Investigating Gender Bias in Sociopolitical News Articles

Ali Mian
Occidental College
Computer Science

Logan McIntyre
Occidental College
Computer Science

Max Peng
Occidental College
Computer Science

***Abstract*—Journalism and print media are an incredibly important and vital tool when it comes to dissemination of information and knowledge to the masses. Newspapers and articles are widely considered reliable and trustworthy with regards to delivering the facts and educating their audiences about important issues and events happening around the world. In times of crises, social change, political movement. Examples of these include election campaigns and debates, changes in policy, the MeToo movement etc. People rely on the news for the unbiased truth and trust the strict and necessary standards of journalism which govern what gets printed and what does not. However, this truth may not be as unbiased as the public believes it to be, and is often ridden with biases associated with personal identity, specifically gender. This paper discusses the use of NLP methods to explore gender bias, specifically in sociopolitical news articles with a history of notable gender markers and gender bias.**

I. Introduction

In our research paper, we will be investigating news articles centred around Hillary Clinton's political campaign in 2016 for President. We felt articles surrounding a political campaign may be more polarized with significant gender markers which we can use for analysis. Furthermore, the news coverage surrounding Clinton's campaign has been met with widespread criticism, specifically relating to bias in reporting as the subject was a female political candidate, which we believe would have significant gender markers. There has been evidence in the past about significant bias in the way female politicians are portrayed in the news, claiming that the media's treatment of female candidates has been very sexist. Our research question is to investigate whether such news articles can be identified according to the author's gender, and how this relates to bias in reporting. This paper discusses how a gender classifier was created and how informative features used to distinguish the gender identity of the author translated to bias in reporting.

II. Related Work

The prevalence of these biases can be seen in countless research papers on the subject. One such paper, Enduring Gender Bias in Reporting on Political Elite Positions: Media Coverage of Female MPs in Belgian News Broadcasts (Hooghe, Jacobs, Claes 2015) [1], explores the biased news coverage of female politicians in Belgium. This paper posed the question 'Has the increase in female representation in parliament resulted in a change in how female politicians are represented in the media?'. This was done by looking at over 6,000 full newscasts recorded between 2003 and 2011. These newscasts were analyzed using multilevel linear regression. In the first level the dependent variable was if the member of parliament (MP) was given speaking time. This was to determine why some MPs were given speaking time, while others were not. The second level dependent variable was the speaking time of the MPS. This analysis provided the following results; bias in news media has persisted despite the increasing number of female MPs. Female MPs receive significantly less speaking time than their male counterparts when they are allotted time, which happens at a much lower rate as well. These results provide an interesting look into what we can expect to find in an analysis of political articles. This research paper provided evidence that there is an anti-woman bias in terms of the representation of female politicians. What we hope to explore is what, if any, kinds of biases towards women exist in print news. Furthermore, televised news coverage is still predominantly directed by men. By exploring news articles written by both men and women, we can see if there is a difference in types/amount of bias in news told from the perspective of a man or woman. Findings here would provide a more nuanced look into news bias, potentially exposing if the bias is institutional or individual. While this paper did not provide a technical framework that we can apply directly to our problem, it gave theories and resources

that inform both what we are looking for and why it is important.

There has been other work done on identifying bias in newspaper articles. Detecting and Identifying Bias-Heavy Sentences in News Articles (Hirning, Shankar, Chen) [2] investigates identifying bias heavy sentences in news paper articles using convolutional neural nets. Articles in this study were classified not according to the news being covered, rather by their news provider. Over 10,000 news articles were collected from five different news sources. Using a classifier, certain bias-heavy words and phrases were identified within the articles. The results proved to be successful, achieving an 84 percent accuracy with a 99.6 percent consistency with the classifier prediction. This model of course was trained to identify general bias within sentences whereas the purpose of our research is to identify bias specifically regarding gender and how it relates to the gender of the author, as opposed to classification according to the news provider.

Another paper we found attempts to investigate how different genders are described in journalism. "Deep and Machine Learning Approaches to Analyzing Gender Representations in Journalism" (Campa, David, Gonzalez) [3] investigates the way people belonging to different gender identities are described in the news. Furthermore, classifying the gender identity of the subject in the news headline is used as a way of gauging gender bias. This research also uses a convolutional neural network as well as Naive Bayes for classification and obtains an 86.7 accuracy on the task of classification. The data consisted of news headlines of approximately 1800 news headlines from various major news outlets. The results were successful and provide us with some interesting ways to build our model and achieve better accuracy. While this model only considers news headlines and the gender of the subject, our research intends to explore the content of the article and the gender of the author and how the gender identity may result in gender bias.

