The Presence of Women in the Gutenberg Dataset over a Thirty-Year Period around 1920

Elika Bozorgi, Matthew Pooser, Hao Yang
Department of Computer Science
University of Georgia
Athens, GA 30602

Abstract

In this paper, we addressed the presence of women in the Gutenberg data set over a thirty-year period around 1920. We show how we used many different collections of word counts to establish a correlation between word count and how it relates to the presence and negative attitude expressed towards women. Our data was split into two pieces, a balanced and unbalanced dataset, and then split further into a pre-1920 and post-1920 dataset. Results are shown in tables and pie charts below.

1 Motivation

1.1 Purpose

Although it is often assumed attitudes toward women improved after receiving the right to vote, the Gutenberg data set may prove otherwise. Women still struggled in many areas including employment, sexuality, and individuality. When looking at women from a male's point of view before and after 1920, we hope to obtain a clearer, more in-depth understanding of the attitude towards them, as well as their overall presence.

1.2 Research Question

How can we analyze the presence of women in the Gutenberg data set over a thirty-year period during the women's suffrage movement?

1.3 Hypothesis

If we look at works of literature before and after the women's suffrage movement in 1920, then we will find an overall improved attitude towards women and their roles in novels.

1.4 Solution

Our solution was to narrow down the Gutenberg data set to primarily American and British white male authors in a thirty-year period during the Women's suffrage movement. We divided our data set into two sub-components: literary texts written before 1920 and after 1920.

We then applied a bag-of-words model to look for trends in negative terms or phrases regarding women to see if they are positive, neutral, or negative while also analyzing trends in respect to women's roles, specific women's names, and specific terms in reference to women during the time period. Our domain experts provided us a sample set of words which we used to construct a toy-like bag-of-words model to test our predictions. We settled on using three sets of bag-of-words for negative terms, women's presence, and women's roles.

After pre-processing the datasets, we compared the results for both time periods and outlined them below. We assume that there will be an overall decrease in the amount of negative words in books with significant presence and women's roles. In addition to the bag-of-words, we also wanted to use NLTK's Vader to determine just how much polarity exists between the datasets instead of using simple word counts. Unfortunately, we ran into complications with Vader.

2 Project Background

2.1 Women's Suffrage Movement

Women's suffrage is the right of women to vote in elections. The women's suffrage movement began in the late 19th century. Women sought to change voting laws to allow them to vote. To achieve that objective, both national and international efforts are made. Women gained the right to vote in the Isle of Man in 1881. Then major Westerns powers extended voting rights to women successively, including Canada (1917), Britain and Germany (1918), Austria and the Netherlands (1919) and the United States (1920). Notable exceptions in Europe were France, where women could not vote until 1944, Greece (1952), and Switzerland (1971).

Figure 1: Picture of women marching

The women's suffrage movement was a decades-long fight to win the right to vote for women in the United States. It took activists and reformers nearly 100 years to win that right, and the campaign was not easy: Disagreements over strategy threatened to cripple the movement more than once. But on August 18, 1920, the 19th Amendment to the Constitution was finally ratified, enfranchising all American women and declaring for the first time that they, like men, deserve all the rights and responsibilities of citizenship.

It is often assumed that attitudes towards women improved after receiving the right to vote, but the truth may tell a different story. There is no surprise that women female status is improving. However, we can still see that women struggle in many areas including employment, sexuality and individuality nowadays. Women are not treated in some districts. As a remarkable symbol in the history of women fight for equality, the declaration on August 18, 1920 in American means a turning point in some degree. We may easily assume that there would be some difference between before the date and after the date about the status of female. And the difference can be reflected in so many domains, such as employment, salary, sexuality and education.

Novels come from real lives. The analysis of the presence of women in novels over a thirty years period around 1920 is a brilliant idea to reflect the change of attitudes towards women. Although it is subjective, it also reflects the objective realities. Particularly, when looking at women from a male's point of view during these time periods, we will be able to obtain a clearer, more in-depth understanding of the attitude towards them, as well as their overall presence.

