Biased Blogging: Linguistic Evidence of Gender Bias in Political Blogs

Erika Kirgios
Adviser: Christiane Fellbaum

Abstract

Blog posts are becoming an increasingly relevant way of gaining information about political candidates. While blogs may be more candid and impassioned than news articles, they also tend to be more biased. Due to the presence of two female front-runners in the 2016 election, it is relevant and important to probe the ***linguistic evidence of gender bias in political blogs****. We aim to assess whether female candidates are portrayed differently than their male counterparts in political blogs. The hope is that this work will make individuals more cognizant of potential biases in their sources of information.*

1. Introduction

The current presidential election is more gender focused than any election before as two of the primary candidates from each party, Carly Fiorina and Hillary Clinton, are female. While newspaper sources try to write from as unbiased a viewpoint as possible, bloggers tend to be more candid about their opinions on future candidates. With the prevalence of social media and trending blogs as a news source for many people today, especially those of the younger generation, it is important to determine whether these sources present gender bias and to analyze how such bias may skew the opinions they present. After all, young adults were one of the biggest demographics to contribute to President Obama's 2008 election. Furthermore, such unfiltered information sources are optimal for sentiment analysis and computational linguistics since they explicitly state opinions rather than restricting themselves to facts. As such, the goal of this project is to analyze the linguistic content of such blogs to identify how language changes based on which candidate is described and to identify key features that may indicate gender bias.

2. Background and Related Work

Psychological and linguistic principles underlie and motivate this study. The theory of benevolent and hostile sexism penned by Peter Glick and Susan Fiske separated gender bias into two categories: one that uses subjectively positive language and tone to describe women but confines them to traditional roles (benevolent) and a second form that is more explicitly negative (hostile) towards women. Since traditional womanhood does not associate femininity with political prowess or leadership, we theorize that posts that demonstrate a high amount of gendered language, both positive and neutral, may be correlated with an overall negative sentiment towards female politicians [8].

Furthermore, the language used when discussing a female candidate may be different than that used to describe a male candidate. This is called marked language, or language that changes when applied to different categories, here female as opposed to male (like actress vs. actor); marked language also describes things that are out of the norm [4]. This change extends beyond just the modification of individual words. The maleness of a male candidate is not surprising; rather, it is expected. Meanwhile, the idea of a female president after generations of male presidents is outside of the norm, priming one to think about gender when writing about a female candidate. As such, we expect to see a higher occurrence of gendered language (both masculine and feminine) in blogs about female candidates than in blogs about male candidates.

In fact, this tendency to discuss successful females in terms of their womanhood, emphasizing their gender, is discussed by science writer Ann Finkbeiner [6]. Her writing was used to create the Finkbeiner Test - to be successful and female is to have your gender and family life emphasized when you are written about, and the test measures the degree to which this occurs [5]. Meanwhile, maleness is less likely to be mentioned in articles about men because it is the expected gender of a leader.

The linguistic patterns used to discuss females also differ from those used to discuss males in terms of warmth and competence. Fiske, Cuddy, and Glick found that the two most salient

dimensions of social judgment of another individual or group are warmth and competence [7]. Different groups fall differently on these quadrants, with those being seen as high warmth, low competence being regarded affectionately; while those who are high competence, low warmth are seen with respect but not affection, and so on [7]. It has been hypothesized that working mothers are seen as high warmth, low competence while female leaders are seen as high competence, low warmth [3]. This paper endeavors to see whether such patterns are evidenced in political blogs.

Computational linguistics has been used to probe gender inequality on Wikipedia, but not on political blogs [12]. Wagner et al examined corpora composed of Wikipedia articles about notable people in several different languages. The dimensions they focused on were: the relative coverage of males and females in Wiki pages, structural bias in the ways male and female pages linked to each other, lexical bias in terms of word frequencies, and visibility bias in terms of how much these articles were promoted by Wikipedia [12]. They found little coverage and visibility bias, but did find structural bias and strong lexical bias - in other words, women were written about quite differently than men were [12]. We are primarily examining lexical bias and hypothesize that we will find the same pattern in political blogs.

This research will also be done in the context of sentiment analysis, a field that has been expanding rapidly in the last decade. The goal of the field is to offer a summary of the opinions related in a text, often assigning an opinion to the overall text. This is often done through Feature-Based Summarization, a technique implemented by Bing Liu which involves pinpointing the features of an object or person under discussion, then analyzing the polarity of the adjectives used to describe those features and the language surrounding those adjectives [10]. Other researchers have focused on creating tools for sentiment analysis, such as SentiWordNet, which uses crowdsourcing to give every synset of WordNet a sentiment score - similar to what we call a polarity score, or positivity/negativity score, in this paper [1].

While our research does not give an overall bias score to political blogs, it does examine frequencies, bias level, and polarity of different words used in these blogs in order to find patterns that might indicate bias. Unlike SentiWordNet, we were not able to give a huge set of words bias

scores - rather, we restricted such scoring to a set of target words. Current research in the field has also moved beyond focus on product reviews to the analysis of opinions presented on social media such as blogs and Twitter, claiming that the combination of unfiltered opinions and readability make these sources both interesting and impactful [9, 11]. However, few have applied these techniques in order to assess the levels of gender bias in a text.

Thus, this paper is unique in its goal: using natural language processing to find linguistic evidence of gender inequality in political blogs.

3. Approach

In order to assess linguistic evidence of gender bias, we first must choose a dataset. We determined to choose blogs evenly across four candidates: a female Republican (Carly Fiorina), a male Republican (Jeb Bush), a female Democrat (Hillary Clinton), and a male Democrat (Bernie Sanders). We wanted to find a lexical set that could help us evaluate the linguistic patterns used to discuss each gender; to do so, we decided to select 51 of the most common adjectives and nouns in the blogs and determine their polarity and their gender bias. Gender biased terms are defined here as adjectives or nouns predominantly applied to one gender rather than the other.

We developed a gradient for gender bias and polarity based on Amazon Mechanical Turk surveys, in which individuals were asked to categorize words on a gender scale or polarity scale. These results were used to put words on a spectrum from -1 to 1 for polarity (with -1 being negative, 1 being positive) and -1 to 1 for gender bias (with -1 being strongly feminine, 1 being strongly masculine). Furthermore, we looked at overlapping categories of words (i.e. positive feminine words, such as *loving*; negative feminine words, such as *shrill*; and neutral feminine words, such as *girly*).

Furthermore, we also decided to run smaller independent analyses of the categories outlined in the Finkbeiner test: words describing gender (i.e. *woman*, *man*) and words describing family or relationships (i.e. father, *mother*, *wife*, *husband*) [5].

4. Data and Methods

4.1. Blog Selection

We wanted the chosen dataset to be distributed across the political and professional spectrum, with a balance of blogs from conservative, liberal, well-established, and small independent blogs. Small independent blogs were chosen from Wordpress, while well-established blogs were chosen from, for example, The New York Times blog, The Washington Post blog, and so on. In order to determine which were the most popular liberal and conservative blogs, we consolidated the top choices from several online lists and searched for the candidates' names within those forums. Examples of liberal blogs used were HuffingtonPost, Daily Kos, and Mother Jones. Examples of conservative blogs were Hot Air, The Foundry, and Gateway Pundit.

On average, these blogs contained 795 words. Those that went beyond 1000 were cut (always removing from the bottom for consistency) and those that were shorter than 600 were not used.

4.2. Word Selection

Once the blogs were selected, we removed hyperlinks from their contents and ran frequency analyses of the words in the blogs - ensuring first that they were all lowercase so that the same word, or Token, with different capitalization would not be separated into two different frequency counts. Of the most 1500 most frequently used words, we removed all stop words, such as "the" and "a." We also "stemmed" the Tokens - combining plurals, such as "woman" and "women." We did not stem verbs or adverbs since we were looking only at nouns and adjectives.

Though these analyses were done on a by-candidate basis, the word selection was conducted from the most frequently used words overall such that the chosen words would be more likely to occur frequently for each of the candidates' blogs. The words were selected from the most frequently used stemmed adjectives and nouns, paying particular attention to words with strong valences.

Ultimately, we arrived at a list of 51 words: 30 adjectives and 21 nouns. We calculated their frequencies on a by-candidate basis.

4.3. Word Categorization Survey

In order to give the selected words gender and polarity scores as described above, we needed a group of individuals to rate each word on its level of gender bias and positivity/negativity. These individuals were found from Amazon Mechanical Turk (MTurk), which is a platform through which businesses and developers can put up short tasks, or HITs, and an associated monetary compensation. Workers then find the tasks and take them.

Since the HIT was to be a survey, we used Qualtrics, a survey platform, to create our questionnaires. Qualtrics can be linked to MTurk by adding a random number generator to the end of the survey flow, which creates a unique, random MTurk code for each survey taker. This embedded data is stored in Qualtrics, and the user must copy and paste it into MTurk such that the survey administrator can match the embedded data in Qualtrics with each worker's MTurk worker ID and the code they provided. Those that are verified will then be compensated.

We imagined that raters would be most reliable if they had fewer words to score, but it would cost more to have surveys with fewer words. To balance these factors, we determined that each survey would require workers to give ratings to 10 of the words (one of them would require 11) and that we would have 12 workers per word. Since the words were not obscure, we thought 12 would be a large enough sample of ratings. Overall, we created 10 of these surveys: 5 for gender ratings, and 5 for polarity ratings. We decided to separate these so that workers would not conflate the two. We also restricted workers to responding to at most 3 HITs in order to ensure that the results were not skewed by one individual.