Additionally, we came across research papers investigating gender classification of authors by their written works. In "Gender Classification with Deep Learning" (Bartle, Zheng) [4], the authors explore Recurrent Neural Networks to build a classifier which can predict the gender of the author based on the written work. In their approach, the model learns to predict gender based on the surrounding context of words. The dataset they use comprises blogs and literature from several centuries. The model receives an 86 percent accuracy on the blog dataset using a WRCNN model, which is comparable with other state of the art implementations.

In addition to finding research papers which explore gender bias in journalism, we found one other which attempts to identify the gender identity of the author according to their written work. This of course is an essential part of our model. In Gender Classification of Literary Works (Abrams, Chavira, Wong) [5], the question of whether an author's gender can be determined based on their work is explored using Naive Bayes. The data consisted of books written by both male and female authors. Book genres included short stories, poem collections and novels. The NLP techniques employed four methods of feature selection and three variants of Naive Bayes. The final results outperformed baseline results which were 79 percent and proved to be successful. Our research considers news paper articles as opposed to literary texts which is an important difference, however, a similar technical framework may be applied.

III. Methodology

A. Data Collection and Preprocessing

Our dataset was downloaded from Kaggle. This data comprises sociopolitical news articles from a variety of different news sources, all between 2016 and 2017. The data includes the publication name, name of the author, date of the article, year and month article was published, a URL for the article, and the content of the article. From this, we selected articles surrounding Clinton's campaign during the year 2016 specifically covered by two major news sources: The New York Times and The Washington Post, with the keyword being "Hillary Clinton". We chose the above two news sources because they use a relatively more objective tone compared to some other publications. When combining multiple news sources, we wanted to make sure that the gender bias is identified based on individual journalists' writing style, rather than the opinion of a specific news source.

This data set, however, did not include the gender of the author. Therefore, we had to annotate the data ourselves, adding an additional column for gender and hand classify the data. We used M for male and F for female. In some cases, more than one author was listed for the article. To classify the gender for multiple authors, we decided that if all authors listed were male or female, we would mark them as M or F respectively. If the article was written by both male and female authors, we classified it as B for both. This was the final distribution:

Gender	Count
M	185
F	101
B	43

As shown, there are more articles written by male authors than female authors. It is not ideal, but there is still a significant number of articles written by female authors to provide enough insight. If the unbalanced labels later are proven to be problematic, we can trim the dataset so that the numbers of male/female labels are balanced. To make the classfication task easier, we only chose to use articles that are labeled as either "M" or "F" for gender.

Using the Pandas library, we were able to create a new data frame with an added column for gender of the author. Moving on to preprocessing and cleaning up the data, we decided to adopt the options to remove stop words and punctuation, as well as to tokenize the content of each article into words. The stopword list was created using NLTK's stopwords library with words that contain the "negative" logic removed. Such words include isn't, aren't, weren't, shouldn't, haven't, etc. Taking those words out of the list was necessary because the sentences without them would give us the completely opposite meanings. For analysis and training our classifier, we were specifically looking at the gender of the author and the content of the article.

B. Technical Approach and Models

Given the size of our dataset, it was important to pick the appropriate model for classification and training. While other works exploring gender classification and bias experimented with Recurrent Neural Networks to train and test on their data, our dataset was comparatively smaller. We decided, therefore, to implement Logistic Regression model and test for accuracy and precision on different types of feature extraction. For our baseline, we decided to move forward with TF-IDF features, trained on a Logistic Regression model.

To provide a brief overview, TF-IDF turns a text into a vector so it can be used in our models. This is done by combining the two metrics in the name of the algorithm, Term frequency and inverse document frequency. Term frequency is, as the name implies, the frequency of each term (word) found in a text. Inverse document frequency refers to to the inverse number of documents that a word appears in. This method is useful because it helps determine not simply which terms are frequent in a given text, but which terms are frequent more unique to the specific text. TF-IDF provided the base feature extraction for our data, while not the most nuanced method, it is a good indicator of if we would find any noticeable differences between articles written by a man or a woman. If there was no noticeable difference in TF-IDF values for articles written by men vs women, then it would suggest that there may not be a strong relationship between the word choice of an author and their gender.

