LIN 350 Project: Final Report

Misty Peng mp46528
Maanasa Darisi mvd456

Differences in Male and Female Journalist Use of Adjectives

The use of language plays an important part in the political and social spheres of America. Politicians are widely known to use ambiguous language and half-truths, while some newspapers and journalists are famously bipartisan. In order to maintain voter integrity, neutral news reporting is essential, which is why we want to look at this issue of adjective use in headlines. We hope to find the validity of this issue as it could be a cause of bias in journalists. As a result, it would also affect public perception of important issues that should remain unbiased in the eyes of American citizens. For example, female journalists could find it more important to discuss female reproductive rights than male journalists, which could lead to an unbalanced narrative lacking a multitude of perspectives. Contrastively, issues covered by both male and female journalists would ideally remain consistent, but articles on border patrol could cause divisive opinions, where female journalists may lean towards the negative effect border patrol has on families while male journalists don't mention that issue as much.

The issue with this, however, is the fact that gender is nonbinary. While our goal is not to reinforce harmful gender biases, existing data has not been revised and categorized any differently. In fact, with this research, we hope to find factors that play into current gender biases and to highlight them so we can move away from this type of bias.

One of the earliest works done in this area of focus was by Robin Lakoff in 1973 which shows the prevalence of this issue. Without the same computer-processed data analysis capabilities that we have now, Lakoff's data came from her observations of her own language and the language of those around her. This source and method of data collection as well her time period makes the results dated, however, still relevant for discussion. In the context of our project, one of Lakoff's conclusions was that women avoided strong emotions and feelings in their speech compared to men. This meant fewer curse words, euphemisms for taboo subjects, and the trivialization of important matters. This research paper remains a strong proponent that men and women do indeed speak differently, now called the 'dominance' approach to language because it focuses so much on social prowess.

Another paper on this area, however, with different results, is the 2018 study done by Barczewska and Andreasen. This research included computer-assisted data analysis and focused on the Michigan Corpus of Academic Spoken English, which is a collection of 152 transcripts with about 1.85 million words. This data, however, is collected from activities on a college campus, some being formal like lectures and labs and others less formal like tours and study groups. This means some differences from our goal to look at headlines, but still with relevant findings in terms of gender. The results did not have specific adjectives that leaned either towards male or female. This shows ambiguity in the realm of words spoken by women and men, unlike Lakoff's conclusions, which is the majority opinion in research like this. After looking at research done in areas of spoken words, we wanted to see if there were any commonalities among that area with written word in terms of headlines for news articles.

We changed our research question to better fit our data. Our original objective was to find the possibility of creating a summary of a news story based on multiple news articles. The issue was that the Kaggle database is of news headlines and not whole articles, so we decided that we could maintain more accuracy and better manage smaller pieces of data like headlines as

opposed to entire articles. From this narrowed scope, we decided to add the factor of gender into our project as well. This led to our new research question which deals with gender differences, if any, when it comes to writing headlines of news articles.

We obtained two sources of data for this project. The first is the Kaggle database [1] consisting of 200k headlines from the years 2012-2018. Our other dataset is from UC Irvine's Machine Learning Repository called the Gender by Name Dataset [2]. This was an addition to what we previously were using because our question changed to include gender. Originally we were also using a dataset drawn from CNN and Daily News articles, however, these included full articles, since we decided to go with smaller pieces of data we did not end up using this dataset. With the research question change, our data now only consists of the Kaggle database of news headlines with just over 200,000 headlines. This data came in a CSV file labeled with a news category, author, date, and short description. After cleaning this data by removing headlines without an author, we were left with approximately 160,000 headlines in our dataset to be matched with a gender.

For the gender aspect of our question, we also needed to add a database of genders matching with names, for lack of a better way to determine the gender of the journalists. This is sourced from UC Irvine's Machine Learning Repository, with a size of about 150,000 names. This way, we were able to connect a headline to a gender through the name of the author. By adding this database, the process of cleaning and organizing the data also includes running through the name gender database to correctly sort names into a male and female list. The name data was originally presented in a CSV file with the columns of name, gender, count, and probability. Some of the same common names appear twice, once categorized as M and another time as F. Because of our project, we needed two distinct and unique lists of names of either M or F with no overlap, since this data file is arranged in decreasing probability, we just took the first appearance of a name and did not include the second appearance, decreasing the total number of entries to 134k instead of the original 147k.

Our first steps were to clean the data, as discussed in the data section. To get rid of irrelevant entries, we did this by deleting articles with no authors and those with authors whose names are not in the UC Irvine database. Next, we organized the labels and headlines, by looping through the name CSV and separating the male and female names based on first occurrences. At first, we assumed that each name would only appear once, but after some preliminary testing, we found that there were duplicates in the instances of some common names, where parents named both male and female babies names like 'John'. After we made unique lists of tuples of male and female names paired with their M/F tag, we matched them to the authors' names and the titles of their articles.

