Analysis of Gender Roles and Bias in Literary Portrayal of Characters

Megha Srivastava meghas@stanford.edu
Stephanie Wang steph17@stanford.edu
Sarai Gould sarai678@stanford.edu

Abstract

The problem of text classification by author gender has been thoroughly researched, with classification reliably achieving up to 80% accuracy. Our goal is to not only the gender of the author, but also the gender of a target character by using a combination of lexical and syntactic features and machine learning techniques to determine the gender of the author and of a target character of a previously unseen sentence. Our techniques can infer the author/character gender combination with approximately 77% accuracy. The same techniques can be used to determine the gender of the documents author with approximately 90% accuracy.

1 Introduction

Gender bias in literature is an issue of immense cultural and historical interest. While there have been many studies on author gender using computational approaches, it appears that the question of how different genders are written about has lacked the same computational rigor. Moreover, most studies on author gender and gender-influenced topics have focused on online forms of communication that capture recent trends, such as Twitter or Reddit (Bamman 2012, Schrading 2015) . While our contribution does not encompass techniques from all of the existing literature, it does show how computational techniques can be applied to the analysis of gender in literary text, allowing the humanities to wrangle large corpora which otherwise could not be exhaustively examined by hand.

2 Prior Literature

Many have sought to analyze the general question of gender in literature from a computational approach. In their paper, McCabe et. al. sought to examine gender disparity in children's books and its correlation to feminist activism (McCabe, 2011). Their definition of gender disparity was simply the percentage of female characters versus male characters in a novels title and story, and thus lacked insight into the context and representation of gender within the text, or how writing may differ among female and male authors.

Literary research can be improved upon with the use of NLP tools, and recently there has been an explosion in the research of gender in text through categorization by stylometric or topic differences. A stylometric approach to text classification by Koppel et. al. sought to automatically determine the gender of the documents author using a feature set of function words and part of speech tags (Koppel, 2002). By applying machine learning algorithms to a corpus of documents, they achieved 80% accuracy on classifying an unseen document as written by male or female. A topic modeling approach by Vogel and Jurafsky found differences in research topics among men and women in the ACL Anthology Network, observing that women published more often on dialog, discourse and sentiment while men publish more on parsing, formal semantics, and finite state models (Vogel, 2012). Other papers focus on differences of gender representation within text. Hota et. al. classified Shakespeare characters as male or female based on their dialog, and obtained accuracies as high as 82% (Hota, 2006).

Our research expands on previous work by focusing on the stylometric differences between male and female authors with the aim of classifying text by not only author gender, but also by gender of a target character within a context of a document. Thus, we explore not only text classification by author, but also differences in gender representation by authors of different genders.

3 Methodology

3.1 Literature Data

A well-known resource for literary books is Project Gutenberg. Project Gutenberg is the oldest digital library, and includes several canonical books in literature. The collection consists mainly of full texts of public domain books - therefore, there were very few texts from recent years. Not only does the large length of text provided by Project Gutenberg serve useful for making general analysis, but its inclusion of famous literary books helps our data capture books that influenced society. Project Gutenbergs book catalog spans a variety of countries of origin and genres, and Project Gutenberg has already assigned categories to each book. Some books were found in multiple categories. We chose to specifically download books from 30 categories including:

> *Africa, Atheism, Australia, Bahai, Buddhism, Bulgaria, Children's Fiction, Children's Myths, Christianity, Czech, Egypt, France, Germany, Gothic Fiction, Greece, Hinduism, India, Islam, Italy, Judaism, Native America, New Zealand, Norway, Plays, Science Fiction, South Africa, South America, United Kingdom, United States, Western*

The publication dates of the books downloaded spanned 99 AD (translations) to 2008. Figure 1 below shows the frequency of publication dates for male and female authors. For each genre category, we created two folders for Male authors and Female authors. We then proceeded to download each .txt file in each category, placing it in the respective folder based on author gender. This was done manually, to ensure 100% accuracy. We also manually embedded the publication year date, author name, and sub-category within the title of each embedded file. Each .txt file contains a Project Gutenberg-added header and additional information before the actual text of the book - however, since such sections did not contain characters or gender pronouns, we did not expect the headers to interfere with our data. Overall, we used 701 texts (12 other texts were unable to be processed for syntactical parses due to memory issues), with the largest category Children's Fiction / Male containing 250 files. Figure 2 displays the frequency of genres by gender in our dataset.