We narrowed down our model of choice in the first round of implementation to Logistic Regression and Multinomial Naive Bayes in conjunction with TF-IDF. We decided to use TfidfVectorizer from sklearn because of its streamlined design. Inside TfidfVectorizer, there are parameters for further fine tuning of the input. For the initial model comparison, we decided to utilize the stopword removal method that comes with the TfidfVectorizer. There was no other parameter tuning for this stage. For this and all future implementations, we used 5-fold cross validation and to return the most unbiased results. Sklearn's cross_validate method was used to achieve this task. Eventually, we decided to use Logistic Regression for further implementation because of the better performance.

Logistic Regression was further fine tuned with a few additional techniques. TfidfVectorizer has the ability to include any combination of n-grams. After some tweaking, we decided on using a combination of unigram + bigram for the vectorized input, as well as setting the maximum number of features to 1000. We also tuned the Logistic Regression model with C=10000. Most informative features by each gender label were also calculated by sorting the coefficient of the features combined with feature names.

To provide more context for the features, we implemented parts-of-speech(POS) tagging. POS tagging involves adding a label to each word in a text that corresponds to a part of speech (ie, noun, verb, etc.) based on the context in which the word is found. While an easy concept to grasp in practice, it is often not straight forward. For example, consider the word fly. In the sentence there is a fly on the wall, the word fly is a noun. In a different sentence, like superman can fly, it is a verb. While this is intuitive for a native speaker, it is something that makes English difficult to learn for both people and computers. By implementing POS tagging on our dataset, in conjunction with TF-IDF, we were able to derive valuable insights as to not only what terms are used more frequently by a given gender, but also what parts of speech each gender prefers.

To ensure the accuracy of the POS tags attached to the end of the words, we decided to create tags

first, then remove the stopwords. Doing this allows the pos_tag method identify the complete structure of each sentence, which includes words such as ”is”, ”about”, ”to”, etc. The parameters in TfidfVectorizer and Logistic Regression this time in conjunction with POS tags were almost unchanged, with the exception that max_iter wass set to 1000. Most informative features were also reported with the POS tags attached.

In addition to TF-IDF, we used another vector representation of text for feature extraction, sepecifically, Doc2Vec. Doc2Vec works is a lot like Word2Vec, however with an added vector for paragraph ID. This type of vector representation generates vector representations of words which carry semantic meaning. Every unique word in the corpus is assigned a vector in the space. With Doc2Vec, when training the word vectors W, the document vector D is trained as well, and in the end of training, it holds a numeric representation of the document. For our model, we decided to implement a continuous bag of words model which is similar to the skip gram algorithm used in Word2Vec.

IV. Results Analysis

Our initial version of the model training on TF-IDF features with a Naive Bayes Classifier did not return promising results, producing an accuracy of 35 percent with Precision and Recall of 1.0 and 0.5 respectively. The results indicated that the model was classifying all the articles as written by women, which was strange and indicated a potential imbalance of features. Following an implementation of the Doc2Vec model, the results improved, however were not promising or substantial.

	Doc2Vec	Fine Tuned Doc2Vec
precision macro	0.49	0.5342
recall macro	0.49	0.5336
f1 macro	0.49	0.5301
precision weighted	0.54	0.5700
recall weighted	0.55	0.6469
f1 weighted	0.54	0.5744

Here are the initial results when comparing Logistic Regression and Multinomial Naive Bayes with no parameter tuning:

	Logistic Regression	Multinomial Naive Bayes
precision macro	0.6292	0.3234
recall macro	0.5243	0.5000
f1 macro	0.4411	0.3928
precision weighted	0.6387	0.4185
recall weighted	0.6643	0.6469
f1 weighted	0.5451	0.5082

After fine tuning the model and training it on Logistic Regression, the results improved significantly. Here are the results with and without the POS tagging:

	TF-IDF + N-grams	TF-IDF + N-grams + POS tagging
precision macro	0.7540	0.7232
recall macro	0.7261	0.7031
f1 macro	0.7247	0.7024
precision weighted	0.7687	0.7416
recall weighted	0.7692	0.7449
f1 weighted	0.7564	0.7344

Here are the most informative features for both male and female writers. The features are extracted from Logistic Regression model in conjunction with TF-IDF + N-grams + POS tagging.

