kernels. Out of these SVM with linear and RBF kernels performed the best. Table 4 reports the averaged 5-fold cross-validation F1-measures for the best combinations of classifiers and feature sets. For each fold, we penalized a mistake on the minority class by a factor of 2.55 (153/60), while penalizing a mistake on the majority class by a factor of 1. This was an important step and as expected had a significant impact on the results. A binary classifier that uses a 0-1 loss function optimizes for accuracy. In a skewed data distribution scenario where F1-measure is a better measure to report, classifiers optimizing for accuracy tend to learn a trivial function that classifies all examples into the same class as the majority class. By penalizing mistakes on the minority class more heavily, we forced the classifier to learn a non-trivial function that achieved a higher F1-measure.

Results in Table 4 show that the features derived from social network analysis metrics (SNA) outperform linguistic features (BOW, LING, and TOPIC) by a significant margin. SNA features also outperform the features proposed by Garcia et al. (2014) by a significant margin (0.68 versus 0.62). Various feature combinations did not outperform the SNA features. In fact, all the top feature combinations that performed almost as well as the SNA features included SNA as one of the feature sets.

7.3 Evaluation on the End Task

We used the IMDB_GMAP + STAN_GMAP gender resource for the first test, the CONSECUTIVE approach for creating an interaction network for the second test, and compared the performance of the baseline versus the best feature set for the third test. Table 5 presents the results for the evaluation on our unseen test set of 52 movies that failed and 38 movies that passed the Bechdel test. As the results show, our best feature and classifier combination outperforms the baseline by a significant margin (0.73 versus 0.80). Note that the end evaluation is easier than the evaluation of each individual test. Consider a movie that fails the first test (and thus fails the Bechdel test). At test time, lets say, the movie is mis-predicted and passes the first two tests. However, the classifier for the third test correctly predicts the movie to fail the Bechdel test. Even though the errors propagated all the way to the third level, these errors are not penalized for the purposes of evaluating on the end task.

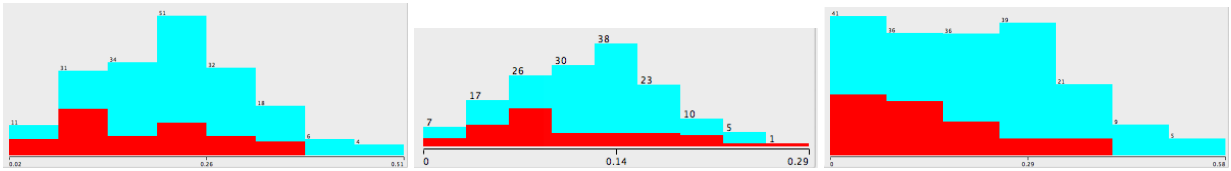

Figure 2: Distribution of three SNA features (left to right): mean degree centrality, mean closeness centrality, and mean betweenness centrality of named women. Red histogram is for movies that fail and the Blue histogram is for movies that pass the third Bechdel Test. The histograms show that the average centralities of women are higher for movies that pass the Bechdel test.

7.4 Discussion

We studied the correlation of our SNA features and the features proposed by Garcia et al. (2014) with the gold class on the set of 183 movies that pass or fail the third test in our training set. The most correlated SNA feature was the one that calculated the percentage of women who formed a 3-clique with a man and another woman ($r = 0.34$). Another highly correlated SNA feature was the binary feature that was true when the main character was a woman in terms of betweenness centrality ($r = 0.32$). Several other SNA features regarding the different notions of centralities of women were among the top. The feature suggested by Garcia et al. (2014), B_F and B_M, were also significantly correlated, with $r = 0.27$ and $r = -0.23$ respectively.

Figure 2 shows the distribution of three of our SNA features: mean degree centrality, mean closeness centrality, and mean betweenness centrality of named women. As the distributions show, most of the mass for movies that fail the test is towards the left of the plot, while most of the mass for movies that pass is towards the right. So movies that fail the test tend to have lower centrality measures as compared to movies that pass the test. Using our classification results, correlation analysis, and visualizations of the distributions of the SNA features, we conclude that in fact, movies that fail the test are highly likely to have less centrally connected women.

8 Conclusion and Future Work

In this paper, we introduced a novel NLP task of automating the Bechdel test. We utilized and studied the effectiveness of a wide range of linguistic features and features derived from social network analysis metrics for the task. Our results revealed that the question, *do women talk to each other about something other than a man*, is best answered by network analysis features derived from the interaction networks of characters in screenplays. We were thus able to show a significant correlation between the importance of roles of women in movies with the Bechdel test. Indeed, movies that fail the test tend to portray women as less-important and peripheral characters.

To the best of our knowledge, there is no large scale empirical study on quantifying the percentage of children's books and novels that fail the Bechdel test. In the future, we hope to combine some of the ideas from this work with our past work on social network extraction from literary texts (Agarwal and Rambow, 2010; Agarwal et al., 2012; Agarwal et al., 2013a; Agarwal et al., 2013b; Agarwal et al., 2014a) for presenting a large scale study on children's book and novels.