To make the surveys themselves as unbiased as possible, we prefaced the gender bias questions with the following statement:

> In this questionnaire, we are not asking you to describe your personal beliefs about the given words; rather, we would like you to think about the way these words are generally used.
>
> If you understand, select "No". Else select "Yes"

We hoped that this statement would both prevent users from changing their answers out of fear of appearing biased and act as an attention control at the beginning of the survey. We then randomized the order of appearance of each question. Each question read "Is this word used most frequently to describe males, females, or both equally?" followed by the target word. The options were on a scale of 1 to 3 to reduce noise and because the question content was conducive to a three-way choice.

The polarity questions also had an initial attention section which simply stated that we wanted to evaluate whether the following words were positive or negative. The questions themselves were "Would you most likely want to be friends with someone described/labeled in this way?" These were evaluated on a 1 to 5 scale because polarity might be easier to evaluate than gender score and there is a more clear middle ground between "neutral" and "strongly negative/positive."

Once the results were collected, we gave each word a score from -1 to 1 for gender bias and polarity by averaging the scores that each word received. When analyzing our data, we split our words into Set 1 and Set 2. Set 1 is the set of all 51 words, while Set 2 is the set of words without factual gender identifiers, such as *man* and *woman*. We did so for two reasons: first, because there was a much higher frequency of factual gender words than any of the other words, which made some of the graphs confusing, particularly because these words were evaluated as having a positive valence; second, because we wanted to see how removing such words would change the frequency distributions. According to Finkbeiner's hypothesis, female leaders' gender should be more salient and therefore mentioned more frequently. Thus, if her hypothesis is correct, removing factual gender words should change the distributions more drastically for blogs about females than blogs about males.

We also found overlapping categories of words, i.e. feminine-negative, feminine-positive, and so on. We did so by finding the mean and standard deviation for both gender scores and positivity/negativity scores. Words less than half a standard deviation away from the mean were considered neutral; between half a standard deviation and a full standard deviation were considered neutral-positive or neutral-negative, and so on; over a standard deviation away from the mean were considered feminine, masculine, negative, or positive. Thus, we had a total of 25 categories, and

found which of these each word fit into (note that several categories were empty).

4.4. Article Categorization Survey

In order to evaluate our results, we wanted to run the same analyses on blogs that had been given a gender bias score and see how well the differential linguistic patterns identified correlated to these assigned scores. To do so, we used a random number generator to select 5 numbers between 0 and 40 for each candidate. These numbers were used to index into an alphabetical list of the blogs used for each candidate, and those 20 blogs were selected to be rated.

As for the word ratings, we used Qualtrics to create a survey that was linked to MTurk. Since these gender bias ratings might be more subjective or prone to error if one does not fully understand what gender bias is, we had 21 respondents per article. The survey flow was created such that for each survey, only four articles of the 20 would appear at random (in order to keep the workers from getting bored and to ensure they would pay attention), and each article would be given to the same number of survey takers.

For each article, we asked "How gender biased does this article seem?"; "Does this article have a positive, negative, or neutral opinion about the candidate?"; and "How frequently was gender stereotypic language used in this article?" The first question was on a scale of 1-5, while the second and third were on scales of 1-3. The articles were prefaced by an attention question, a formal definition of gender bias and of gender stereotypic language, and paragraphs from a strongly biased blog and an unbiased blog. These will be included in the appendix.

4.5. Other Evaluations

We also examined the by-gender likelihood of mentioning specific categories of words discussed by Finkbeiner: for example, frequency of appearance of spouses' names, words about family, and words about gender. These probabilities were calculated with

$$P(wordcategory|g_1) = \frac{P(wordcategory, g_1)}{P(wordcategory)}$$

Here, g_1 is the gender being assessed, and $P(wordcategory, g_1)$ is the joint probability of seeing a word from the given category and seeing a blog of gender g_1. Meanwhile, $P(wordcategory)$ is the overall probability of seeing the word in the blogs.

5. Results

5.1. Overall Frequencies

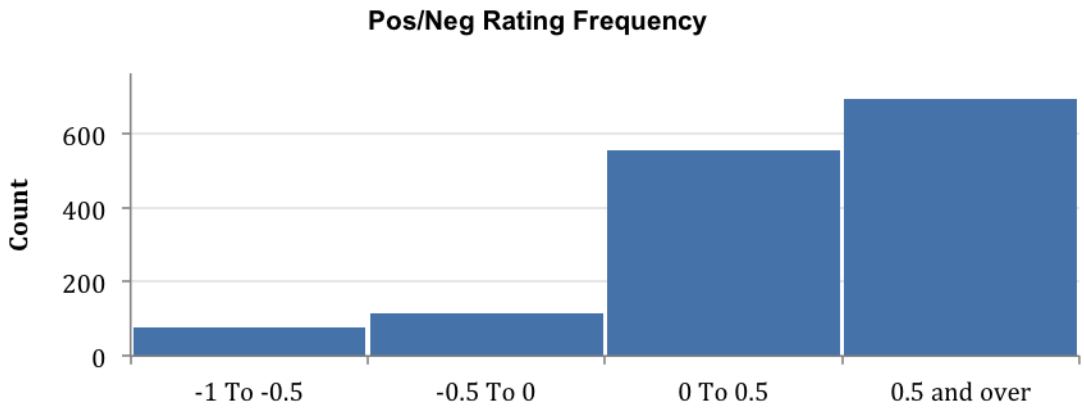

Figure 1: A histogram depicting the frequency with which words of given polarity scores were used across all 160 blogs

In Figure 1, words were grouped in bins of size .5, with each bin holding the number of times words within that rating range were used. Thus, the first bin represents the total frequency across the 160 blogs of words that received ratings from [-1, -.5). The graph demonstrates that words with positive connotations were used with much greater frequency than words with negative connotations.

Figure 2 groups gender ratings into bins of size .5. Again, the frequencies represented here were taken across all 160 blogs. Thus, each bin holds the total frequency of words assigned gender scores between [a, b). The graph suggests that strongly feminine language was used most frequently, then strongly masculine language, with more neutral language being least frequently used. Note that the high counts of feminine words could be attributable to the high frequency of the use of the words *woman*, *female*, and *Mrs./Ms.* – which happened particularly often in the blogs about females.

This is evidenced through comparison to Figure 3, in which only the frequencies of Set 2 words

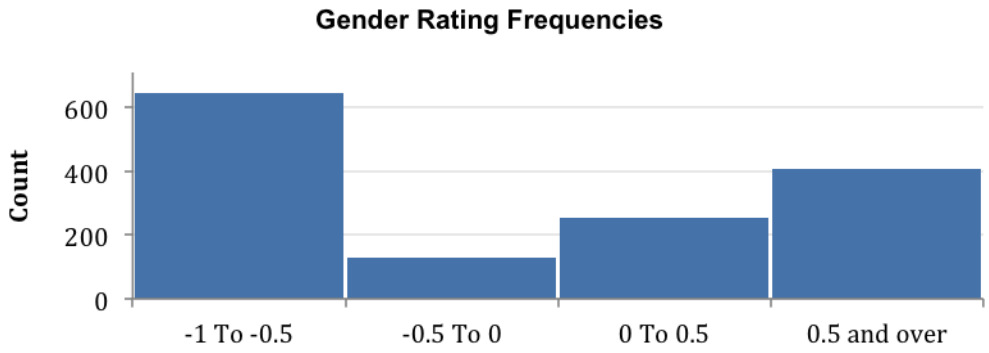

Figure 2: A histogram depicting the frequency with which words of given gender scores were used across all 160 blogs

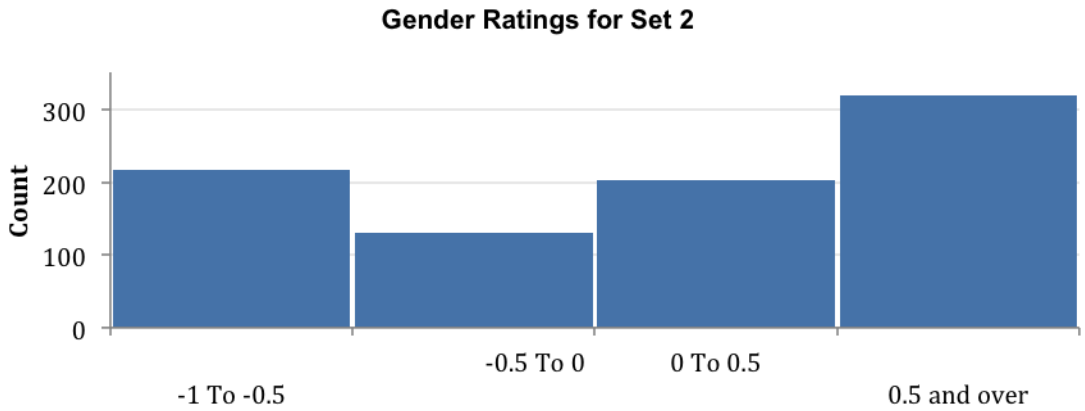

Figure 3: A histogram depicting the frequency with which words of given gender scores in Set 2 were used across all 160 blogs

is depicted. Recall that Set 2 includes all Tokens except those that factually indicate gender, such as *woman* or *man*. Figure 3 demonstrates that the words used in political blogs are more frequently strongly masculine and are neutral-masculine about as often as they are strongly feminine once factual gender words are removed. In the next sequence of graphs, we see that such sex-specifying words were more frequently used in blogs about Clinton or Fiorina, supporting our hypothesis that their gender would be more frequently mentioned than that of male candidates because it is out of the norm.