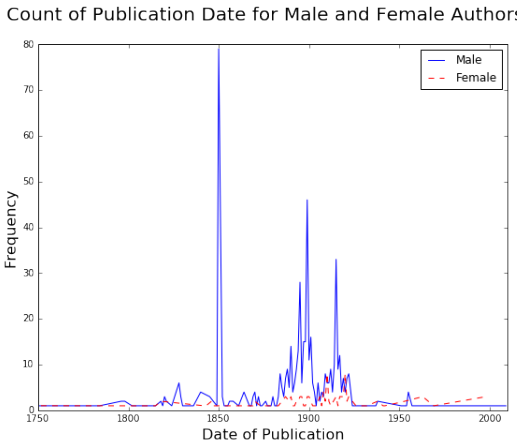

Figure 1: Plot of publication dates of male and female authors

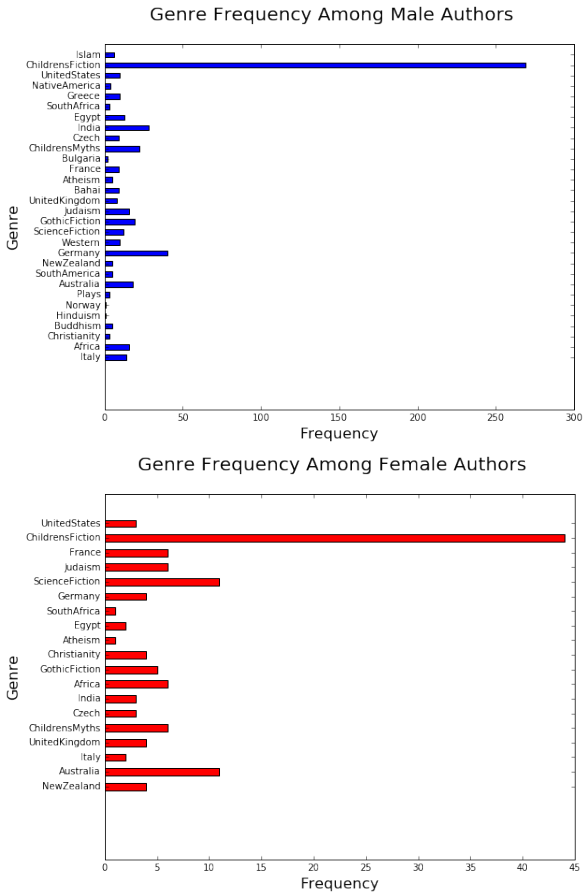

Figure 2: Counts of genre frequency among male and female authors

Figure 3 shows that our dataset contains many skews reflective of the history of literature. For example, there was not a single genre in which there were more female authors than males. The gender disparity differed among the genres, so we expected that some genres would be strong features for predicting author gender.

