***A Catchy Title about Ladies in Movies***

*Subtitle: Something equally catchy about sad gender differences*

By Michelle Kim, I-Ching Wang, and Nikki Haas

***Abstract***

Though there have been many gender studies on popular media, none have focused on contextual and linguistic differences in representation. Through this study, we aim to explore these differences in language and vocabulary in movies, and create a prediction algorithm that will accurately be able to determine whether a line in a script is spoken by a man or woman based on the surrounding context and complexity. Although we have only just begun to scratch the surface of this analysis, we have explored meta statistics with a statistical inference and logistic regression and have determined that our data is best explored via deep learning in neural nets.

***Intro***

For young girls growing up in our society, there are few things more disenfranchising than going to the movies and watching a hapless princess wait for others to act in her place.  Some women in film will be given a voice of their own, but too often men are given the responsibility of conveying the important ideas and plans.  At least, that is how it feels, but is it really true?  For our project, we want to test this hypothesis.  We will use natural language processing and deep learning via convolutional neural nets to evaluate screenplays for hundreds of popular movies from many genres over a period of several years and compare the complexity of scripted lines for male and female roles.

This project is important because young girls would benefit from seeing films that more accurately reflect the various roles women play in society as leaders and innovators.  Not only is length of dialogue important, but the content.  There have been many instances where women roles are given fewer lines, and in addition, the majority of their lines were about their male leads instead of anything substantial that contributes to the plot.  If we can prove that popular movies are following the trope of simplifying women using stringent statistical methods, perhaps we can help Hollywood see its limitations and effect change.

***Background***

In 1985, the cartoonist Alison Bechdel created the Bechdel test for works of fiction. This test checks to see if a) two women in a movie, television show, play, or novel b) speak to each other c) about something other than a male character. Since 1985, this test has been used for online resources, movie critics, and the media to judge the entertainment industry’s representation of women.



Figure 1: Bechdel Test by Alison Bechdel

There are already various studies that focus on the length of screen time characters have based on gender, as well as how much revenue and ratings those types of movies get. We’ve taken a look at some of those studies and found that many times, women had far less screen time than men, even if the storyline has them listed as the main character. However, movies where screen times between the two genders had less of a difference seemed to perform better when they did exist. None of these studies went into an exploration of the actual content of the dialog that was spoken. Our study works to confirm a lot of the trends that these previous studies have shown, and will delve deeper into the differences on complexity of the dialogue between men and women as well as develop a prediction aspect based on the content.

Because there are known differences in the dialogue of men vs women in scripts, we are interested in seeing if these differences widen across genres of movies. By implementing NLP algorithms to process scripts, nature of the line being spoken, and looking at the Bechdel test, we aim to see if our model can properly identify if a line belongs to a man or woman.

That all being said, the better model we are able to create indicates that wider the gender difference is. Essentially, based on current society, we hope to create a model that is accurately discriminating between men and women based on dialogue which we know to have gender differences based on previous research. However, as we learn over time, the model will hopefully perform worse as we slowly move to close the gap between intellectual conversations in media and entertainment that is widely consumed.

***Methods***

*Data Collection*

Our model uses the Film Corpus 2.0 which is a collection of movies from various genres up to the year 2015.