2.2 Works of Literature

In order to find whether attitudes towards women and their roles in novels improved, we look at works of literature before and after the women's suffrage movement in 1920. Novels that were published around the timeline, 30 years span before and after women had the right to vote, are picked. We also specifically pick authors who were American and British because the women's rights happened around the same time for them. We want to look at the novels in a holistic way. So we find some specific authors: Algernon Blackwood: The Wolves of God and Incredible Adventures; J. M. Barrie: Peter Pan in Kensington (1902); Alvin Bunin; Edgar Burroughs: Tarzan of the Apes (1912) and The Gods of Mars (1913); G. K. Chesterton: The Innocence of Father Brown (1911) and The Man Who Knew Too Much (1922); Joseph Conrad: The Secret Agent (1907) and Nostrum: A tale of the Seaboard (1904); Gayle Porter Hoskins; Aylmer Maude; Bertrand Russel; George Bernard Shaw ; H.G. Wells; P.G. Wodehouse: Psmith in the City (1910) and Leave it to Psmith (1923); Emile Zola. In addition, we found some authors that has a couple of books. Edgar Burroughs: The chess men of Mars (1922); Tarzan the untamed (1920); The mucker (1921); The outlaw of torn (1927); At the Earth's Core; The Monster Men; The Mad King; Thuvia, Maid of Mars; Pellucidar; The Oakdale Affair; The Land That Time Forgot; Out of Time's Abyss; The Moon Maid; Tarzan and the Terrible; Tarzan and the Golden Lion; The Efficiency Expert.

After we select related authors and their works of literature, we should filter speaking parts of women and find the derogatory/negative terms or phrases regarding women. In addition, we narrow down to men authors. Since looking at women from a male's point of view, we will be able to obtain a clearer, more in-depth understanding of the attitude towards them, as well as their overall presence. We assume man authors would show more respect and less prejudice to women after the women's suffrage movement. The words used to describe women character in novels will change. For example, in the novels published before the women's suffrage movement, the women's roles in society are always housewives, mothers, cooking etc. Words used to describe women character are such specific terms: emotional, hysteria, delusional, beauty, house work/keeper, caretaker, mother, children, youth, marriage, wife, housekeeper, Mister, Sir, depression, irrational, impulsive, Wench, temptress, manners, Flapper (Flap), Moll, speakeasy, Bearcat, wallflower. In contrast, in the novels published after the women's suffrage movement, more women are engaged in business, law, economics, education and doctor, which are dominated by men. And more respectful and judicial words are used to describe women character.

2.3 Women's Suffrage after 1920

Project Gutenberg (PG) is a volunteer effort to digitize and archive cultural works. Digital literature works are needed to assess the presence of women in novels. It is founded in 1971 by American writer Michael S. Hart and is the oldest digital library. Most of the items in its collection are the full texts of public domain books. The Project tries to make these as free as possible, in long-lasting, open formats that can be used on almost any computer. As of 23 June 2018, Project Gutenberg reached 57,000 items in its collection of free eBooks. So we can make full use of PG to get the digital novels of related authors.

3 Input Data

We used the Project Gutenberg subset available on DigiUGA's GitHub. There were 79 books written before 1920 and 27 books written after 1920 before any type of balancing. No female writers are in either dataset, and this was confirmed by our domain experts and simple

Google searches. See below for some examples in no order from both sets of books:

Table 1: Pre-1920 Books

Book	**Author**
The Cause of it All	Leo Tolstoy
The Power of Darkness	Leo Tolstoy
The Food of the Gods…	H. G. Wells
The New Machiavelli	H. G. Wells
Psmith in the City	P.G. Wodehouse
Our Knowledge of the External World	Bertrand Russell
An Essay on the Foundations of Geometry	Bertrand Russell
Mysticism and Logic	Bertrand Russell
Political Ideals	Bertrand Russell
The Trial	Franz Kafka
Three Days in the Village	Leo Tolstoy
The War in the Air	H. G. Wells
The War of the Worlds	H. G. Wells
Why Men Fight	Bertrand Russell
Nana	Emile Zola

Table 2: Post-1920 Books

Book	**Author**
The Wolves of God	Algernon Blackwood
The Man Who Knew too Much	G. K. Chesterton
The Analysis of Mind	Bertrand Russell
The Problem of China	Bertrand Russell
Free Thoughts and Official Propaganda	Bertrand Russell
Leave it to Psmith	P. G. Wodehouse
Jill the Reckless	P. G. Wodehouse
The Adventures of Sally	P. G. Wodehouse
The Chessman of Mars	Edgar Burroughs
Siddhartha	Hermann Hesse
Tarzan and the Golden Lion	Edgar Burroughs
Tarzan the Terrible	Edgar Burroughs
Tarzan the Untamed	Edgar Burroughs
The Girl on the Boat	P. G. Wodehouse
Pellucidar	Edgar Burroughs

These 15 authors will be used in a balanced set and were chosen to demonstrate the most common number of authors between the two datasets.

Some authors showed up in both before and after 1920 books, like P. G. Wodehouse and Bertrand Russell, and were studied more closely to provide deeper insight into how their attitude towards women may have changed throughout the women's suffrage movement.