Moving on from here, our next steps were to create prediction-based spaces for the headlines as a whole, headlines written by female journalists, and headlines written by male journalists. To create the prediction-based spaces, we tokenized the headlines from the entire database and used Gensim's word2vec to create the spaces. We used two lists of headlines, one from all authors with female names and a list for those by authors with male names. Then, in order to compare male and female usage, we created a space using the tokenized words from these two lists. We also tokenized words from headlines under every category to create a prediction-based space for the headline dataset as a whole.

---

[1] Rishabh Misra, *News Category Dataset*, https://www.kaggle.com/rmisra/news-category-dataset

[2] UC Irvine Machine Learning Repository, *Gender By Name Dataset*, https://archive.ics.uci.edu/ml/datasets/Gender+by+Name

We also further separated the data by headline categories that were provided in the CSV file. For example, we took only the headlines labeled ‘Sports’ and compiled them into three spaces, one for all authors, just male authors, and just female authors. The nearest neighbors are shown below. Some results are not surprising, such as the female neighbors like ‘Khloe’ and ‘Kourtney’ (famous Kardashian sisters that are often in the news and linked to sports players), and ‘Rumors’ (according to the female stereotype that likes to follow gossip and create rumors). The neighbors from the all-headline space are also not surprising, with words like ‘Hospitalization’ and ‘Odom’ (a professional American basketball player named Lamar Odom) The male neighbors are a little less accurate with the nearest neighbor being ‘Kendrick’ from the singer Kendrick Lamar, who does not play sports, as opposed to basketball player Lamar Odom. The other results further stray from sports with most of the neighbors being names that aren’t usually related to sports like ‘Aniston’ (Jennifer Aniston, an actress) and ‘Aguilera’ (Christina Aguilera, a singer). This creates some doubt in the accuracy of our first try at prediction-based spaces.

All headlines:

```
Word to find: Lamar
('Odom', 0.8225144743919373), ('Kendrick', 0.7553653120994568), ('Hospitalization', 0.6741722226142883),
('Goulding', 0.65189528465271), ('Morrissey', 0.6327471733093262), ('Hiddleston', 0.6230356693267822),
('Harden', 0.6204155683517456), ('Khloe', 0.6133443117141724), ('Kimye', 0.6114490628242493),
('Meek', 0.609896183013916)
```

Male headlines:

```
Word to find: Lamar
('Odom', 0.8225144743919373), ('Kendrick', 0.7553653120994568), ('Hospitalization', 0.6741722226142883),
('Goulding', 0.65189528465271), ('Morrissey', 0.6327471733093262), ('Hiddleston', 0.6230356693267822),
('Harden', 0.6204155683517456), ('Khloe', 0.6133443117141724), ('Kimye', 0.6114490628242493),
('Meek', 0.609896183013916)
```

Female headlines:

```
Word to find: Lamar
('Kendrick', 0.9326701760292053), ('Odom', 0.8943759799003601), ('Aniston', 0.7940096259117126),
('Jake', 0.7872006893157959), ('Mic', 0.7824608087539673), ('Aguilera', 0.780041515827179),
('Stuns', 0.7747182846069336), ('Carrey', 0.7651447057723999), ('Jessica', 0.7640302777290344),
('Diane', 0.7636460661888123)
```

After conducting a sanity check on this data, we found that the Pearson correlation number is not large enough to prove consistent accuracy in the spaces that we created. Our next step was to use a more accurate method for our dataset because it is on the smaller side.

After the prediction-based space, we wanted to look at the cultural biases and directions in the headline word embeddings, and we did this in two different ways to see if the results would be similar between both. First, we used a pre-computed Gensim space to calculate the mean cosine similarity between the words used in headline articles to two groups of anchor terms. The first group was “man”, “he”, “him”, and “boy”. The second group consisted of “woman”, “she” , “her”, and “girl”. In order to do this, the data has to be further grouped out. We took the list of headlines and sorted them out based on three of the many categories listed. These were Sports, Politics, and Parenting. We chose these three specifically as each seem to be topics of life that are heavily gendered. Next, we further separated these three lists of headlines

by gender, so that we now had six sets of headlines, for example, headlines written by males in the sports category. Like before, we tokenized the headlines to get the individual words. Next, using NLTK's POS tagging, we separated just the adjectives from each of these lists of words, so now a list of data would consist of something along the lines of adjectives used by females in the politics category.

With each of these six sets, we calculated the cosine similarities between every word and the anchor terms. After taking the difference of the mean cosine similarities, we ended with 6 values, as shown below.

Difference in Cosine Similarities: Adjectives

	Male	Female
Sports Adjective	0.038	0.033
Politics Adjectives	0.032	0.029
Parenting Adjectives	0.004	-0.002

A higher number means that the words in that category were, on average, more similar to Group A, "man", "he", "him", and "boy". The negative number seen with adjectives in parenting articles by female authors indicates that that set of words had a mean cosine similarity difference that put the influence closer to anchor term group B, "woman", "she" , "her", and "girl". As expected, in all three categories of article types, there was a higher correlation to anchor term group A from adjectives used by males than females. The p-value calculated between these two sets of numbers was greater than 0.05 (0.83), signifying that the difference is not statistically significant, the results do still align with our original expectations and there is a slight difference between the two, with a bias in the direction we assumed there would be one in.