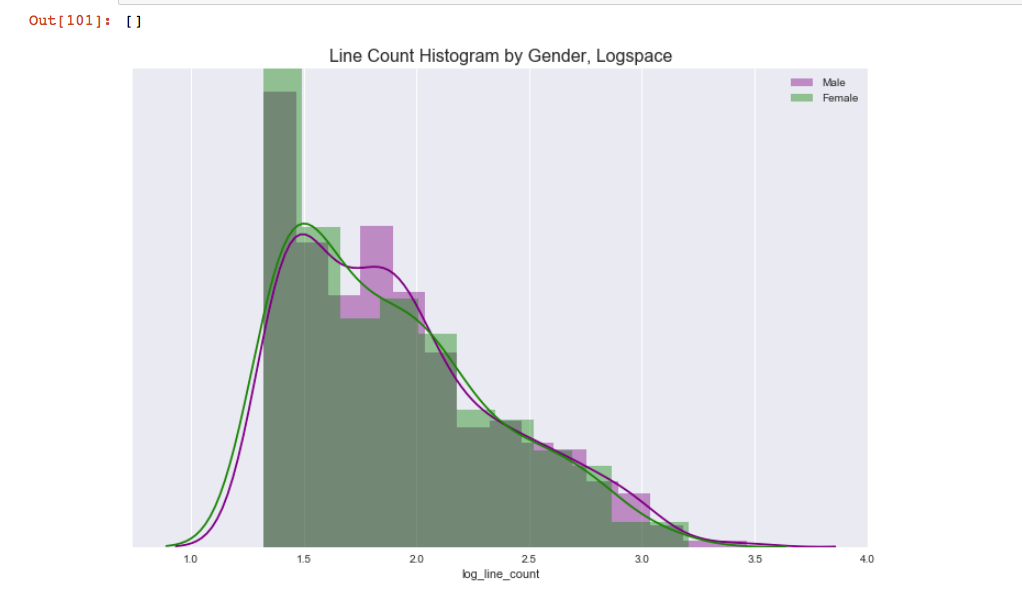
From this Film Corpus, we took all the dialog files and parsed out information about the title of the movie, the genre, and individual lines. Each line represents roughly 10 second of on screen time. We have written a python pre-processing script that matches the lines with the character name that is speaking it. Most of the cues such as Cut to, exit stage, etc have been removed as much as possible. although since each script is written a little differently, there are still some lingering characters in some places. In addition, we have only kept characters that had a total of 20 or more lines. Characters that had too few lines didn’t have much of a vocab was deemed as not important enough to the story line.

We then randomly selected roughly 1000 characters to characterize the gender. This was based on the actor or actress that represented the character, though in few cases, there may be a mismatch. This is more common with LGBT characters or characters in animation, where a character may be voiced by an actor or actress of the opposite gender. The rest of the characters were used to build and test our model.