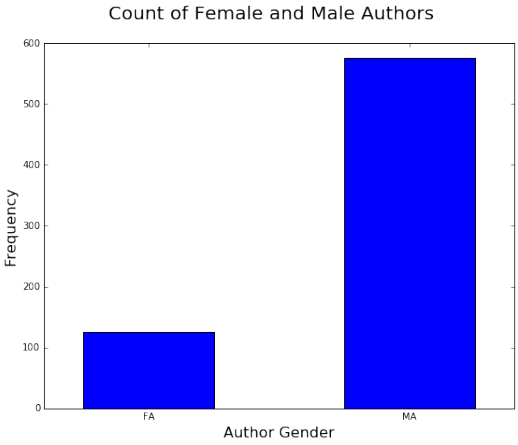

Figure 3: Count of total male and female authors

3.2 Processing Data

We first used NLTKs sentence tokenizer on each file to separate each sentence by line (Bird, 2006). We believe that the scope of a character and their gender is best viewed by sentence: a portion of a sentence is too small to capture dependencies that are far apart, and multiple sentences may contain too many characters and pronouns to have the level of specificity necessary to understand how a single character is portrayed. To determine which sentences contained words and names indicating gender, we used both a name gender classification tool and pattern matching for gendered-pronouns. We modified Stanford Professor Dan Jurafskys Gender Classification tool from his course CS124: From Languages to Information to return Male or Female if a provided word had a greater than .180 likelihood of matching a male/female name. The dataset of names were taken from the U.S. 1990 Census Data. For both Male and Female data, .180 was a high threshold and no names above the threshold were also commonly used words. Therefore, we expected there to be very few instances of words being mis-tagged as a characters name. We also matched and extracted all sentences that matched gendered pronouns and other indicators. These are found in Figure INSERTNUM. We tagged every match and gender-classified name with the appropriate 00FEMALE00 or 00MALE00 tag.

3.3 Dependency Parsing

For each sentence in each file, we generated dependency parses. Due to both the amount of sentences and the length of many sentences (we had several sentences over a 150 words), this was a time-intensive process. We used the Stanford NLP Groups English Probabilistic Context-Free Grammar, which on average parses a 30-word sentence in .6 seconds (Klein, 2003). Each sentence in all 701 files was eventually parsed for its dependencies in around 60 hours. Each sentence was associated with a list of dependencies as a feature. Each dependency was represented as a triple - the type of the dependency, and its two targets. Because we used the parser on the plain text of each file to retain all syntactical clues that may aid the parser, we did a post-processing step where each gendered pronoun/character name was replaced with 00UNKNOWN00. We used a non-gendered label because the dependencies are used as features, and therefore should not include any gender-indicative clues. Therefore, for every sentence in our entire dataset of literary texts, we had the following information:

1. The sentence itself
2. The sentences dependencies
3. The publication year of the text the sentence belongs to
4. The genre of the text the sentence belongs to
5. The gender of the author of the text the sentence belongs to
6. The gender of the character in the sentence

We used variations of the first 4 pieces of information to predict the last 2: specifically, our goal was to classify each sentence by the gender of the character and the gender of its author.

4 Classification Results

We used a Maximum Entropy Classifier with k-cross validation. Due to the large size of our data, the time training our entire data was around 3 hours. However, we also narrowed our dataset by genre and publication year for different experiments - further work could examine an even broader range of genres and publication date to identify interesting social differences in literature.

During preliminary training, we found that, as hypothesized above, genre was a significant feature when training for the entire dataset, with a classification accuracy of .770. For example, genres with no female authors had extremely negative weight for male author labels. Future work could standardize the amount of text for each genre - however, due to the extremely low numbers of

female-authored works, this would result in a significantly smaller data set. Thus, while genres significance does reflect genuine bias in literary history (or at the very least, the literary history Project Gutenberg is able to access), we chose to remove genre as a feature for future experiments.

We specifically report, compare, and analyze classification results for the following 5 experiments:

1. Classification of character gender/author gender for Entire Data
2. Classification of only author gender for Entire Data
3. Classification of only author gender for Childrens Fiction Data
4. Classification of character gender/author gender for Childrens Fiction Data
5. Classification of character gender/author gender for Science Fiction Data
6. Classification of character gender/author gender for Gothic Fiction Data

In this paper, we only show features for Male-Character-Male-Author and Female-Character-Male-Author labels when classifying character genders. Furthermore, we only present the top features after filtering out proper-nouns - it is important to note that name-unigrams held significant weight as features, yet the names were too unique to take care of during pre-processing. We will further discuss handling names in the error analysis section. We first trained and ran our classifier over

Experiment 1: Classification for Entire Text, All Labels

Features: unigrams, sentence length, dependencies, publication year

Labels: FCFA, FCMA, MCFA, MCFCFA, MCFCMA, MCMA (of the form gender_character_gender_author)