With the second method of looking at cultural biases and directions in these word embeddings, we computed a "gender direction" as the direction from woman to man. This was done using the same 6 sets of adjectives in the same three categories of articles. Using scalar projections of different vectors, we could calculate how far along the woman to man vector each one fell. We again received six values, for which a negative number indicates it being closer to "man" than "woman", and a positive value indicates the opposite.

Looking at Gender Differences: Adjectives

	Males	Females
Sports Adjective	-0.255	-0.012
Politics Adjectives	-0.064	0.032
Parenting Adjectives	-0.183	-0.121

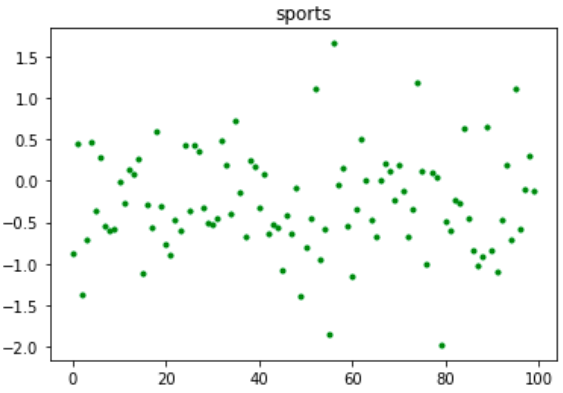
Again, we see that the adjectives used by male authors point closer to the direction of "man" than adjectives used by female authors. Overall five out of six categories resulted in positive values, meaning they are closer to the masculine direction. This happened before when

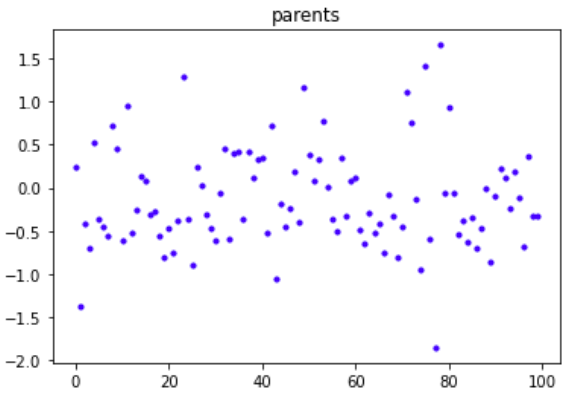
using cosine similarities as well, however, with the gender direction vectors, it was the politics category that results with a number that is flipped from the rest. Again, the p-value was greater than 0.05, which means that the difference between these two sets of values is not statistically significant but as seen before, the difference that is there is as expected. It seems that on average, the female and male authors are influenced by their surroundings enough that their writings may reflect this gender influence. Overall, it seems most of the adjectives used are closer to the male anchor terms or the "man" directions, as seen with both methods of looking at cultural biases. We calculated these values again for all words, not just adjectives to see if the amount of bias shifts in any way, using the cosine similarity method.

Difference in Cosine Similarities: All words

	Males	Females
Sports Adjective	0.049	0.044
Politics Adjectives	0.042	0.039
Parenting Adjectives	0.018	0.014

As seen in the table above, when taking into account all words used, excluding stopwords, the bias leans even more heavily towards anchor term group A. This indicates that adjectives specifically are something that authors are more influenced to choose depending on their gender than other parts of speech.





The above plots show the gender direction values for the first 100 words of the sports and parents categories. This is for all words excluding stopwords, not just adjectives. As seen here, most words have a value on the y-axis that falls under 0, which indicates the direction being closer to "man". Even for topics like parenting, which is a facet of life that women have generally always been in charge of, the bias goes in the direction of the masculine anchor group and "man" direction. This may be because published writing is very different from spoken speech, which means that authors are often using words they would not usually pick. Female authors may tend to use less flowery language than they would in real life in order to sound more professional, and it makes sense that professional

language has a gender bias towards the masculine anchor group/direction as professional spheres have been headed by men for decades.

Our data analysis did not go as far as to show exactly why these differences in bias occur between male and female authors, although we predicted a few, including the need for more professional sounding language and the topics that authors of different genders choose to discuss. A person's gender impacting the language that they choose to utilize has become a persistent issue in recent times, with many working women trying to bridge the gap by actively speaking with less female influenced language. Similarly to Barczewska and Andreasen's study in 2018, we found ambiguity rather than a defined pattern of differences between male and female journalists' habits of writing headlines. This shows that more research needs to be done, possibly with access to more data, however, our research has positive indications that there are biases that can be brought to attention and resolved.

References

Barczewska, Shala & Andreasen, Agata. (2018). Good or marvelous? Pretty, cute or lovely? Male and female adjective use in MICASE. Suvremena Lingvistika. 44. 194-213. 10.22210/suvlin.2018.086.02.

Lakoff, Robin. “Language and Woman’s Place.” Language in Society, vol. 2, no. 1, Cambridge University Press, 1973, pp. 45–80, http://www.jstor.org/stable/4166707.