EXAMINATION OF GENDER BIAS IN NEWS ARTICLES

by

Damin Zhang

A Thesis

Submitted to the Faculty of Purdue University

In Partial Fulfillment of the Requirements for the Degree of

Master of Science

Department of Computer and Information Technology

West Lafayette, Indiana

December 2021

THE PURDUE UNIVERSITY GRADUATE SCHOOL

STATEMENT OF COMMITTEE APPROVAL

Dr. Julia T. Rayz, Chair

Department of Computer and Information Technology

Dr. John A. Springer,

Department of Computer and Information Technology

Dr. Baijian Yang,

Department of Computer and Information Technology

Approved by:

Dr. John A. Springer

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

LIST OF ABBREVIATIONS

IAT	Implicit Association Test
LDA	Latent Dirichlet Allocation
MAC	Mean Average Cosine
ML	Machine Learning
NLP	Natural Language Processing
NN	Neural Network
SEAT	Sentence Encoder Association Test
STM	Structural Topic Modeling
TM	Topic Modeling
WEAT	Word Embedding Association Test
XWEAT	Cross-lingual Word Embedding Association Test

GLOSSARY

Natural Language Processing - a branch of artificial intelligence that deals with the interaction between computers and humans using the natural language. Its objective is to read, decipher, understand, and make sense of the human languages in a manner that is valuable

Tokenization - the process of splitting a phrase, sentence, paragraph, or an entire text document into smaller units, such as individual words or terms

Topic Modeling - a text-mining tool for discovery of hidden semantic structures in a text body

ABSTRACT

Reading news articles from online sources has become a major choice of obtaining information for many people. Authors who wrote news articles could introduce their own biases either unintentionally or intentionally by using or choosing to use different words to describe otherwise neutral and factual information. Such intentional word choices could create conflicts among different social groups, showing explicit and implicit biases. Any type of biases within the text could affect the reader's view of the information. One type of biases in natural language is gender bias that had been discovered in a lot of Natural Language Processing (NLP) models, largely attributed to implicit biases in the training text corpora. Analyzing gender bias or stereotypes in such large corpora is a hard task. Previous methods of bias detection were applied to short text like tweets, and to manually built datasets, but little works had been done on long text like news articles in large corpora. Simply detecting bias on annotated text does not help to understand how it was generated and reproduced. Instead, we used structural topic modeling on a large unlabelled corpus of news articles, incorporated qualitative results and quantitative analysis to examine how gender bias was generated and reproduced. This research extends the prior knowledge of bias detection and proposed a method for understanding gender bias in real-world settings. We found that author gender correlated to the topic-gender prevalence and skewed media-gender distribution assist understanding gender bias within news articles.

CHAPTER 1. INTRODUCTION

Humans convey knowledge and thoughts in the form of natural language. Many artificially intelligent systems rely on natural language to mimic and learn human perspectives of the world. As the building blocks for such systems, natural language texts have been massively collected for training language models. However, not only the knowledge was collected, but human bias was also contained in the data. Therefore, systems trained with these data learned not just human knowledge, but also human bias.

1.1 Historical Content

When talking about fairness issues in artificially intelligent (AI) systems, we focused on the bias or stereotypes of the models. AI models are ubiquitous and affect our lives in significant ways. In the past few years, due to the occurred scandals, fairness issues had become a necessary consideration of AI models. Amazon's recruiting tool was shown to be biased against female applicants (Dastin, 2018). Google Translate was accused of stereotyping based on gender, for example, it presupposed that all engineers are male and nurses are female (Olson, 2018). The reason behind such malfunctions was the systems learned history bias or stereotypes residing within the training data, and such history bias reflects human bias. Training data, and information contained within it, affect the models that are learned. The data Amazon used to train its recruiting tool was from 10-year applicants' resumes, heavily skewed to male applicants. According to Huang (2018), top U.S. tech companies had a gender gap in hiring where males outnumber females in technical positions such as software developers. Amazon's experimental recruiting engine learned such a pattern in the data and penalized resumes containing terms describing females. Similarly, the fairness issue of Google Translate was caused by the true-to-life biased training data. Google Translate was trained on natural language which linked masculine properties to objects that happen to be grammatically masculine and feminine properties to objects that happen to be grammatically feminine. It then learned such gender stereotypes and projected them into the whole world of nouns (McCurdy & Serbetci, 2020).