Accuracy: 0.713

	precision	recall	f1-score	support
FCFA	0.614	0.385	0.473	23110
FCMA	0.524	0.276	0.362	38986
MCFA	0.553	0.231	0.326	22800
MCFCFA	0.634	0.215	0.321	7049
MCFCMA	0.520	0.256	0.343	20313
MCMA	0.747	0.953	0.838	201222
avg / total	0.678	0.713	0.671	313480

Top Features (FCMA)	Weights	Bottom Features (FCMA)	Weights
compound_merle_00unknown00	2.793172	nmod:poss_comrades_00unknown0	-2.178605
compound_00unknown00_mount	2.528944	nmod:poss_bride_00unknown00	-2.221308
amod_wife_my	2.397321	compound_00unknown00_king	-3.152674
nmod:of_board_00unknown00\n	2.320409	nmod:poss_wife_00unknown00	-4.849197

Top Features (MCMA)	Weights	Bottom Features (MCMA)	Weights
compound_00unknown00_steam	1.850770	compound_00unknown00_queen	-3.603606
compound_00unknown00_telegraph	1.841681	nmod:poss_lover_00unknown00	-3.871616
amod_00unknown00_spiritual	1.832724	compound_00unknown00_miss	-4.938606
nmod:poss_husband_your	1.801239	compound_00unknown00_santa	-5.002877
rajah	1.778931	nmod:poss_husband_00unknown00	-5.820444

every sentence in our data file to predict both author gender and character gender (Experiment 1). When looking at the data, we see expected results regarding king, queen, wife, and husband. However, the classifier does pick up some interesting results - for example, rajah, steam, and telegraph were all features that appeared either by themselves or as compounds that we did not expect. Since many of the literature revolves around colonialism and fiction with a historical angle, the classifier is likely picking up noise. While the overall accuracy of .713 is decent, it dramatically improved when we classify solely on author gender.

Experiment 2: Classification for Entire Text, Author Gender

Features: unigrams, sentence length, dependencies, publication year

Labels: Female Author (FA), Male Author (MA)

Hyper-parameter Accuracy: **Accuracy: 0.930**

	precision	recall	f1-score	support
FA	0.796	0.528	0.635	19378
MA	0.941	0.982	0.962	149757
avg / total	0.925	0.930	0.924	169135

Top Features (MA)	Weights	Bottom Features (MA)	Weights
0000	4.805266	paddock	-1.977515
de	2.027389	totem	-1.991515
colonel	1.984995	root_root_went\n	-2.008694
1700	1.841898	compound_00unknown00_mr	-2.126881
king	1.818415		

With an accuracy of .930, our classifier does remarkably well in classifying author gender across the entire corpora. The top weighted features include two century categories which, based on previously mentioned analysis, correspond to time periods where there were virtually no female authors. The presence of king and colonel give insight into the nature of books written by male authors - revolving around imperialism and battles, while female author texts tend to focus on more domestic themes.

4.1 Results from Children's Fiction

We also decided to narrow in our classification to simply Childrens Fiction. This genre is big in scope given that there are many themes discussed in childrens literature, but we also hypothesized that there may be bigger discrepancies within gender bias, given the more simplifying, objectifying way authors may treat various themes.

Experiment 3: Classification for Children's Fiction Text, Author Gender

Features: unigrams, sentence length, dependencies, publication year

Labels: Female Author (FA), Male Author (MA)

Accuracy: 0.933

	precision	recall	f1-score	support
FA	0.840	0.666	0.743	22728
MA	0.945	0.978	0.961	133591
avg / total	0.930	0.933	0.930	156319

Top Features (MA)	Weights	Worst Features (MA)	Weights
0000	3.002057	grown-up	-2.056211
colonel	2.062386	1900	-2.218947
rifle	2.012736	compound_robinson_tom	-2.339785
compound_00unknown00_santa	1.991027	horrid	-2.345557
dollars	1.984062	silly	-2.353276

Indeed, performance in distinguishing author gender was slightly better when narrowing our

data to be within. Childrens Fiction. We again found interesting discrepancies, such as rifle being a strong indicator for male-author childrens book, versus more descriptive language like horrid and silly that were more skewed to female authors.