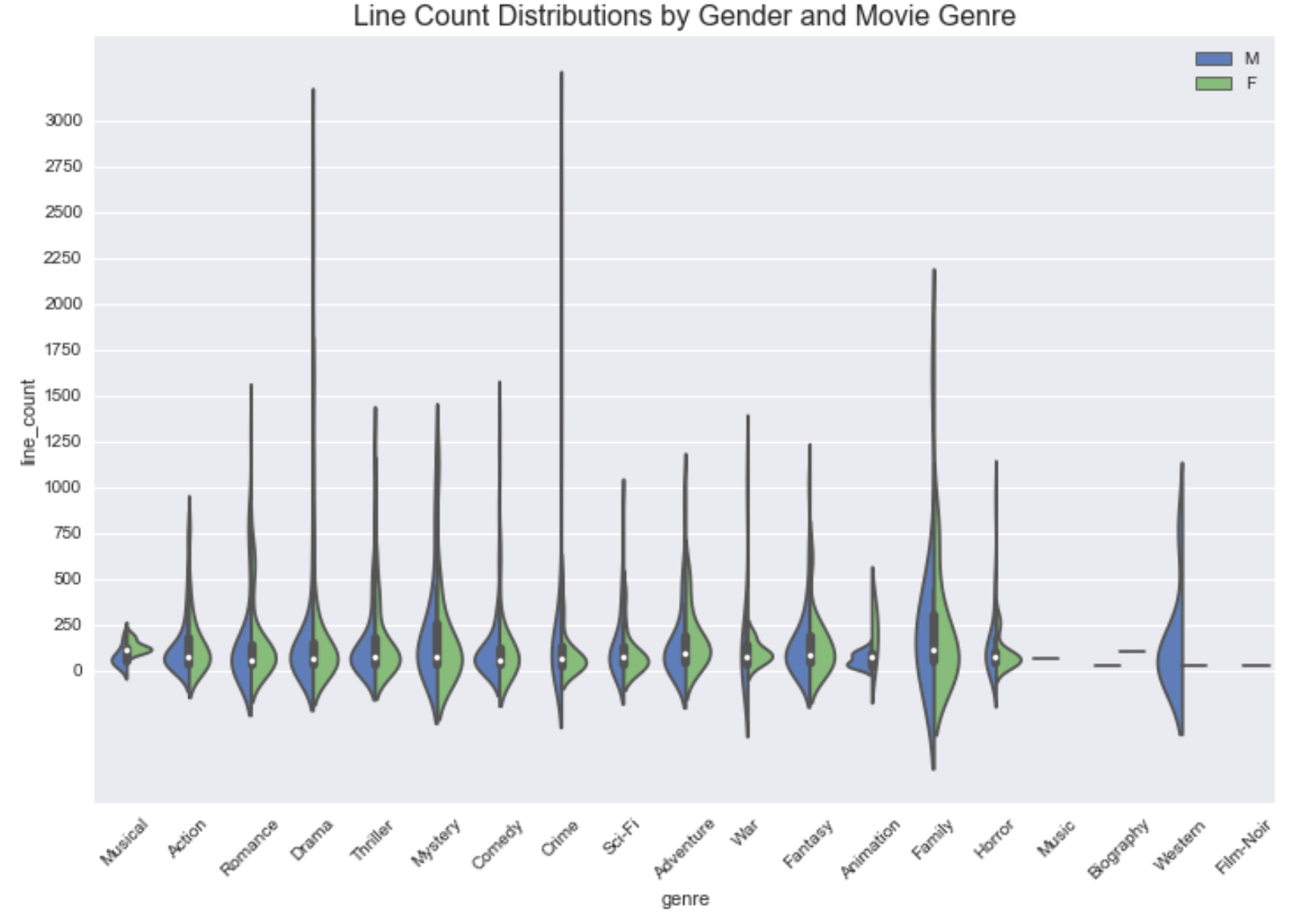
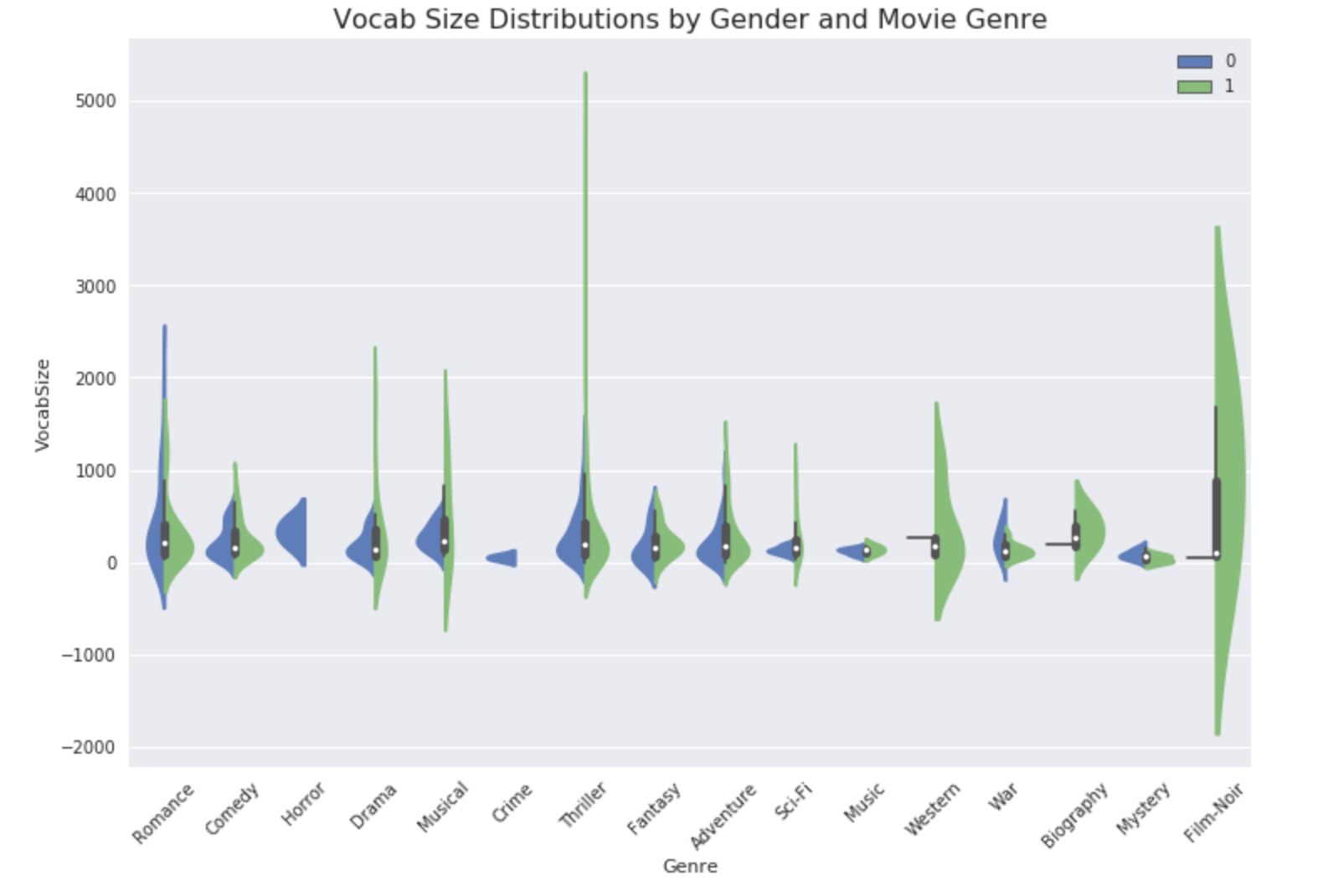
***Results and Discussion***

*Exploratory Analysis*

Before starting to look at any modeling, we first took a look at how our raw data look after we manually classified the genders on the randomly selected data. We found that from the raw data, we were already seeing some gender differences in the data collection.