5.2. Gender Rating Frequencies by Candidate Gender

Comparing Figures 4.1 and 4.2, we see that for the female candidates, the vast majority of strongly feminine words used simply denoted their gender.

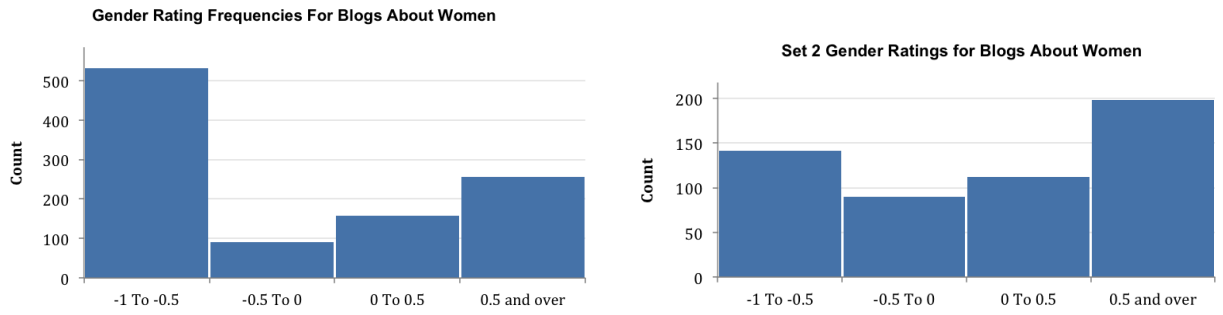

Figure 4: Figure 4.1 on the left is a histogram depicting the frequency with which words of given gender scores appeared across the 80 blogs about women. Figure 4.2 is a histogram that shows the same frequencies for Set 2 words only

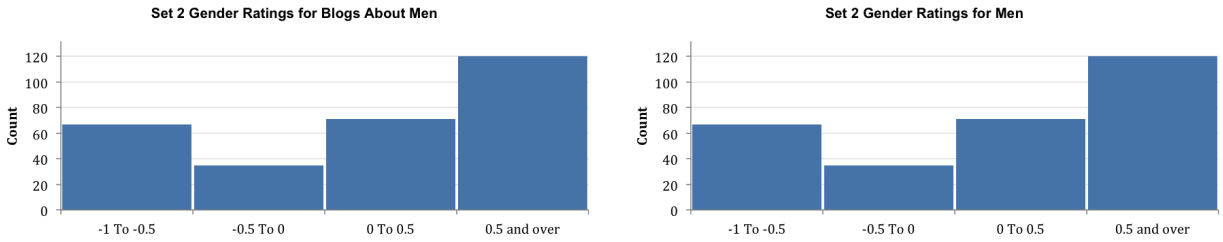

Figure 5: Figure 5.1 on the left is a histogram depicting the frequency with which words of given gender scores appeared across the 80 blogs about men. Figure 5.2 is a histogram that shows the same frequencies for Set 2 words only

Figure 5.1 shows that strongly masculine words are used most commonly in blogs about male candidates - the opposite was true for Figure 4.1, the frequency distribution across blogs about women. Meanwhile, Figure 4.2 and 5.2 are much more similar in distribution. As hypothesized, the change in frequency distribution for blogs about women between all words and Set 2 words was far more drastic than that in blogs about men. This supports the hypothesis that the gender of female candidates is referred to more frequently than that of male candidates, perhaps because it is more surprising to see a female running for president. When factual gender words were removed, strongly masculine words were the most commonly used across the board. This suggests that the language used in political blogs is most frequently gendered towards the masculine, perhaps because words describing strong political candidates - such as *prominent* or *formidable* - were considered by our raters to be masculine words.

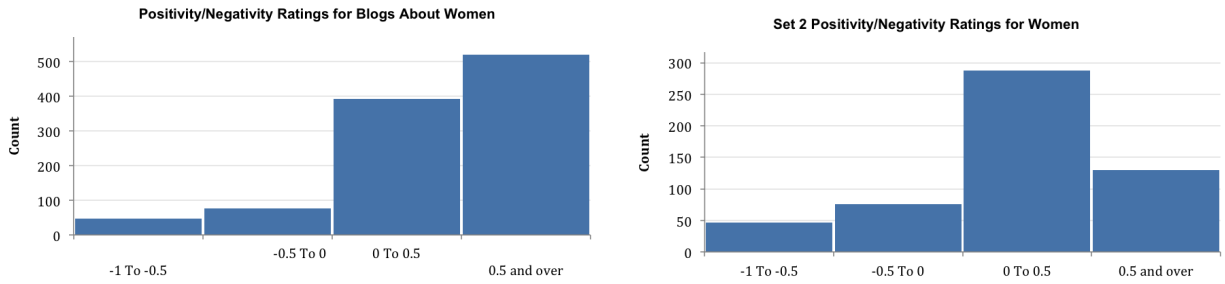

Figure 6: Figure 6.1 on the left is a histogram depicting the frequency with which words of given positivity/negativity scores appeared across the 80 blogs about women. Figure 6.2 is a histogram that shows the same frequencies for Set 2 words only

5.3. Positivity/Negativity Rating Frequencies by Candidate Gender

In Figure 6.1, we see that strongly positive words were used most frequently in blogs about women, with neutral-positive words following. The gap between neutral-positive and neutral-negative word frequencies is quite large. However, this information may be skewed by the fact that our raters rated words as positive that may normally be considered neutral, such as *woman*. In fact, when comparing Figure 6.1 to Figure 6.2, we see that strongly positive words are no longer used very frequently - likely because the feminine factual gender words were given scores over .5.

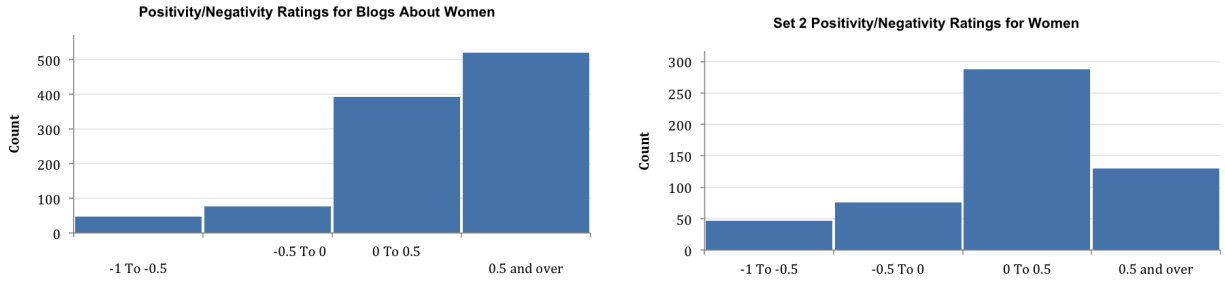

Figure 7: Figure 7.1 on the left is a histogram depicting the frequency with which words of given positivity/negativity scores appeared across the 80 blogs about men. Figure 7.2 is a histogram that shows the same frequencies for Set 2 words only

Figure 7.1 shows a similar distribution to Figure 6.1; however, the difference between strongly positive and neutral-positive words is less pronounced. When comparing Figure 7.2 and 6.2, we see that with factual gender words removed, the language used to describe men had a higher proportion of strongly positive words to words of other valences than that of women. Furthermore, the change in distribution between Figures 7.1 and 7.2 was, again, less pronounced than that between the

distributions in Figures 6.1 and 6.2 due to the high frequency of factual gender words in blogs about females.

5.4. Gender Score on Positivity Negativity Score

Figure 8 examines the location of our 51 words on four quadrants, with the upper-left quadrant being negative masculine words, the upper-right being positive masculine words, the lower-left being positive feminine words, and the lower-right being positive feminine words. Note that words closer to the origin and the axes are neutral on one or both dimensions.

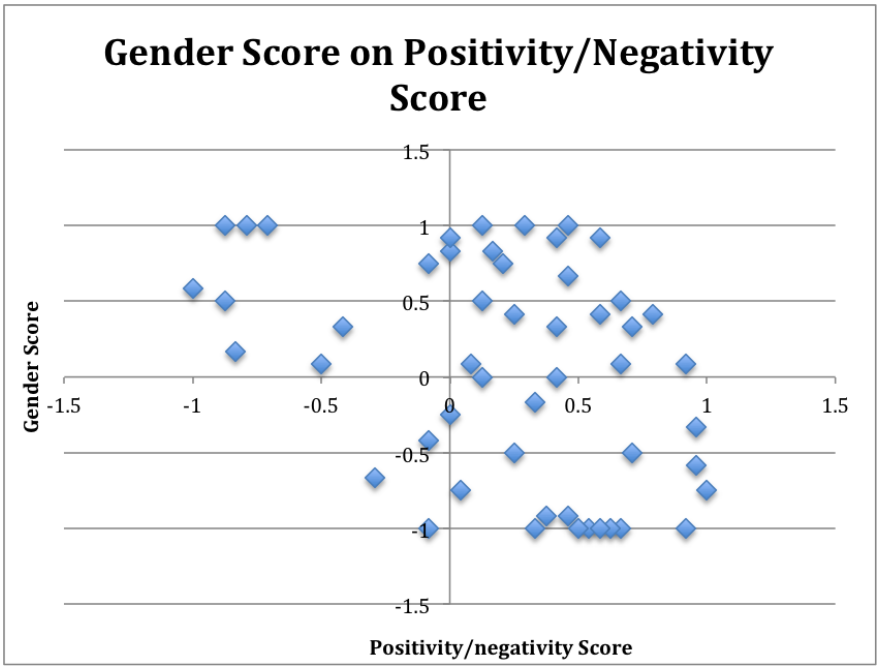

Figure 8: The x-axis represents polarity while the y-axis represents gender bias. Each of the 51 words is plotted on these two dimensions

Note that the 51 words were most often considered positive, and that the lower-left quadrant is particularly sparse. None of the words in that quadrant are particularly far from the y-axis, which implies that for the most part, they are neutral rather than negative. In fact, none of the words that scored below -.5 on the valence scale were gendered feminine. This pattern will be explored more closely in the next section.