We also tried to classify childrens fiction text into both author gender and character gender labels. Similarly, performance improved relative to the overall text, supporting our hypothesis that the childrens fiction data would be more easily skewed.

Experiment 4: Classification for Children's Fiction Text, All Labels

Features: unigrams, sentence length, dependencies, publication year

Labels: FCFA, FCMA, MCFA, MCFCFA, MCFCMA, MCMA (of the form gender_character_gender_author)

Accuracy: 0.766

	precision	recall	f1-score	support
FCFA	0.648	0.475	0.548	13297
FCMA	0.535	0.226	0.318	14979
MCFA	0.534	0.205	0.296	6778
MCFCFA	0.508	0.109	0.180	2631
MCFCMA	0.528	0.211	0.301	8828
MCMA	0.797	0.970	0.875	109806
avg / total	0.728	0.766	0.725	156319

Top Features (FCMA)	Weights	Bottom Features (FCMA)	Weights
candy	2.711559	compound_00unknown00_don	-1.661208
nmod:poss_deck_00unknown00	2.283690	det_00unknown00_every	-1.725603
compound_00unknown00_northern	2.065569	nmod:poss_comrades_00unknown00	-1.830699
nmod:poss_husband_00unknown00	2.037690	compound_00unknown00_00unknown00	-2.157311
advmod_eagerly_00unknown00	1.969501	nmod:poss_wife_00unknown00	-4.167310

Top Features (MCMA)	Weights	Bottom Features (MCMA)	Weights
compound_00unknown00_old	1.866423	compound_00unknown00_northern	-3.377896
professor	1.794379	nmod:of_board_00unknown00	-3.604916
comrade	1.766874	compound_00unknown00_santa	-4.240189
compound_00unknown00_steam	1.764170	compound_00unknown00_miss	-4.832424
colonel	1.681532	nmod:poss_husband_00unknown00	-5.423610

Although Childrens Fiction did have a skew towards male authors, it was the largest category in our dataset - also supporting its high performance. We therefore decided to compare this genre with two other genres - Science Fiction, which was one of our more balanced categories, and Gothic Fiction, which has one of our largest skews.

4.2 Results from Science Fiction

This Science Fiction genre had a balance of 12 texts from male authors, and 11 texts from female authors. Our classification achieved an accuracy of .649, which performed less than both the overall data set and Childrens Fiction. The best performing labels were Male Character Female Author (precision .664, recall .715) and Male Character Male Author (precision .649, recall .833). In this dataset, the number of training examples for male characters in science fiction text is greater than the number of training examples for female characters for both male and female authors.

We analyzed the accuracy of four different feature sets with the Science Fiction classification, results shown in figure 4. It is interesting that including the date of publication increases accuracy substantially. Our dataset contains a large number of male authored text from the 1800s, and this was picked up by our model and used as a highly weighted feature indicating male authorship. Interestingly, sentence length appeared as a high-weighted feature in this classification while it was not as important in previous classifications. The shorter sentence length seems to correlate more with Female Characters. The features also appear genre-specific: the dependency The Terran refers to a Science Fiction species.

Experiment 5: Classification for Science Fiction Text, All Labels

Features: unigrams, sentence length, dependencies, publication year

Labels: FCFA, FCMA, MCFA, MCFCFA, MCFCMA, MCMA (of the form gender_character_gender_author)

Accuracy: 0.649

	precision	recall	f1-score	support
FCFA	0.479	0.143	0.221	237
FCMA	0.500	0.024	0.046	165
MCFA	0.664	0.715	0.689	1155
MCFCFA	0.552	0.158	0.246	101
MCFCMA	0.579	0.136	0.220	81
MCMA	0.649	0.833	0.730	1299
avg / total	0.628	0.649	0.608	3038

Top Features (FCMA)	Weight	Bottom Features (FCMA)	Weight
3	1.043406	again	-0.743203
dice	0.992255	18	-0.706965
advmod_looked_scared\n	0.953136	1800	-0.668030
Nothing	0.942098	24	-0.652123
dance	0.901463	0000	-0.616436