Knowledge from fields outside Computer Science could be great compliments to develop methods and tools for fair models. In prior research, the definitions varied depending on the task (Mehrabi et al., 2019). Many prior works proposed to solve fairness issues in ML models or neural network (NN) models. But some works defined bias or stereotype only base on a specific setting, for example, bias within carefully selected datasets, and such definitions could not be applied on other datasets, making the conclusions unreliable. The bias or stereotype revealed from manually selected datasets did not reflect the bias or stereotype in the real world, and the definitions extracted from the results did not correlate with how humans create such bias or stereotype. Frye (1983); Nussbaum (1999) stated that fairness from a gender-theoretical perspective should be understood as the structural framework of power asymmetries. Within the training data of AI systems, fairness should be first achieved by balancing each gender group, where each gender group has equal influence on the same object. Therefore, it was necessary to combine knowledge from relevant literature outside Computer Science, for example from Cognitive Science and Linguistics.

ML or NN models for bias classification tasks were often trained on human-labeled data. It was a natural consequence that human-labeled data can reflect the true bias or stereotype in natural language (Caliskan et al., 2017), but models trained on such data would also contribute to shaping such bias or stereotypes when the models were deployed into use (Devinney et al., 2020). Bias or stereotypes are also potentially harmful behaviors that reinforce the subordination of groups that were already disadvantaged (Crawford, 2017). Therefore, we wanted to explore how bias or stereotypes are present in the data and what were the factors making such bias or stereotypes generated and reproduced.

Methods for bias detection and mitigation are either early stage or late stage. Early-stage methods focused on the dataset while the late-stage methods focused on the algorithms or the models (Shah et al., 2019). Bias mitigation on models that were trained on biased data did not perform as well as removal bias from the training data (Zhou et al., 2021). Therefore, to remove bias from the training data, we focused on understanding how they were generated and reproduced. More specifically, we examined the news articles dataset and the possibility of using topic modeling (TM) to explore how the author's gender affects the appearance of gender bias or stereotypes in news articles.

The topic model is a statistical generative model that reveals a set of implicit topics within the documents in the dataset during the training procedure. Because the training data contains representational bias or stereotypes, some topics tend to be gendered by representing traditionally feminine or masculine content (Dahllöf & Berglund, 2019). We used the structural topic model (Roberts et al., 2019) to explore the association between words and different gender properties. Structural topic model includes covariates into the estimation process to help explore how topics were influenced by predefined covariates.

1.2 Theoretical Grounding

Objects can be understood by humans through descriptions in natural language, without the descriptions the objects would be meaningless to humans. Generally, the world itself was immaterial, but the language described had material effects (Foucault, 1990; Hall, 1997). Similarly, AI systems are implemented to learn the patterns within the training data, and the systems themselves do not have biased or stereotypical preferences. What makes the systems have biased or stereotypical behaviors are the biased or stereotypical patterns within the training data. In NLP, we referred to the data to be a set of texts written in natural language. The representation of things was created when people used language to describe the world, for example, the news articles. Therefore, the natural language had material effects and the way we described or represented groups is related to power relations and could affect the resources distribution (Devinney et al., 2020; Foucault, 1990).

Gender was composed of different meanings and should not be understood in a single perspective (Scott, 2015). According to (Lamas, 2015; Serret Bravo, 2006), culture is formed as a network of meanings that produce social realities. Additionally, the same objects or phenomenon could have different meanings in different political environments, for instance, the definition of gender bias could be different in monarchistic countries from capitalist countries. Therefore, political and cultural intersections comprised the context of gender bias or stereotype. Representations of gender bias or stereotype should then be understood within political and cultural backgrounds (Butler & Lourties, 1998), rather than using pre-defined templates generated by a human.

In a broad sense, bias or stereotypes were spread through our natural language (Maass & Arcuri, 1996), and natural language constructed our understanding of the world. News articles published by media produced and reproduced bias or stereotypes in societies (Foucault, 1990). In the Social Cognitive Theory proposed by Bandura et al. (1963), one of the most important theoretical frameworks, media was understood as a medium of reproducing bias or stereotypes. It stated that human behavioral patterns and attitudes could be observed either in real life or in popular media. Since it was expensive to observe how humans use language to generate and reproduce bias or stereotypes in real life, news articles from mainstream media were more practical for this research. Therefore, the way mainstream media portray women and men in stereotypical roles could be used to examine how gender bias or stereotypes generate and reproduce in news articles (del Teso-Craviotto, 2006). When women and men read news articles with biased or stereotypical content, such representations of gender would be enhanced and perpetuated (del Teso-Craviotto, 2006).

The reason that humans should not be eliminated from the process of understanding bias was that the bias was inherently human and often vague and subtle to detect. Pure mathematical definitions of being biased could only verify or falsify if a specific bias or stereotype appears (Devinney et al., 2020). Prior works had different definitions of bias in NLP and they are often inconsistent and implicit (Blodgett et al., 2020). Therefore, instead of simply identifying the existence of an expected bias, using Topic Modeling combined with qualitative analysis helps to discover what kind of gender bias or stereotypes reside in a corpus.