Figure 9 again plots the words on the gender-polarity quadrants; however, here each different bullet style represents a different candidate. Furthermore, we found the median frequency of the 51

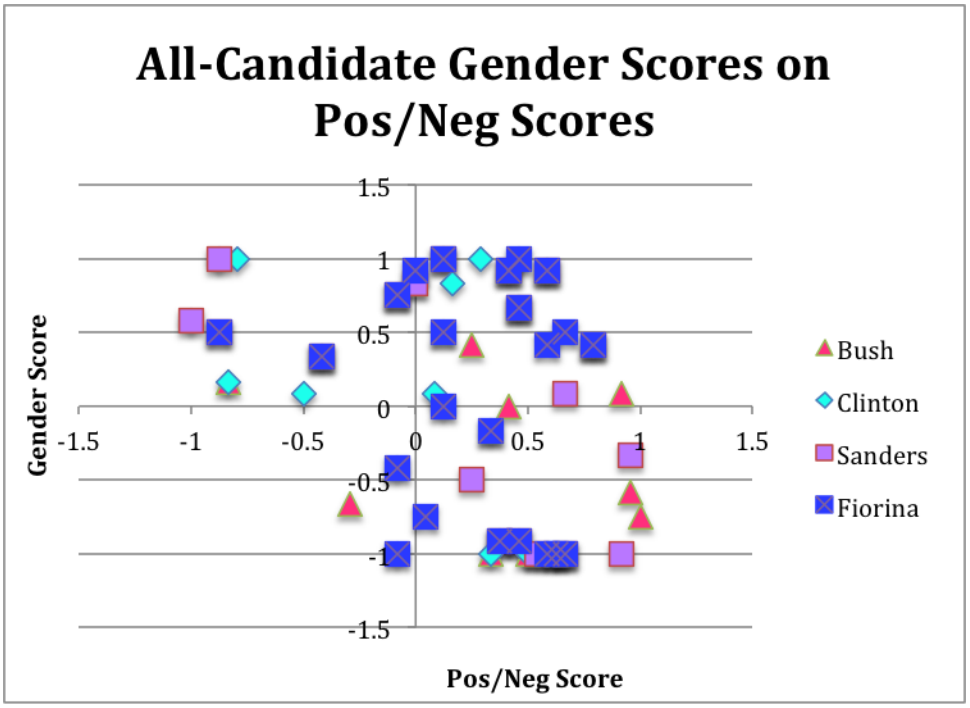

Figure 9: The x-axis represents polarity while the y-axis represents gender bias. Each of the 51 words is plotted on these two dimensions, with bullets of different shapes and colors representing different candidates.

words for each candidate, and only words used more frequently than the median were plotted for each candidate. Here we see that many more words on the neutral-positive end of the spectrum are used for all candidates. A higher frequency of neutral-masculine or masculine words were used for all candidates.

5.5. Overlapping Word Categories

Feminine Positive words were the most commonly used Feminine words, though this tendency was less exaggerated when looking only at the Set 2 words. In both cases, no Feminine Negative words appeared. This is likely because the words *woman* and *female* were considered Positive. Interestingly, the Feminine Positive words were, for the most part, terms such as *friendly* or *loving* - words that suggest warmth rather than competence.

There is a strong difference between Figure 11 and Figure 10. The frequency distribution of Masculine word types for both all words and those in Set 2 is concentrated in the Masculine-Neutral category, followed by the Masculine Neutral-Positive category. When considering only the Set 2

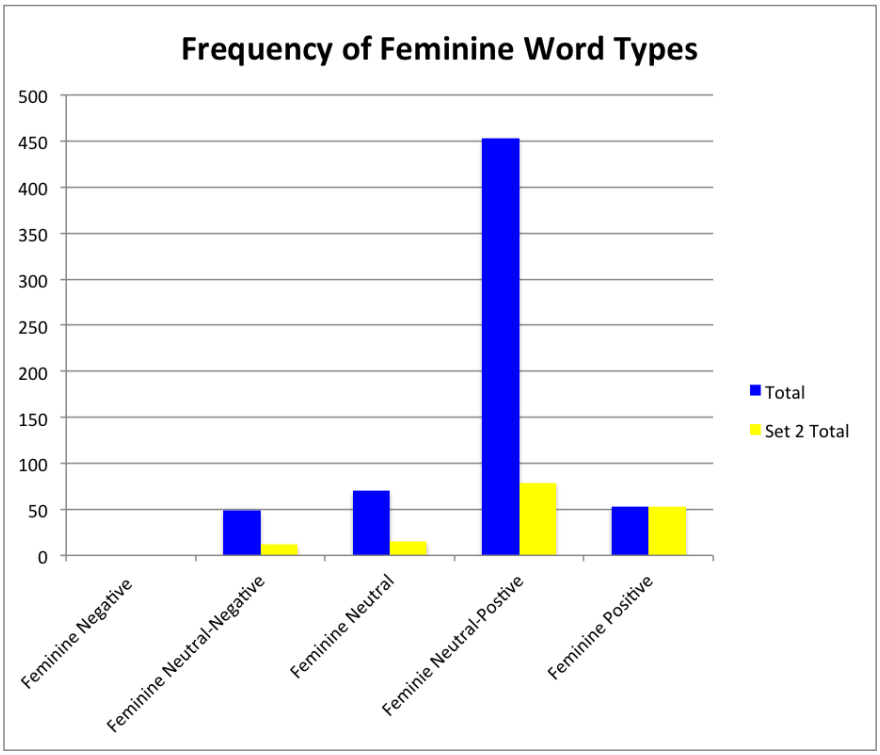

Figure 10: The frequency across all blogs of words in each of the overlapping Feminine categories (i.e. Feminine Neutral, Feminine Neutral-Positive, etc) is shown across all Tokens and those in Set 2.

words, there are no Masculine Positive words used at all, marking a strong contrast between the distribution of Masculine and Feminine words.

Interestingly, words in the Masculine Neutral category included words such as *chief/boss*, *powerful*, and *formidable*. These words all carry the connotation of competence, but not of warmth. Thus we can clearly see that our strongly feminine target words are high on warmth and low on competence, while the strongly masculine target words were high on competence.

In the following section, we examine the by-candidate-breakdown of frequency of words in each overlapping category.

5.6. Overlapping Category Frequencies By Candidate

Figure 12 shows that Fiorina and Clinton's blogs have the greatest combined frequencies of all the Feminine categories by a quite a large margin. In other words, the vast majority of the strongly female-gendered language was used to describe females, and a very small proportion was used to

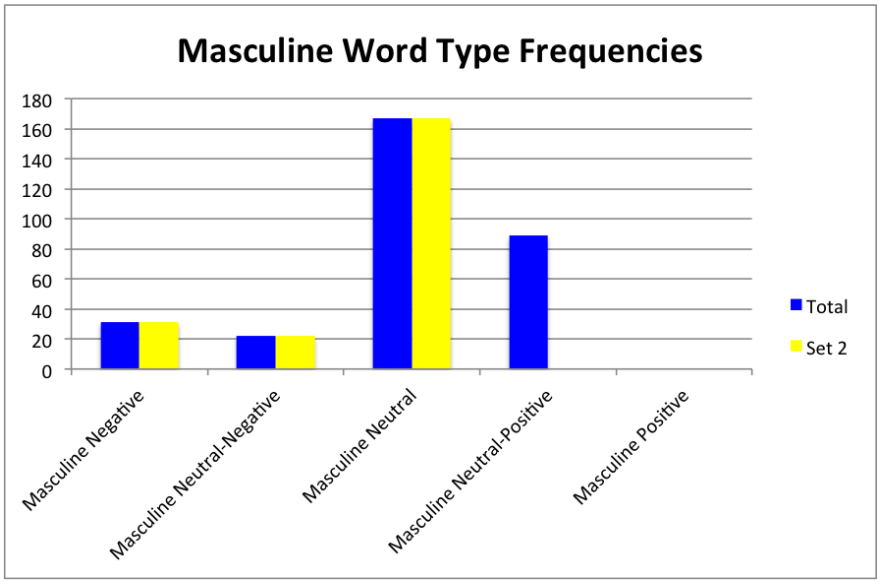

Figure 11: The frequency across all blogs of words in each of the overlapping Masculine categories (i.e. Masculine Neutral, Masculine Neutral-Positive, etc) is shown across all Tokens and those in Set 2.

describe males.

Meanwhile, Figure 13 shows a bit more variability. Nearly all categories of gender neutral target words were used more frequently to describe females than males, but the Neutral Positive category was dominated by the males. In the Appendix, the Neutral Feminine and Neutral Masculine distributions can be found - interestingly enough, these are the categories in which the words were used more often in blogs about male candidates than about female candidates.

The distribution of frequencies of Masculine word types is shown in Figure 14. For all categories, there is a higher occurrence of the terms in the feminine blogs than in the masculine blogs. Interestingly, this includes the Masculine Positive category, which we saw in the previous section was composed only of factual gender terms, such as *man*. This seems to support our initial hypothesis that mentions of gender and language on both extremes of the gender spectrum would occur most frequently in blogs about female candidates.