For sake of brevity, the full list of all books and their respective authors will be uploaded onto this project's GitHub.

4 Pipeline

4.1 Tools

NLTK was used extensively to finish off the preprocessing step (and sentiment analysis) for the project. Google Colab was partially used when memory would run out during program execution. Spyder was the main IDE used in writing up the Python scripts while Jupyter Notebook was used for fast scripts such as plotting the results. Basic Python libraries such as NumPy were also used.

4.2 Procedures

4.2.1 Data Collection

We cloned the DigiUGA's Project Gutenberg repository and copied every single text file in the dataset above. They were further divided into a balanced and unbalanced dataset and passed into preprocessing and word counting.

4.2.2 Preprocessing

We followed a hybrid manual and automatic approach to preprocessing. For the manual part, the header and footer of every book was deleted from the text files.

For the automatic step of preprocessing, we used NLTK to do the usual preprocessing tasks of tokenizing the words, making them all lower-case, stripping punctuation, and removing stop words [4]. We did not use stemming in the pre-processing phase.

4.2.3 Experiment 1: Bag-of-Words

We wrote a script which would calculate the total word counts using Python's collections Counter if there was a match in the presence, roles, and negative dictionaries. The results have been printed in Tables 3-6. See section 5.3 for examples of words in these dictionaries.

5 Results

5.1 Bag-of-Words Generated from Balanced Dataset

Table 3: Balanced Pre-1920 Counts

Bag	Count
Roles/occupations	1170
Presence	2615
Negative	92

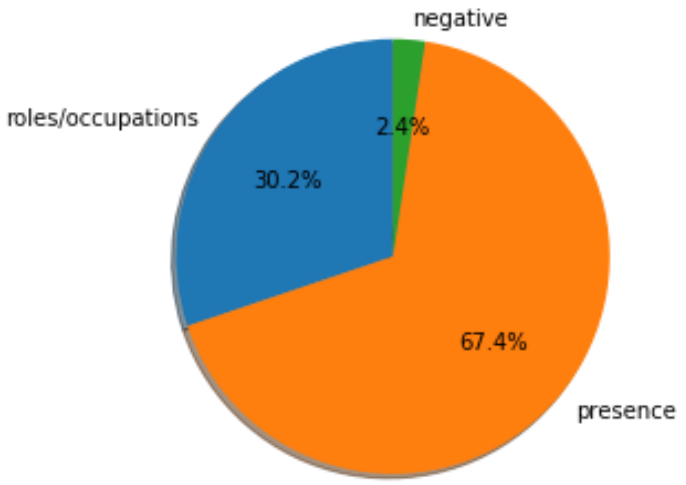

Figure 2: Relative Frequencies of Balanced Pre-1920 Counts from Table 3's numbers

Table 4: Balanced Post-1920 Counts

Bag	Count
Roles/occupations	1007
Presence	4041
Negative	98

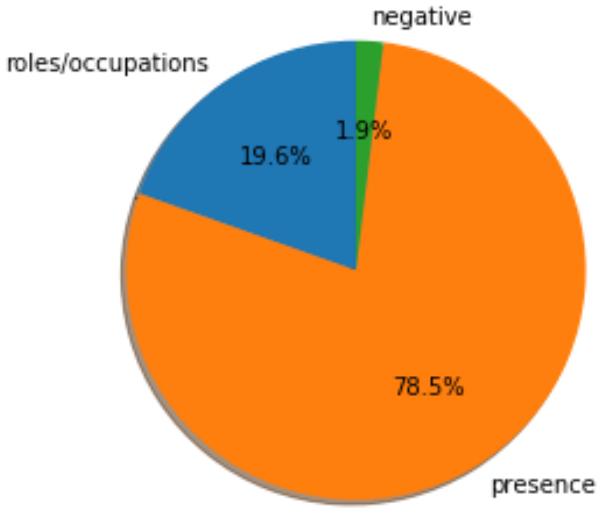

Figure 3: Relative Frequencies of Balanced Post-1920 Counts from Table 4's numbers