Acknowledgments

We would like to thank anonymous reviewers for their insightful comments. We would also like to thank Owen Rambow, Caronae Howell, Kapil Thadani, Daniel Bauer, and Evelyn Rajan for their helpful comments. We thank Noura Farra for suggesting the title for the paper. The idea originated from a class project (Agarwal, 2013). We credit Michelle Adriana Marguer Cheripka and Christopher I. Young for the initial idea. This paper is based upon work supported in part by the DARPA DEFT Program. The views expressed are those of the authors and do not reflect the official policy or position of the Department of Defense or the U.S. Government.

References

Apoorv Agarwal and Owen Rambow. 2010. Automatic detection and classification of social events. In *Proceedings of the 2010 Conference on Empirical Methods in Natural Language Processing*, pages 1024–1034, Cambridge, MA, October. Association for Computational Linguistics.

Apoorv Agarwal, Augusto Corvalan, Jacob Jensen, and Owen Rambow. 2012. Social network analysis of alice in wonderland. In *Proceedings of the NAACL-HLT 2012 Workshop on Computational Linguistics for Literature*, pages 88–96, Montréal, Canada, June. Association for Computational Linguistics.

Apoorv Agarwal, Anup Kotalwar, and Owen Rambow. 2013a. Automatic extraction of social networks from literary text: A case study on alice in wonderland. *In the Proceedings of the 6th International Joint Conference on Natural Language Processing (IJCNLP 2013)*.

Apoorv Agarwal, Anup Kotalwar, Jiehan Zheng, and Owen Rambow. 2013b. Sinnet: Social interaction network extractor from text. In *Sixth International Joint Conference on Natural Language Processing*, page 33.

Apoorv Agarwal, Sriramkumar Balasubramanian, Anup Kotalwar, Jiehan Zheng, and Owen Rambow. 2014a. Frame semantic tree kernels for social network extraction from text. *14th Conference of the European Chapter of the Association for Computational Linguistics*.

Apoorv Agarwal, Sriramkumar Balasubramanian, Jiehan Zheng, and Sarthak Dash. 2014b. Parsing screenplays for extracting social networks from movies. *EACL-CLFL 2014*, pages 50–58.

Apoorv Agarwal, Sarthak Dash, Sriramkumar Balasubramanian, and Jiehan Zheng. 2014c. Using determinantal point processes for clustering with application to automatically generating and drawing xkcd movie narrative charts. *Academy of Science and Engineering (ASE)*.

Apoorv Agarwal. 2013. Teaching the basics of nlp and ml in an introductory course to information science. In *Proceedings of the Fourth Workshop on Teaching NLP and CL*, pages 77–84, Sofia, Bulgaria, August. Association for Computational Linguistics.

David Bamman, Jacob Eisenstein, and Tyler Schnoebelen. 2012. Gender in twitter: Styles, stances, and social networks. *arXiv preprint arXiv:1210.4567*.

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*, pages 352–361, Sofia, Bulgaria, August. Association for Computational Linguistics.

David Bamman, Jacob Eisenstein, and Tyler Schnoebelen. 2014. Gender identity and lexical variation in social media. *Journal of Sociolinguistics*, 18(2):135–160.

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

David M Blei, Andrew Y Ng, and Michael I Jordan. 2003. Latent dirichlet allocation. *the Journal of machine Learning research*, 3:993–1022.

Na Cheng, Rajarathnam Chandramouli, and KP Subbalakshmi. 2011. Author gender identification from text. *Digital Investigation*, 8(1):78–88.

Kay Chick and Stacey Corle. 2012. A gender analysis of ncss notable trade books for the intermediate grades. *Social Studies Research and Practice*, 7(2):1–14.

Roger Clark, Jessica Guilmain, Paul Khalil Saucier, and Jocelyn Tavarez. 2003. Two steps forward, one step back: The presence of female characters and gender stereotyping in award-winning picture books between the 1930s and the 1960s. *Sex roles*, 49(9-10):439–449.

Malcolm Corney, Olivier de Vel, Alison Anderson, and George Mohay. 2002. Gender-preferential text mining of e-mail discourse. In *Computer Security Applications Conference, 2002. Proceedings. 18th Annual*, pages 282–289. IEEE.

Cristian Danescu-Niculescu-Mizil and Lillian Lee. 2011. Chameleons in imagined conversations: A new approach to understanding coordination of linguistic style in dialogs. In *Proceedings of the 2nd Workshop on Cognitive Modeling and Computational Linguistics*, pages 76–87. Association for Computational Linguistics.

David García and Dorian Tanase. 2013. Measuring cultural dynamics through the eurovision song contest. *Advances in Complex Systems*, 16(08).