5.7. Category Analysis

Based on the Finkbeiner Test, we created three categories of words: spouse mentions, relationship words, and gender words. Spouse mentions is straightforward: it is simply the frequency with which

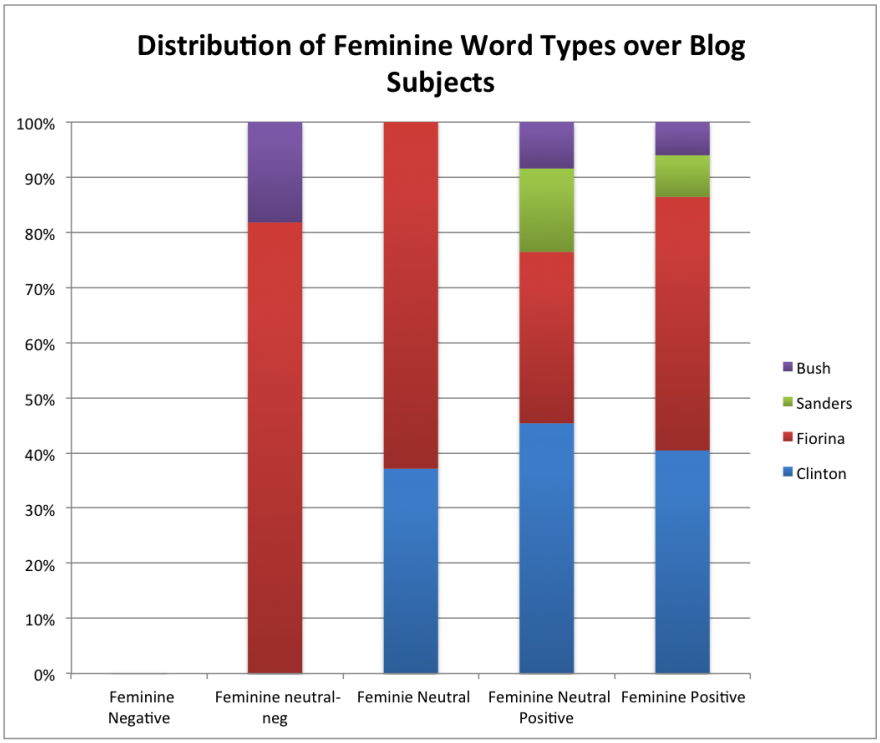

Figure 12: For each of the overlapping Feminine categories, the relative frequencies across each candidate's blog posts is charted

a candidate's spouse (or ex-spouse) is mentioned in blog posts about them. Relationship words are words such as *marriage*, *mother*, *father*, and so on. Finally, gender words are words that indicate a candidate's gender, such as *woman* in blogs about female candidates and *man* in blogs about males.

We found that the conditional probability $P(wordcategory|w)$, were w is a female candidate, was always significantly higher than the probability $P(wordcategory|m)$. In fact, for spouse mentions these probabilities were .959 and .041, respectively; for relationship words they were .705 and .295, respectively; and for gender words the probabilities were .912 and .088, respectively. Figure 15 shows the relative frequencies of each category in blogs about males and blogs about females. It is clear in the figure that these categories are used much more often in blogs about female candidates than in those about males.

5.8. Evaluation

We first wanted to see whether there was a significant difference between the gender bias scores given to the articles about female candidates and those given to the articles about male candidates.

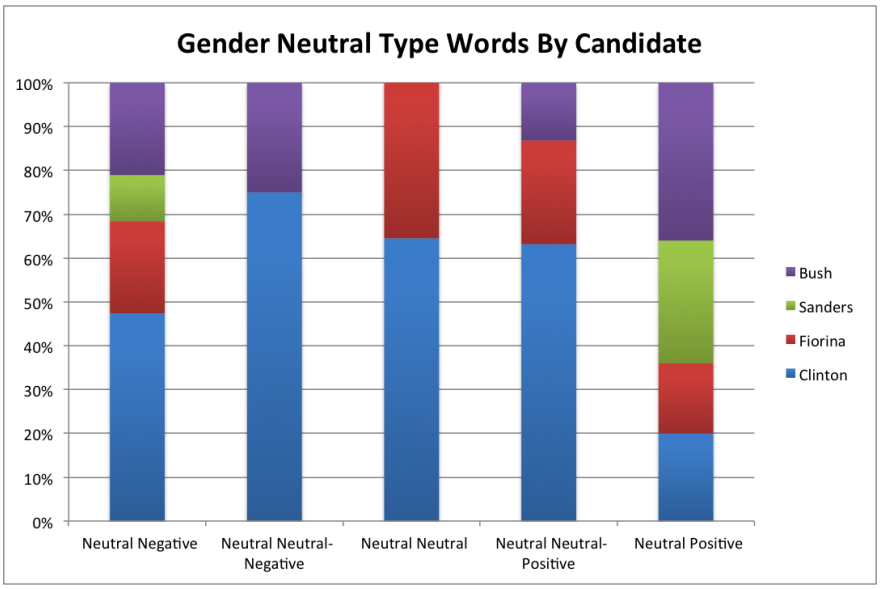

Figure 13: For each of the overlapping Neutral categories, the relative frequencies across each candidate's blog posts is charted

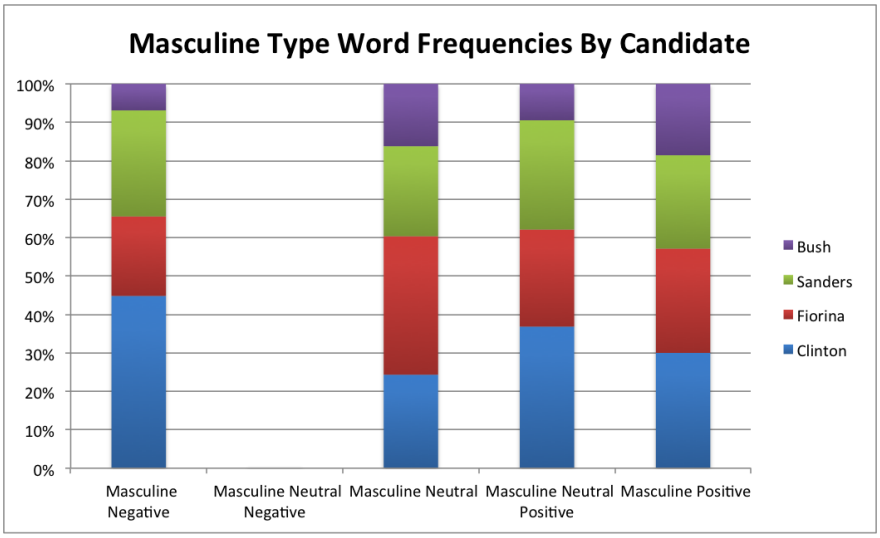

Figure 14: For each of the overlapping Masculine categories, the relative frequencies across each candidate's blog posts is charted

We conducted a two-tailed t-test on the mean gender bias scores given to each of the blogs for both genders. The mean gender bias score for the female blogs was .47216 (on a scale from 0 to 1), while that for males was .23033. The variances were .02455 and .01034, respectively. The two-tailed t-test yielded p = .0007. Thus, there is a significant difference in the level of gender bias in political blogs about females vs males.

We then conducted a linear regression to assess the effect of frequency of strongly gendered

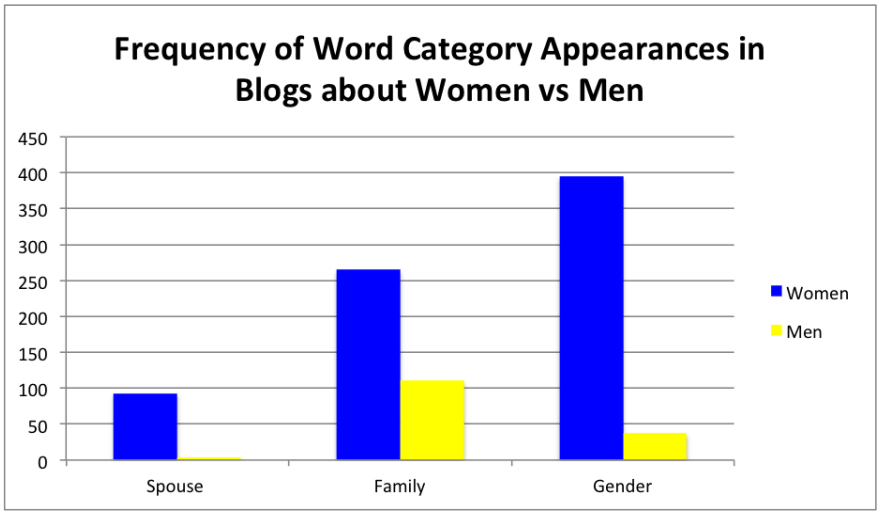

Figure 15: The bottom three bins represent spouse, family, and gender words, respectively. Here the frequency of appearance for females and males is confronted

Regression Statistics	
R	0.6863
R-square	0.471
Adjusted R-square	0.44316
S	0.14197
N	21

GenderBiasScore = 0.23529 + 0.01667 * FreqGenderedWords

ANOVA

	d.f.	*SS*	*MS*	*F*	*p-level*
Regression	1.	0.34097	0.34097	16.91694	0.00059
Residual	19.	0.38296	0.02016		
Total	20.	0.72393			

Figure 16: The outcome of the linear regression run on Gender Bias Rating vs Frequency of Gendered Language

words (both strongly masculine and strongly feminine) on the given gender bias ratings. We found that the frequency of gendered language had a statistically significant correlation with perceived gender bias ratings, with a p-value of .00059 and R^2 of .471, as shown in Figure 16. The regression is shown in Figure 17.