Coefficient	Most Informative Features for Female
-18.649103	foundation-nn
-14.358238	keep-vb
-13.268541	clintons-nns
-12.692948	voters-nns
-12.619410	clinton-nn foundation-nn
-12.352482	briefing-nn posted-vbd
-12.272360	mr-nn sanders-nns
-11.995323	posted-vbd
-11.921507	briefing-nn
-11.887587	million-cd
-11.815164	women-nns
-11.523852	material-nn
-11.332599	mother-nn
-11.148910	turned-vbd
-10.844189	mr-jj trump-nn
-10.783651	leading-vbg
-10.654950	know-vbp
-10.570964	bill-nn clinton-nn
-10.091250	response-nn
-10.015097	act-nn

Coefficient	Most Informative Features for Male
16.494725	debate-nn
11.327887	americans-nns
10.889115	gay-nn
10.888439	clinton-jj
10.599619	weapons-nns
10.278732	hard-jj
10.044489	worse-jjr
9.928051	best-jjs
9.787263	results-nns
9.760599	trade-nn
8.964542	iraq-nn
8.961640	takes-vbz
8.828273	morning-nn
8.763980	hillary-nn
8.711025	making-vbg
8.709075	democratic-jj
8.696901	obama-nn
8.679345	democrats-nns
8.382556	media-nns
8.126026	despite-in

When looking at the list of most impact features, there is a stark contrast in both word choice and POS choice. Out of the top 20 terms most associated with an article being written by a women there is only 1 adjective and it is an arguable one (the mr in mr trump) while in comparison one fifth of the most telling terms from an article written by a man are adjectives. The opposite is true of verbs, with female writers using them at a higher rate than their male counterparts. Outside of the verb-adjective discrepancy, the part of speech's used are predominantly nouns for both sex's. The word choice, however, provides some significant differences.

The top 5 most important features for female authors were the following: foundation, keep, Clinton, voters, Clinton foundation. While the 5 most important features for males where: debate, Americans, gay, Clinton, weapons. While there are some similarities (primarily Clinton unsurprisingly) this also paints a clear distinction, not in just how authors of different genders pick what types of words they use but also in what type of stories they are reporting on. For this let us focus in on the 5th most important feature for the two genders, Clinton foundation vs weapons. This is not a case of different authors choosing different words to describe the same news, this gets at a difference in what is being reported. A significant portion of the articles we had about Hillary Clinton written by women were about the Clinton foundation, a non for profit founded by Hillary Clinton's husband Bill Clinton. While the articles written by men focused more on foreign policy, the presidential race, etc. This discrepancy in the topics being reported on was one of the primary reasons that the articles were able to be classified by gender.

When embarking on this research one big question was, "given the strict standards and guidelines journalists must follow, will there be enough room for any significant gender difference?" It is possible the answer to this question is no, there may not be enough room in news articles (from respected publications) for an authors gender to come through. The differences in adjectives vs verbs aside, there does not seem to be a significant difference in how men and women describe the same news story. What is clear, is that the news they are reporting on is often not the same. More specifically given an world event covered by one-hundred people, there may not be a clear split with half the reporters being men and half being women. Focusing again on the features that best classified an authors gender, there seems to be too strong of a correlation between an authors gender and what they are reporting on than can be ignored. From our dataset women are mostly reporting on stories that involve Hillary Clinton and mothers, women, the Clinton foundation, and the basics of the presidential race (who is leading, etc,). In contrast articles written by men are stories about the debates, foreign trade policy, etc. This is a divide significant enough that it allowed the model to predict an authors gender almost 75 percent of the time. That is a large enough gap that it can't be chalked up to chance, there is a clear gender bias, not for or against Hillary, not in how Women are portrayed, but in what stories female reporters are given.