David Garcia, Ingmar Weber, and Venkata Rama Kiran Garimella. 2014. Gender asymmetries in reality and fiction: The bechdel test of social media. *International Conference on Weblogs and Social Media (ICWSM)*.

Sebastian Gil, Laney Kuenzel, and Suen Caroline. 2011. Extraction and analysis of character interaction networks from plays and movies. Technical report, Stanford University.

Angela M Gooden and Mark A Gooden. 2001. Gender representation in notable children's picture books: 1995–1999. *Sex Roles*, 45(1-2):89–101.

Moshe Koppel, Shlomo Argamon, and Anat Rachel Shimoni. 2002. Automatically categorizing written texts by author gender. *Literary and Linguistic Computing*, 17(4):401–412.

Heeyoung Lee, Angel Chang, Yves Peirsman, Nathanael Chambers, Mihai Surdeanu, and Dan Jurafsky. 2013. Deterministic coreference resolution based on entity-centric, precision-ranked rules. *MIT Press*.

Janice McCabe, Emily Fairchild, Liz Grauerholz, Bernice A Pescosolido, and Daniel Tope. 2011. Gender in twentieth-century childrens books patterns of disparity in titles and central characters. *Gender & Society*, 25(2):197–226.

Andrew Kachites McCallum. 2002. Mallet: A machine learning for language toolkit. http://mallet.cs.umass.edu.

Jean-Baptiste Michel, Yuan Kui Shen, Aviva Presser Aiden, Adrian Veres, Matthew K Gray, Joseph P Pickett, Dale Hoiberg, Dan Clancy, Peter Norvig, Jon Orwant, et al. 2011. Quantitative analysis of culture using millions of digitized books. *science*, 331(6014):176–182.

Saif M Mohammad and Tony Wenda Yang. 2011. Tracking sentiment in mail: how genders differ on emotional axes. In *Proceedings of the 2nd Workshop on Computational Approaches to Subjectivity and Sentiment Analysis (ACL-HLT 2011*, pages 70–79.

Claudia Peersman, Walter Daelemans, and Leona Van Vaerenbergh. 2011. Predicting age and gender in online social networks. In *Proceedings of the 3rd international workshop on Search and mining user-generated contents*, pages 37–44. ACM.

Vinodkumar Prabhakaran, Emily E Reid, and Owen Rambow. 2014. Gender and power: How gender and gender environment affect manifestations of power. In *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing, Doha, Qatar, October. Association for Computational Linguistics*.

Cicely Scheiner-Fisher and William B Russell III. 2012. Using historical films to promote gender equity in the history curriculum. *The Social Studies*, 103(6):221–225.

H. Andrew Schwartz, Johannes C. Eichstaedt, Margaret L. Kern, Lukasz Dziurzynski, Stephanie M. Ramones, Megha Agrawal, Achal Shah, Michal Kosinski, David Stillwell, Martin E. P. Seligman, and Lyle H. Ungar. 2013. Personality, gender, and age in the language of social media: The open-vocabulary approach. *PLoS ONE*.

S.L. Smith, M. Choueiti, E. Scofield, and K. Pieper. 2013. Gender inequality in 500 popular films: Examining onscreen portrayals and behindthescenes employment patterns in motion pictures released between 2007 and 2012. *Media, Diversity, and Social Change Initiative: Annenberg School for Communication and Journalism, USC*.

Cassidy R Sugimoto, Vincent Lariviere, CQ Ni, Yves Gingras, and Blaise Cronin. 2013. Global gender disparities in science. *Nature*, 504(7479):211–213.

Robert Turetsky and Nevenka Dimitrova. 2004. Screenplay alignment for closed-system speaker identification and analysis of feature films. In *Multimedia and Expo, 2004. ICME'04. 2004 IEEE International Conference on*, volume 3, pages 1659–1662. IEEE.

Claudia Wagner, David Garcia, Mohsen Jadidi, and Markus Strohmaier. 2015. It's a man's wikipedia? assessing gender inequality in an online encyclopedia. Arxiv preprint arXiv:1501.06307.

Stanley Wasserman and Katherine Faust. 1994. *Social Network Analysis: Methods and Applications*. New York: Cambridge University Press.

Lenore J Weitzman, Deborah Eifler, Elizabeth Hokada, and Catherine Ross. 1972. Sex-role socialization in picture books for preschool children. *American journal of Sociology*, pages 1125–1150.

Chung-Yi Weng, Wei-Ta Chu, and Ja-Ling Wu. 2009. Rolenet: Movie analysis from the perspective of social networks. *Multimedia, IEEE Transactions on*, 11(2):256–271.

Yiming Yang. 1999. An evaluation of statistical approaches to text categorization. *Information retrieval*, 1(1-2):69–90.

Patrick Ye and Timothy Baldwin. 2008. Towards automatic animated storyboarding. In *AAAI*, pages 578–583.

Slavoj Žižek. 1989. *The sublime object of ideology*. Verso.