We also conducted a linear regression to determine the correlation between the raters' perceived frequency of strongly gendered language and our calculated frequency of strongly gendered target words. We found that the two were significantly correlated, with a p-value of .00055 and R^2 of .475. The regression is shown in Figure 18. This may indicate that frequency of gendered language is one of the indicators that humans use to assess gender bias in political blogs.

Figures 19.1 and 19.2 show the frequency of words with given gender scores across the blogs used on Amazon Mechanical Turk about women and men, respectively. The distributions are quite similar to those of 4.1 and 5.1, supporting the assumption that the surveyed blog posts were a

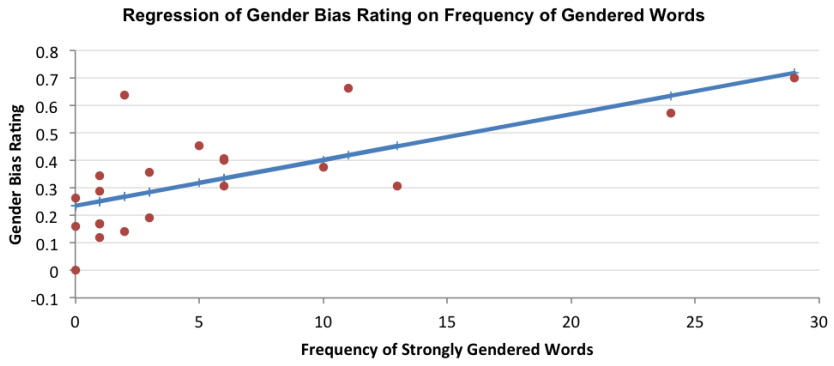

Figure 17: The regression of Gender Bias Rating on Frequency of Gendered Language

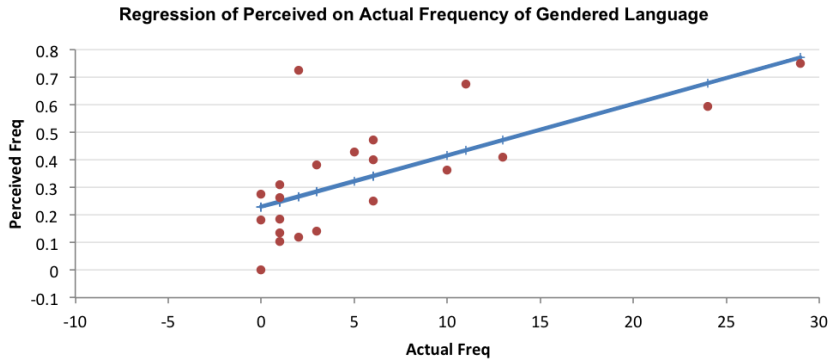

Figure 18: The regression of Perceived on Actual Frequency of Gendered Language

representative sample of all blog posts used.

6. Conclusion

Blog posts are becoming an increasingly relevant source of information, particularly for the younger generation. Hundreds of new trending blogs about the current presidential election pop up on Facebook and Twitter each day, making it easy for individuals to click, read, and form an opinion. Furthermore, the 2016 presidential election is among the first to involve two female front-runners. As such, it is important to examine whether these females are being treated differently in these blog posts - particularly, whether the language used to describe them differs from that used to describe men.

Overall, we found that there were significant differences in the linguistic patterns used to describe females. Particularly, there is a tendency on the part of bloggers to reference a female candidate's gender far more frequently than that of male candidates. Essentially, authors of these posts seem to

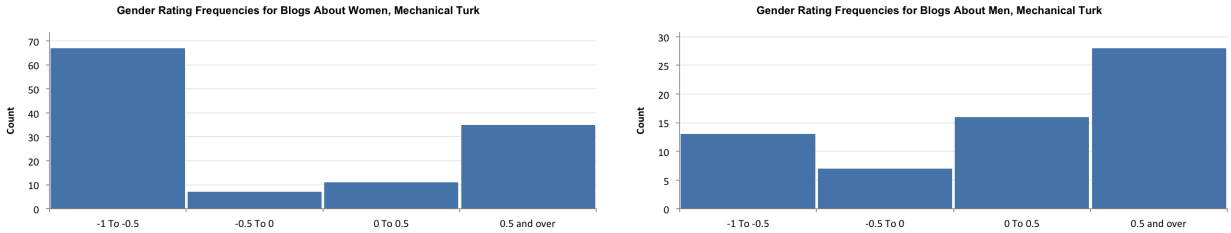

Figure 19: Figure 19.1 on the left is a histogram depicting the frequency with which words of given gender scores appeared across the 10 surveyed blogs about women. Figure 19.2 is a histogram that shows the same frequencies for the 10 surveyed blogs about men.

operate under the assumption that a male candidate is the null hypothesis, while a female candidate is surprising and therefore must be discussed in the context of her femininity. We hypothesized and found that such "surprise" would also lead to female candidates being discussed in more strongly gendered terms overall.

Gendered language included both the categories described by the Finkbeiner Test and words that were determined to be gender biased by our MTurk surveys. Words in the family, relationship, and gender identification categories were far more likely to appear in blogs about females rather than blogs about males, as was demonstrated by large the conditional probabilities $P(wordcategory|g = female)$. Meanwhile, the strongly gendered target words identified through the MTurk surveys also more frequently appeared in the blogs about females. Interestingly, of these words the strongly feminine words were usually adjectives that denoted affection or kindness or nouns such as *lady* - words that usually have high warmth, low competence connotations. On the other hand, the strongly masculine words, such as *powerful* and *formidable* carried connotations of high competence, low warmth. These patterns are worth investigating more thoroughly in the future.

This increased frequency of highly gendered language was also found to significantly correlate with higher gender bias scores as assigned by human raters. Furthermore, the ratings on frequency of strongly gendered words given by the human raters correlated significantly with the actual frequency of these words. Together, these findings may indicate that human raters accurately assess the frequency of gendered words and this frequency is a strong indicator for the perceived level of gender bias in a blog post. Considering the greater frequency of strongly gendered words in blogs about female candidates than blogs about male candidates, this may indicate that the blogs focused

on female candidates show more gender bias. In fact, the difference between the average gender bias score for the blogs about females and that for the blogs about males was found to be statistically significant, with a higher level of average gender bias in posts focused on female candidates.

6.1. Limitations

The primary limitation to this work is that we did not assess the context in which the words were used - in other words, we did not check whether the words' valence may have been affected by surrounding words, such as *not* or *very*. We also could only examine the usage of a limited set of words - our target words - though these were carefully selected to be representative and effective in our analyses. Finally, the phrasing of the Polarity Survey questions may have skewed the results. The question was: "Would you want to be friends with someone described/labeled in this way?" For most words, this worked well. However, for words such as *ugly*, individuals may not have wanted to appear superficial, so they rated the token more positively than they might otherwise have done.

6.2. Future Work

In the future, it would be interesting to give a competency rating to these words, such that we could more accurately examine how the language to describe female politicians and male politicians falls on warmth vs competence measures. It would also be interesting to use our data to identify salient features of a blog post in order to make a classifier that could accurately predict the level of gender bias of a blog post. In order to do so, we would also need to collect a larger data set and increase the size of the training set by having more blogs evaluated by human raters.

7. Acknowledgements

I want to thank my adviser, Christiane Fellbaum, for her support, revisions, and lively conversation. I could not have asked for a better mentor in my first journey into independent work. Thank you to the creators of the Natural Language Toolkit - your book and tools were invaluable to me [2]. I also want to thank all those I love on campus who stayed up late studying with me, let me bounce ideas off of them, and dealt with my overcaffeinated self. I couldn't have done this without all of you!

References

[1] S. Baccianella, A. Esuli, and F. Sebastiani, "Sentiwordnet 3.0: An enhanced lexical resource for sentiment analysis and opinion mining." in *LREC*, vol. 10, 2010, pp. 2200–2204.

[2] S. Bird, "Nltk: the natural language toolkit," in *Proceedings of the COLING/ACL on Interactive presentation sessions*. Association for Computational Linguistics, 2006, pp. 69–72.

[3] A. J. Cuddy, P. Glick, and A. Beninger, "The dynamics of warmth and competence judgments, and their outcomes in organizations," *Research in Organizational Behavior*, vol. 31, pp. 73–98, 2011.

[4] D. Fertig, "Markedness and morphological change."

[5] A. Finkbeiner. (2013) The finkbeiner test.

[6] A. Finkbeiner. (2013) What i'm not going to do - doublexscience.

[7] S. T. Fiske, A. J. Cuddy, and P. Glick, "Universal dimensions of social cognition: Warmth and competence," *Trends in cognitive sciences*, vol. 11, no. 2, pp. 77–83, 2007.

[8] P. Glick and S. T. Fiske, "The ambivalent sexism inventory: Differentiating hostile and benevolent sexism." *Journal of personality and social psychology*, vol. 70, no. 3, p. 491, 1996.

[9] N. Godbole, M. Srinivasaiah, and S. Skiena, "Large-scale sentiment analysis for news and blogs." *ICWSM*, vol. 7, p. 21, 2007.

[10] M. Hu and B. Liu, "Mining and summarizing customer reviews," in *Proceedings of the tenth ACM SIGKDD international conference on Knowledge discovery and data mining*. ACM, 2004, pp. 168–177.

[11] A. Pak and P. Paroubek, "Twitter as a corpus for sentiment analysis and opinion mining." in *LREC*, vol. 10, 2010, pp. 1320–1326.