A. Comparison with related work

Our results differ from some of the findings reflected in [1], where Hooghe, Jacobs, Claes were able to prove that there is an anti-woman bias in terms of the representation of female politicians. The paper explored Media Bias Theory which suggests that media "play an integral role in the campaign by framing, shaping, ignoring, or presenting the candidates to the public". This theory applies not just to broadcast news but print as well. Proper, unbiased, news coverage helps make sure that candidates are judged on the quality of their leadership. Given our current findings and how male reporters, when describing Hillary Clinton often refer to stories about weapons, debates, foreign policy etc, and

how female reporters refer to stories revolving around the foundation, competition with Bernie sanders, voters etc. Our findings reveal that while there may be some bias and differences in how men and women portray female candidates, there is also bias in what stories are assigned to male and female reporters. Comparing our gender classification to the work done in [5], our model's results are comparatively less impressive. There is a difference in approach of course. Abrams, Chavira, Won use PMI for feature extraction where as we use TF-IDF and POS tagging. However additionally, their model investigates works of literature, which may reflect greater differences in how men and women write based on the kinds of topics they choose to write about. This of course is to some extent reflected in our findings as well, indicating a noticeable difference in topics assigned to female journalists and topics assigned to male journalists. However, news articles undergo a very strict standard of journalism, whereas literature entails freedom of expression on behalf of the writer, a difference which potentially explains the difference in results from classification. As far as technical approaches are considered, our research followed methods of feature extraction similar to the work done in [4] where the authors also use POS features and paragraph2vec, which is very similar to our doc2vec approach. Just like their results, our results also significantly improved when adding POS features, producing the best results. Our POS features also helped indicate the nature of bias present in the articles we were investigating, displaying the most relevant features differentiating how both men and women write. Additionally, an explanation for why their results are slightly better than our model's are due to the fact that their dataset comprised blogs and literature, which are more obvious in indicating biased writing due to freedom of expression on part of the writer.

V. Threats to Validity

Currently, our model is attaining around 70 percent accuracy and precision. This of course means that our model is still unable to correctly classify gender around 30 percent of the time, which of course threatens the reliability of our results. Furthermore, we could have explore other methods of classification and feature extraction such as a hybrid model with rule based approach to classification as well, taking into account previously researched biases and differences in writing between male and female journalists. Additionally, we could have considered a larger dataset by incorporating more articles, and using deep learning and neural networks to train the model. Furthermore, our dataset only considers news articles from two sources, Additionally, it is difficult to detect more subtle biases in news media and journalism. However, give the nature of political articles, how nuanced or polarised certain perspectives may be, especially pertaining to topics such as the one we're investigating, the results we obtained are realistic and align with expectations regarding detectable bias in journalism and how often it is observed. With a 30 percent misclassification, it is entirely possible and reasonable to assume that the writings weren't as polarised or neutral in terms of writing and any obvious gender markers or biases. Additionally, it is of course possible for exceptional cases where some female writers may express opinions or write in a manner similar to how male writers usually write, and vice versa.

VI. Conclusion

Overall, the research was a success. While precision was not perfect, it was successfully enough that we can safely say that this is an area where more research would be beneficial. As with all good research we have ended up with more questions than we started. We set out to see if gender biases would allow us to classify articles based on the gender of the author. The answer to this question, it seems, is yes. What is not a more interesting and pressing question is what are these biases? Are they personal biases of the author? Are they institutional biases? These questions have significant implications, news is supposed to be unbiased, how can this be the case if there is a segregation between men and women and the news stories they are allowed to report on. While we have made good headway into finding answers to these questions, there are still far to many unknowns to make definitive statements. In the future, we hope to investigate similar questions of gender classification and how it relates to bias, across a larger data set spanning more news sources. It would also be interesting to investigate how the nature of this bias shifts as move our focus from more neutral news sources to famously right wing news sources such as Fox News, and whether there are differences in reporting between both genders. Additionally, we could consider alternative technical approaches with a larger dataset, such as convolutional or recurrent neural networks.

References

[1] Hooghe, Jacobs, Claes *The International Journal of Press/Politics* Volume 20 issue: 4 (2015) "Enduring Gender Bias in Reporting on Political Elite Positions: Media Coverage of Female MPs in Belgian News Broadcasts"

[2] Hirning, Shankar, Chen Stanford University (2017) "Detecting and Identifying Bias-Heavy Sentences in News Articles"
[3] Campa, David, Gonzalez Stanford University (2019) "Deep and Machine Learning Approaches to Analyzing Gender Representations in Journalism"
[4] Bartle, Zheng Standord University (2015) "Gender Classification with Deep Learning"
[5] Abrams, Chavira, Wong Stanford University (2000) "Gender Classification of Literary Works"