In this research, we perceived gender as politically and culturally constructed rather than static sexual characteristics of women and men. Following Butler (2011); Devinney et al. (2020), gender was constructed through repeated behaviors over time and the acts that produced our understanding of gendered categories. Therefore, words associated with women and men contributed to reproducing the ideas of feminity and masculinity even though they did not necessarily reflect real-world settings.

We used two gender categories in this research: female and male. We investigated news articles ranging from 2013 to early 2020 from mainstream media.

1.3 Research Question and Objective

This research explored the relationship between gendered topics and the topics highly associated with male or female authors by using the structural topic model. It aimed to answer the following questions:

1. Are there notable differences between the topics that authors of different genders cover?
    - Are female prevalent topics associated with the feminine sphere?
    - Are male prevalent topics associated with the masculine sphere?

Overall, this work aimed to contribute towards fairness in NLP systems by extending the application of topic modeling on unannotated text.

1.4 Assumptions

Assumptions are made during this research:

- Due to the bias or stereotypes in the training documents, some of the topics are gender prevalent.

1.5 Limitations

This research focused on examining gender bias in mainstream media in the United States, thus ignoring other smaller size media. Since the dataset used in this research contained only the textual format news articles from mainstream media, we did not consider other types of news information from other sources, for example, video format news on Facebook or Twitter. We also did not consider neutral gender in this research as we did not find suitable word representations for this group.

1.6 Importance

This research focused to reveal the representation of gender bias or stereotypes in textual data and explore how such bias or stereotype generate and reproduce using topic modeling technique.

CHAPTER 2. BACKGROUND

This review aimed to provide a comprehensive scope of recently proposed bias-related and topic modeling techniques.

Bias had become a vital issue in NLP systems. Research works of bias in NLP were divided into two categories: bias in the embedding spaces (Bolukbasi et al., 2016; Caliskan et al., 2017; Gonen & Goldberg, 2019; May et al., 2019) and bias in downstream tasks (Blodgett et al., 2020). Analysis of bias in downstream tasks ranged from language modeling (Bordia & Bowman, 2019), sentiment analysis (Kiritchenko & Mohammad, 2018), machine translation (Stanovsky et al., 2019), coreference resolution (Rudinger et al., 2018), to toxicity detection (Dixon et al., 2018; Park et al., 2018). There were three directions focusing the bias in embedding spaces: bias measurements, bias detection, and bias mitigation (also known as debias or bias neutralization) (Pryzant et al., 2020). In this section, to align with the research goal, we would focus on bias measurement methods and bias detection methods.

2.1 Bias Measurement

Implicit Association Test (IAT) was a measurement designed to detect implicit stereotypes in social psychology by Greenwald et al. (1998). IAT measured the implicit bias of the participant through a series of seven tests. The first task asked the participant to categorize the stimuli word into two categories. The second task was similar to the first one by changing the stimuli word. The third task asked the participant to categorize the stimuli word into two categories, each category combined category words from the previous two tasks. Task four repeated task three procedures but with more repetitions of words, names, or images. The fifth task was similar to the first task with reversed the positions of categories. Task six repeated the content the task three, but the categories words were in opposite pairings. Similar to task four, the seventh task repeated task six with more repetitions of words, names, or images. The implicit bias was measured by the associations between categories and the stimuli word. For example, if the participant has more positive associations with the male group than with the female group, he or she will categorize “husband” and “hardworking” faster than “wife” and “hardworking”.

Word Embedding Association Test (WEAT) proposed by Caliskan et al. (2017) compared the similarity between two sets of target words and two sets of attribute words to determine which target words set was closer to an attribute word set. It was an extension of the IAT. To determine the association, a WEAT score was computed based on cosine similarity. WEAT was used to show that word embeddings trained on large corpora would replicate the biases measured by IAT. For example, "mother" was closer to "housework" than to "science", while "father" was closer to "science" than to "housework".

A cross-lingual version of WEAT (XWEAT) proposed by Lauscher and Glavaš (2019) is an extension of WEAT. Mean Average Cosine (MAC) similarity was another extension of WEAT. It modified WEAT for a multi-class setting (Manzini et al., 2019).