Top Features (MCMA)	Weight	Bottom Features (MCMA)	Weight
0000	2.448105	det_foanna_the	-0.941584
root_root_said\n	1.097934	blue	-0.981144
black	1.009642	det_terran_the	-1.009817
huge	0.981770	sea	-1.109663
0.915153	0.915153	1800	-2.691125

Feature Set	Accuracy
UNI/LEN/DEP/PUB	64.9%
UNI/LEN/DEP	62.1%
UNI/LEN/PUB	64.4%
UNI	63.6%

Figure 4: Science Fiction classification accuracy on different feature sets

4.3 Results from Gothic Fiction

Finally, we ran classification on our most skewed genre - Gothic Fiction. This skew appeared to have a negative effect - Gothic Fiction performed worse than all other classifications. Before we filtered out the names, Gothic Fiction gave enormous weight to names. This can be supported by looking through the text files - the material deals heavily with names, often complicated, unique names

Experiment 6: Classification for Gothic Fiction, All Labels

Features: unigrams, sentence length, dependencies, publication year

Labels: FCFA, FCMA, MCFA, MCFCFA, MCFCMA, MCMA (of the form gender_character_gender_author)

Accuracy: 0.575

	precision	recall	f1-score	support
FCFA	0.553	0.498	0.524	1633
FCMA	0.427	0.257	0.321	1923
MCFA	0.469	0.370	0.413	1350
MCFCFA	0.660	0.225	0.335	926
MCFCMA	0.469	0.138	0.213	921
MCMA	0.615	0.889	0.727	5562
avg / total	0.554	0.575	0.534	12315

Top Features (FCMA)	Weight	Bottom Features (FCMA)	Weight
nmod:poss_sister_my	1.915850	chief	-1.491250
cc_grew_but	1.890909	misery	-1.529181
case_count_'s	1.849870	compound_falkland_00unknown00	-1.537951
nmod:poss_husband_00unknown00	1.819506	compound_00unknown00_lord	-1.693987
quo	1.791940	nmod:poss_wife_00unknown00	-1.858223

Top Features (MCMA)	Weight	Bottom Features (MCMA)	Weight
reverend	1.690874	amod_00unknown00_lovely	-1.846338
sin	1.661296	partiality	-1.863762
nmod:poss_wife_00unknown00	1.551834	nmod:poss_lover_00unknown00	-2.040852
school	1.536203	compound_daae_00unknown00	-2.470495
compound_00unknown00_don	1.520137	nmod:poss_husband_00unknown00	-3.663067

that our processing step would not have caught.

4.4 Feature Reduction

We also wanted to see how many features were actually useful in labeling our text. We did this by selecting only a top percentage of features to train and test with. Using chi squared feature selection, the top .50 features were selected and run in the train and test trial. Further iterations used only .20 of the top features and next .10, .05 and finally .01 of the full feature set. The results of reducing the number of features in each iteration are shown in Figure 5. For Childrens Fiction, a full feature set

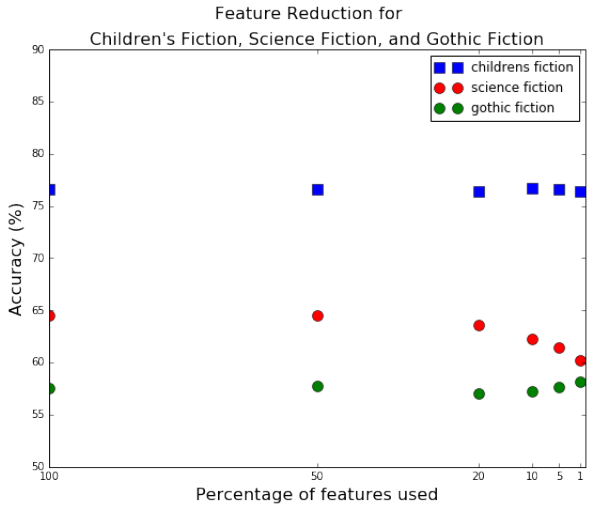

Figure 5: Accuracy obtained using chi squared feature selection

gave .766 accuracy while using only the top .01 of features gave .764 accuracy. In Gothic Fiction, accuracy using .01 of the features reached a peak of .582, which is higher than the .575 accuracy obtained using the full feature set. In both Childrens Fiction and Science Fiction there is an extremely slow drop off in accuracy, suggesting that a select few features contribute most to our categorization task.