5.2 Bag-of-Words Generated from Unbalanced Dataset

Table 5: Unbalanced Pre-1920 Counts

Bag	Count
Roles/occupations	8132
Presence	21479
Negative	893

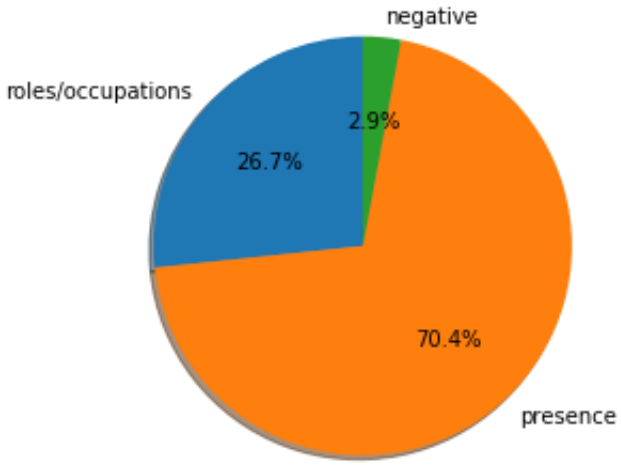

Figure 4: Relative Frequencies of Unbalanced Pre-1920 Counts from Table 5's counts

Table 6: Unbalanced Post-1920 Counts

Bag	**Count**
Roles/occupations	1508
Presence	6937
Negative	152

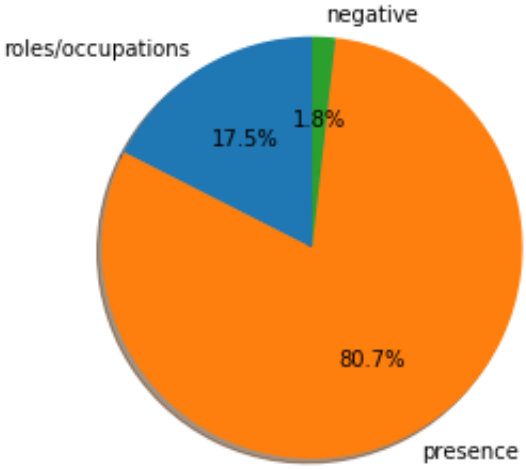

Figure 5: Relative Frequencies of Unbalanced Post-1920 Counts from Table 6's counts

5.3 Dictionaries

Negative words: mistress, virago, frump, harridan, spinster, wench, tomboy, flapper, witch, looker, kitten, and other expletive words

Roles/Occupations: housekeeper, caretaker, house, politics, nurse, cook, suffrage, strike,

war, reform, vote, factory, teacher, school, maid, secretary, sewer, etc.

Presence words: flapper, daughter, madame, mother, girl, may, mary, anne, elizabeth, catherine, beth, sarah, eliza, etc.

See `wordcounts.py` for the full list of words for these three dictionaries.

5.4 Problems

Our laptops at first took a significant amount of RAM to run these Python scripts. Overtime, my laptop seemed to be able to run all the scripts fine within a reasonable amount of time however.

NLTK's Vader did not like our data set, and we had to drop it from the write-up unfortunately. See Section 6 for Future Work involving NTLK's Vader

5.5 Insights

Looking at the balanced data set results for the bag-of-words models, the relative frequency of the negative words used from before 1920 to after 1920 dropped from 2.4% to 1.9% respectively. This may be insignificant, but nevertheless it shows a marginal decrease in the amount of negative words used in the balanced corpora. Additionally, presence increased substantially from 67.4% to 78.5% which shows some correlation between the data and these presence words, suggesting women are incredibly present in these set of books. However, the presence dictionary contains the word 'may' which may be affected if we used stemmed words in our pre-processing stage.

These results are mostly populated by authors who have books from both time periods: pre-1920 and post-1920. These white male authors increased the amount of discussion for women while reducing the relative frequency of negative terms found in the corpora. On the flip side, the amount of pre-1920 occupational roles dropped from 30.2% to 19.6%. This suggests that the amount of occupational job discussion dropped since women retained their jobs and moved away from conventional roles.

The results from the unbalanced corpora show similar trends with all three groups but with much lower numbers. For instance, presence had a total count of 21,479 presence words in the unbalanced pre-1920 corpora but the balanced pre-1920 corpora had a total of 2,615 words. We still included it for any curious inquiry into our unbalanced corpora.

6 Future Work

Our future work would involve using NLTK's Vader to get a better accuracy for the intensity of the defined negative words instead of simply relying on word counts. We also plan to train a LSTM on our data so we can give it a prompt to see what the model would spit out in response.

7 References

[1] Women's suffrage. `en.wikipedia.org/wiki/Women%27s_suffrage`

[2] Woman Suffrage Movement. `womenshistory.org/resources/general/woman-suffrage-movement`

[3] Project Gutenberg. `en.wikipedia.org/wiki/Project_Gutenberg`

[4] Brownlee, Jason. (2017) How to Clean Text for Machine Learning with Python. `machinelearningmastery.com/clean-text-machine-learning-python`