In an untransformed space, the line count for men and women followed a power law. Women peaked slightly closer to 0 than men, indicating that there was a subtle trend towards women having smaller roles. The data was then transformed to log-log space. We can see that the even in log-space, the line count for male and female actors is quite similar in our dataset. We randomly selected from our main dataset, so the law of large numbers says that our mean for this set will tend to be close the mean in our main set.



We suspect that the genre may show differences between men and women, and we can see from the above violin chart that this may be the case for family, romance, and fantasy, as those datasets appear to have normal distributions for men and women. However, since our dataset was sparse, we did not investigate this further at this time. Because of this, we plan on classifying more data to create a better starting data set.

*Baseline Model*

Our first thought was to see how accurate our model would be by calling every character male. From our random selection, had already seen quite a heavy skew towards male characters, as shown in the exploratory analysis. By assigning “male” to every prediction, we found that this already gave us a 65% accuracy in calling the correct gender of the character, without looking at any contextual data at all. This is already very telling about the wide gap of just the number of characters in scripts that have speaking lines at all.

*Naïve Bayes and Logistic Regression*

Our first model implementation used Naïve Bayes. This was mostly to have a baseline for our study to improve on. Outright, we can see the approach perform poorly on our data. We weren’t expecting good results using this method for vocabulary, and we weren’t surprised to find that this model did not give a much better accuracy than to just call every character a male. We also tried using logistic regressions models to see if we could build a model with better accuracy. We reviewed logistic regression with and without genre on the line count and the vocab size. In general, the models fared very poorly and tended to assume all test subjects were male. The vocab size over the line count however had slightly better results with an accuracy of 70%. Because of these results, perhaps classifying more data as well as getting more contextual information around the vocabulary should help build better models.

*Bag of Words and Neural Nets*

We plan on using a bag of words method to try and get a better model. By looking at lines vocabulary types as well as lines before and after and complexity of the dialogue by calculating the number of syllables of words being used, we may see better results. We also plan on doing a Convolutional Neural Net model and deep learning techniques to develop an algorithm that may be able to successfully determine the gender of a speaker.

***Limitations***

Our sampling method relied on only one draw from our population due to time constraints. We had to manually assign genders to each character, and we have over 20,000 characters in our dataset. We attempted to find a set that included the gender for each character in the IMDB and IMSDB that we could use to enrich our script set, but to no avail. Our statistical inference and therefore and linear model that we could build upon this set was thus limited. Again, all of this information helped will map our next steps which will include: classifying more data, using NLP models such as bag-of-words to analyze the complexity of the actual words used by the characters in their moves, and implementing deep learning techniques such as CNN to gain further insight.

***Resources***

Bechdel, Allison. Bechdel Test. <http://dykestowatchoutfor.com/>

Bechdel Test Movie List: <https://bechdeltest.com>

Kim, Yoon. Convolutional Neural Nets for Sentence Classification. <http://www.aclweb.org/anthology/D14-1181>

Anderson, Hannah and Daniels, Matt. Film Dialogue <https://pudding.cool/2017/03/film-dialogue/>

Geena Davis Foundation. The Reel Truth: Women Aren’t Seen or Heard. https://seejane.org/research-informs-empowers/data/

*Links to data sources*

UC Santa Cruz. Film Corpus 2.0 <https://nlds.soe.ucsc.edu/fc2>

Polygraph’s Film Dialogue Set https://github.com/matthewfdaniels/scripts

*Link to our github*

<https://github.com/mathaholic/w266_final_project>

*Contact the Authors*

[icwang@berkeley.edu](mailto:icwang@berkeley.edu)

[mathaholic@berkeley.edu](mailto:mathaholic@berkeley.edu)

[michelle.m.kim@ischool.berkeley.edu](mailto:michelle.m.kim@ischool.berkeley.edu)