[12] C. Wagner *et al.*, "It's a man's wikipedia? assessing gender inequality in an online encyclopedia," *arXiv preprint arXiv:1501.06307*, 2015.

8. Appendix

9. Preface to Survey on Article Gender Bias

Broadly, gender bias covers a variety of behaviors related to the differential treatment of males and females. Bias can be conscious or unconscious, and may manifest in many ways, both subtle and obvious. This survey is focused on gender bias in politics. According to research by Dr. Cecilia Mo at Vanderbilt, voters have a more difficult time seeing women as leaders, so they must be more qualified/professional than male candidates in order to succeed in an election.

We are investigating the presentation of gender bias in blog posts about presidential candidates. We are interested in the overt/obvious signs of bias, as well as the more subtle signs, such as using different language to describe males and females, undue focus on a female candidate's gender or body or appearance, or the need to more strongly justify statements about a female candidate's qualifications. For males, gender bias could be exemplified by gender-stereotypic descriptions.

We have previously identified some gender-stereotypic words as, for example, "loving" and "friendly" for a female, or "strong" and "tough" for a male.

A strongly, overtly biased example blog clip:

> "Wow, is Hillary Clinton a bitch or what? The big news today is that the 2016 presidential candidate actually answered some questions from the press. She actually didn't answer anything and dismissed the press because she honestly believes that her lifetime of scandal is nobody's business but hers. . . .Clinton seemed like she was answering a question, but in reality just admitted that she has no idea what she is doing. If she can't answer a simple question with out acting like a psycho bitch, she clearly doesn't have what it takes to run the most powerful country in the world and she definitely doesn't deserve the job."

Note: this is an extreme example of a biased blogs. Blogs do not have to be this strongly worded to present with gender bias

An unbiased snippet from a blog about Hillary Clinton is this:

> "There was a mass shooting on Wednesday afternoon in San Bernardino, California, at Inland Regional Services, a center for people with developmental disabilities...Democratic presidential candidate Hillary Clinton responded quickly to the breaking news on Twitter, pushing the need for further gun control in light of the latest in a long string of mass shootings. Clinton happened to be speaking about the need for gun control at a campaign stop in Florida just as the attack was unfolding, per ABC News' Liz Kreutz. In a tweet, Kreutz quoted Clinton as saying that "90 americans a day die from gun violence, homicide, suicides, tragic avoidable accidents. 33 thousand Americans a year die. It is time for us to say we are going to have comprehensive background checks, we are gonna close the gun show loopholes."

Please read the following short blogs and rate each of them based on how gender biased the blog's views and/or language are.

If you have understood these instructions, please select "No"

10. Target Words and Graphs

Word	Gender Score
Strong	.917
Great	.417
Pretty	-1
Stupid	.167
Firm	.75
Moral	.083
Proud	.5
Impressive	.417
Sexist	1
Powerful	.1
Evil	.583
Friendly	-.583
Sexual	.083
Nice	-.75
Aggressive	1
Formidable	.75
Weak	-.667
Smart	.083
Ridiculous	.083
Qualified	.333
Threatening	1
Ugly	-.25
Tough	.833
Professional	.333
Loving	-1
Natural	-.5
Driven	.5
Honest	-.333
Prominent	.417
Dynamic	0

Table 1: Word List: Adjectives.

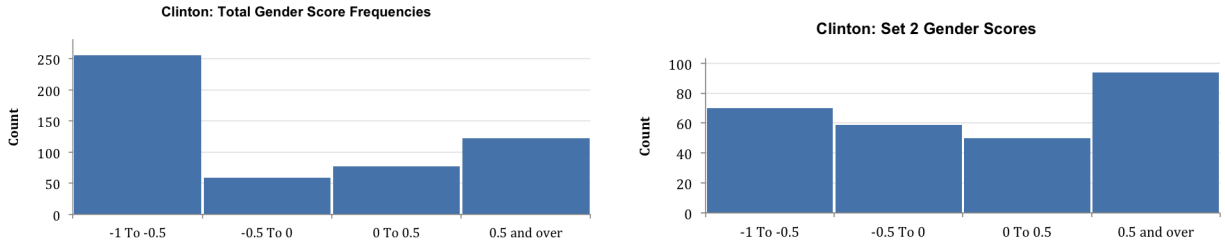

Figure 20: 20.1, on the left, is a histogram depicting the frequency of target words with given gender scores in blogs about Hillary Clinton. 20.2 shows the same for Set 2 target words.

Word	Gender Score
Woman	-1
Man	.917
Married	-.167
Mrs./Ms.	-.917
Leader	.667
Gender	0
Feminist	-1
Secretary	-.917
Female	-1
Husband	1
Lady	-1
Chief/boss	.917
Control	.333
Male	1
Hair	-.5
Mother	-1
Body	-.417
Wife	-1
Sexism	.5
Heels	-.75
Suit	.833

Table 2: Word List: Nouns.

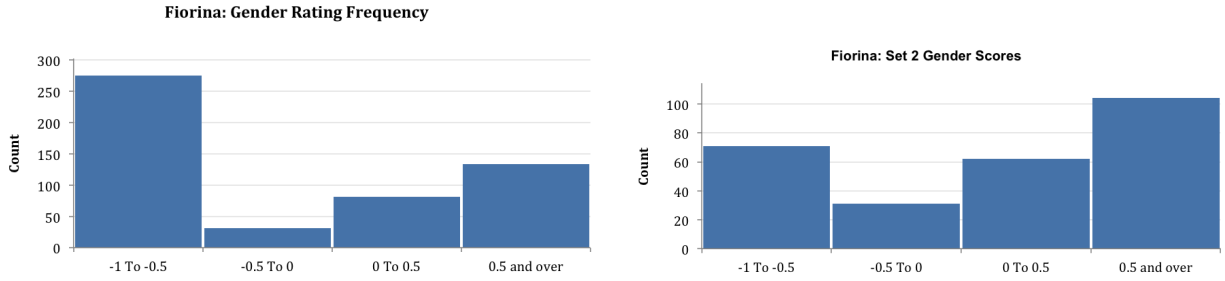

Figure 21: 21.1, on the left, is a histogram depicting the frequency of target words with given gender scores in blogs about Carly Fiorina. 21.2 shows the same for Set 2 target words.

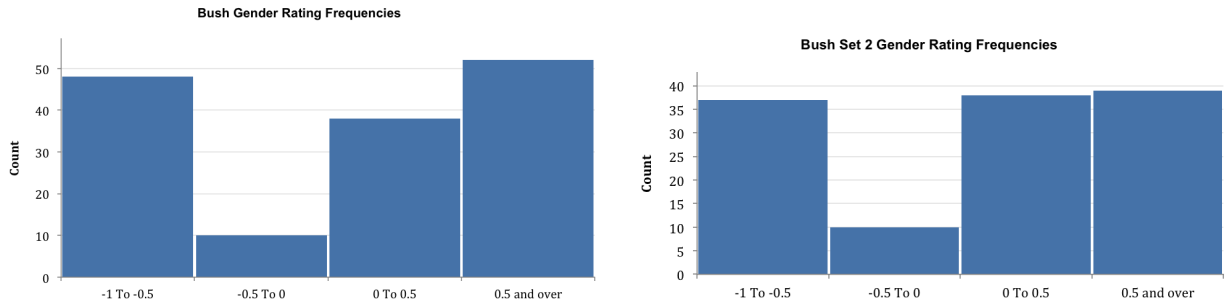

Figure 22: 22.1, on the left, is a histogram depicting the frequency of target words with given gender scores in blogs about Jeb Bush. 22.2 shows the same for Set 2 target words.

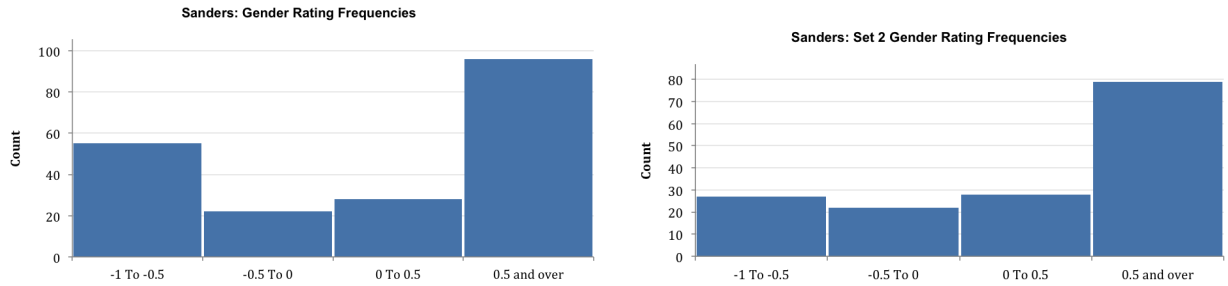

Figure 23: 23.1, on the left, is a histogram depicting the frequency of target words with given gender scores in blogs about Bernie Sanders. 23.2 shows the same for Set 2 target words.