5 Analysis

Our results demonstrate that there are, indeed, differences between different genres in classification. Not only does language and theme vary across genre, but gender disparity also differs. Future work includes comparing genres of different cultures - such as Australian fiction versus South African fiction. Our results also show that author-only classification is extremely accurate compared to including character labels. Even when using hyper-parameters for our Children's Fiction to avoid over-sensitivity, we achieved high-results. However, the immense amount of unique characters and places in fiction could simply contribute to this result.

Although we searched for and reported only features that were not proper nouns, to get a better sense of features beyond names and places, in all of our classifications these unique names performed the best. Unfortunately, in the literary world it is difficult to account for every imagined character and place due to the infinite possibilities due to an authors creation. For example, names such as Boswellister, Brabo, and Atley did not appear in the gender-classification tool, let alone met our threshold. However, names such as Sierra had incredibly high probability to be matched as a Female name, yet resulted in most of the books in the Africa genre to mis-tag Sierra-Leone. Thus, a balance needs to be created between accounting for many names without creating false-positives - especially important due to the strength of individual names as features. However, our classifier did show promise in being used as a tool to add more to gender corpora, due to its ability to assign gender to uncommon names.

To try to improve our results, we also ran our classifiers using features that were binarized - instead of frequency metrics, representing each feature as present or not-present in the overall matrix. However, our results were almost exactly the same when using this approach - indicating that given our dataset, the presence of a feature in a sentence is more important than its frequency.

6 Conclusion

Overall, we showed the vast possibilities of using computational tools to analyze literary texts. We were only able to offer a glimpse of the several techniques researchers can use to analyze literary data and gender. The intersection between the humanities and Natural Language Processing creates a scope to use computational power and speed to understand human history and our social interactions. All our methods and dataset are available to the public, and we hope to have inspired other researchers to view literature as a valuable, rich source of data for Natural Language Understanding tasks.

References

Bamman, David, Jacob Eisenstein, Tyler Schnoebelen. Gender in Twitter: Styles, Stances, and Social Networks. *Gender in Twitter: Styles, Stances, and Social Networks.* Sept. 2012. Web. 06 June 2016.

Bird, Steven. NLTK: the natural language toolkit. *Proceedings of the COLING/ACL on Interactive presentation sessions* (COLING-ACL '06). Association for Computational Linguistics, Stroudsburg, PA, USA, 69-72. 2006. Web. June 2016.

Hota, Sobhan R., Shlomo Argamon, and Rebecca Chung. "Gender in Shakespeare: Automatic Stylistics Gender Character Classification Using Syntactic, Lexical and Lemma Features." *ResearchGate*. N.p., 2006. Web. 03 May 2016.

Klein, Dan and Christopher D. Manning. Accurate Unlexicalized Parsing. *Proceedings of the 41st Meeting of the Association for Computational Linguistics*, pp. 423-430. 2003. Web. June 2016.

Koppel, Moshe, Shlomo Argamon, and Anat R. Shimoni. "Automatically Categorizing Written Texts by Author Gender." *Automatically Categorizing Written Texts by Author Gender* (n.d.): n. pag. 2002. Web. May 2016.

McCabe, Janice, Emily Fairchild, Liz Grauerholz, Bernice A. Pescosolido, and Daniel Tope. "Gender in Twentieth-Century Children's Books." *Gender in Twentieth-Century Children's Books.* N.p., Apr. 2011. Web. 03 May 2016.

Schrading, Nicolas, Cecilia Alm, Ray Ptucha, Christopher Homan. An Analysis of Abuse Discourse on Reddit. *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing.* Sept. 2015. Web. 06 June 2016.