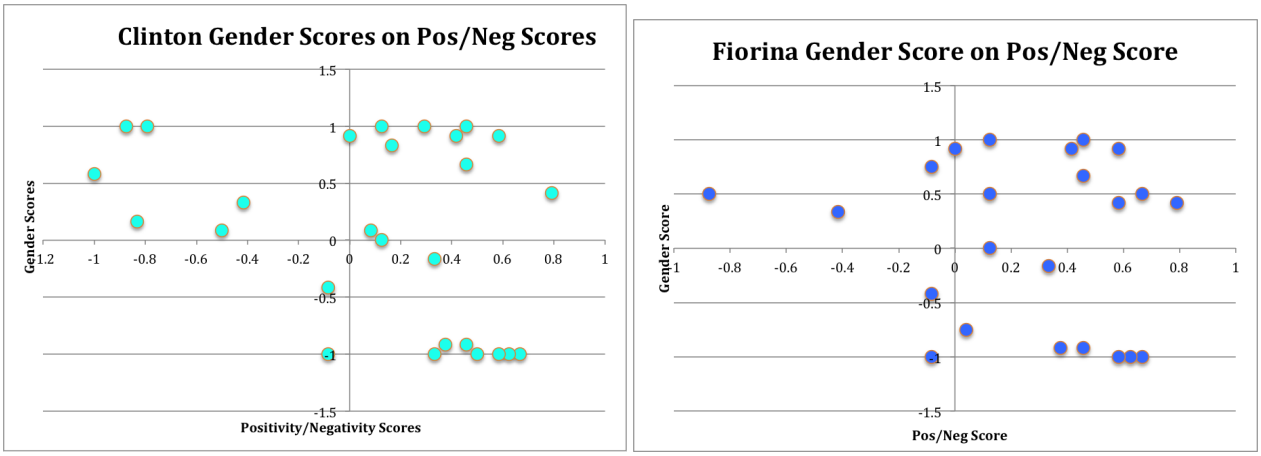

Figure 24: 24.1, on the left, shows the gender score on polarity score of the words used more frequently than the median in blogs about Hillary Clinton. 24.2 shows the same for Carly Fiorina

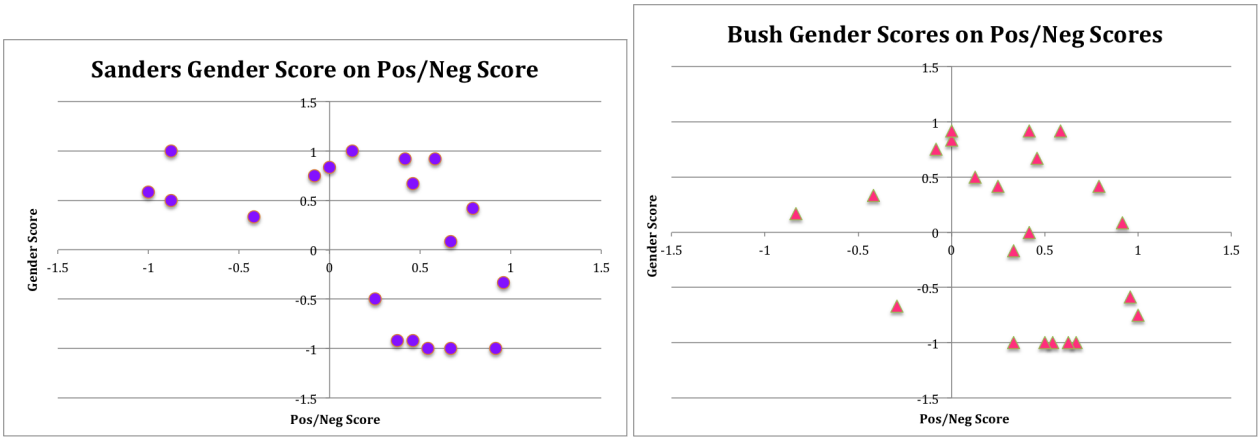

Figure 25: 25.1, on the left, shows the gender score on polarity score of the words used more frequently than the median in blogs about Bernie Sanders. 25.2 shows the same for Jeb Bush

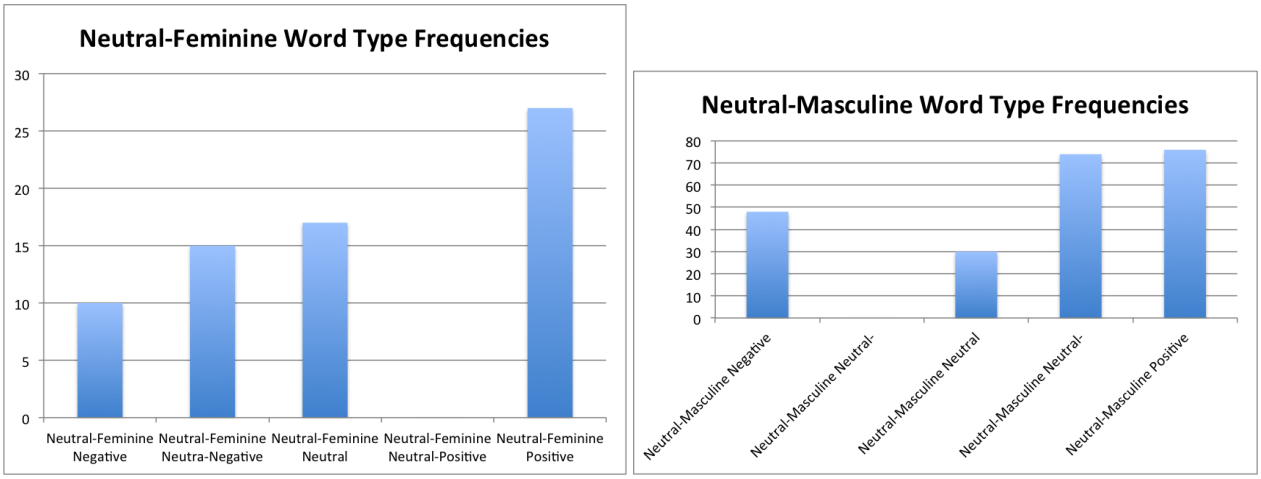

Figure 26: 26.1, on the left, shows the frequencies of Neutral-Feminine word types. 26.2, on the right, shows the same for Neutral-Masculine word types

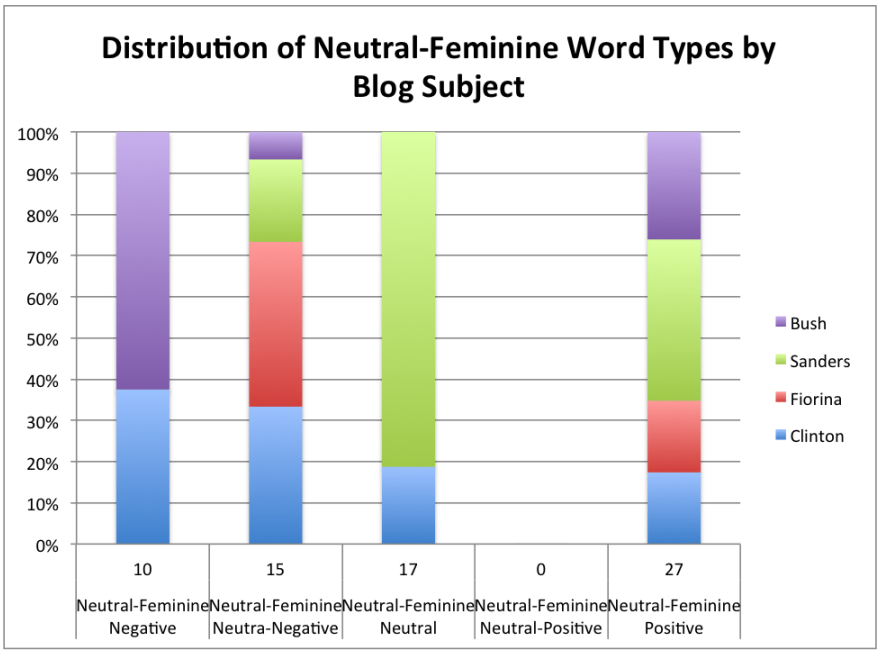

Figure 27: This figure shows the by-candidate proportions of total frequencies of Neutral-Feminine word types

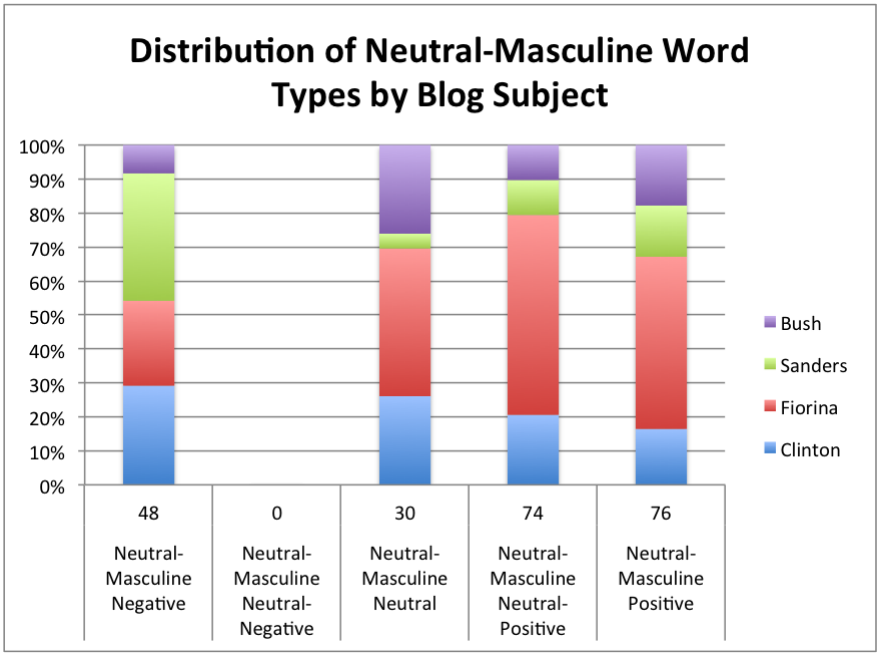

Figure 28: This figure shows the by-candidate proportions of total frequencies of Neutral-Meminine word types