Based on WEAT, the Sentence Encoder Association Test (SEAT) proposed by May et al. (2019) used generated sentence embedding from pre-defined biased sentence templates to replace the word embedding. The new embedding association test was based on both WEAT and SEAT to assess social and intersectional biases at the contextual word-level (Tan & Celis, 2019). Since WEAT based measurements required manually picked target and attribute words list or pre-defined templates, it could prejudice the measurement by introducing selectivity bias. WEAT based measurements had focused on the measure at word-level and sentence-level, but document-level and corpus-level measurements were lacking. Though the latter two could be decomposed by using multiple sentence-level measurements to evaluate document-level bias, or multiple word-level measurements to evaluate corpus-level bias, it was still a need to develop a larger scale measurement to the growing data size.

2.2 Bias Detection

Various techniques had been proposed to detect bias, including quantitative and qualitative methods. Despite the type of target biases and the type of task, the methods could be summarized into three general approaches:

- Embedding detection
- Dataset detection

- Crowdsourcing detection

In the next three subsections, each approach would be introduced in detail.

2.2.1 Embedding detection

This approach identified a sub-space of a word embedding by defining n orthogonal unit vectors. These n unit vectors would form a sub-space. When there was only one unit vector, the subspace turned into a binary direction. Gender bias was usually defined to have binary values: “female’ or “male”. In the work by Bolukbasi et al. (2016), gender bias was measured by the distance between gender-neutral words and gender-specific words to remove bias within the word embedding. For example, if the distances between the gender-neutral word e.g. “housework” and gender-specific words e.g. “wife” and “husband” were different, the word embedding may contain gender bias. This method was claimed to remove gender bias within the word embedding, however, Bordia and Bowman (2019) found it did not completely eliminate bias in a word embedding. Instead, they extended the work by introducing a regularization loss term that minimizes the projection of encoder-trained embedding onto the gender subspace. Both methods evaluate the performance using the analogy, which is later proved not effective as expected (Nissim et al., 2020).

2.2.2 Dataset detection

Recent works had also focused on creating task-specific datasets for bias detection. Such datasets usually contained bias sentences or stereotypes that could be used to train models. For bias mitigation tasks, it had been proved that debias a model trained on biased data was not as effective as simply relabeling the data to remove existing biases (Zhou et al., 2021).

Data augmentation was an approach to balance the target groups in a dataset. Such a method had been applied to reduce the unintended bias without compromising overall quality (Dixon et al., 2018). Wino-gender Schemas, an extension of Winograd Schemas (Levesque et al., 2012), was proposed to reveal cases where coreference systems might be more or less likely to recognize a pronoun as coreference with a particular occupation based on pronoun gender

(Rudinger et al., 2018). It created two-sentence templates by referring to the pronoun with either the occupation or the participant. Each sentence template would have three gender pronoun instantiations and two participant instantiations (Rudinger et al., 2018). Wino-Bias was a benchmark dataset, created for coreference resolution tasks focusing on gender bias (Zhao et al., 2018). The dataset aimed to detect whether occupations relate more to one gender group over the other by exploring the stereotypical and anti-stereotypical relationships. Similar to Wino-Bias but focused on four domains (gender, profession, religion, and race), StereoSet was a dataset that measured the discriminatory behavior in a model while ensuring the underlying language model performance remained strong (Nadeem et al., 2020). The major issue was the datasets currently used to measure specific biases were too small to conclusively identify bias except in the most egregious cases (Ethayarajh, 2020).

2.2.3 Crowdsourcing detection

Since bias could be subtle and hard to detect, human intervention was often required to determine if the statements were biased or not. StereoSet was created by filling crowdsourcing data into the template blanks (Nadeem et al., 2020). Crowdsourced data was also utilized in many prior works to check if human interpretation aligned with the model results. For example, crowd-worker evaluation was used for comparison on analogy tasks (Bolukbasi et al., 2016).

Though this approach provided insightful perspectives, the data was also prejudiced. There was no control over the selection criterion of crowd-workers and therefore, various factors like cultural background could affect the overall consistency.

2.3 Topic Modeling

Topic modeling is a type of statistical modeling for discovering abstract “topics” that occur in a collection of documents. By examining a document with a particular topic, topic model would discover the words that appear in the document more or less frequently. For example, “car” and “driver” would occur more often in documents about traffic, “doctor” and “nurse” would occur in documents about the hospital. For documents with multiple topics, for example, 80%

about hospital and 20% about traffic, we would expect 4 times more hospital words than traffic words. By examining the documents, the topic model would cluster similar words together to form the "topics". It was often used for the discovery of hidden semantic structures in a text body. An early topic model was proposed by Papadimitriou et al. (2000) in 1998. Later in 1999, the Probabilistic Latent Semantic Analysis (PLSA) was introduced by Hofmann (1999).

2.3.1 Latent Dirichlet Allocation

Latent Dirichlet Allocation (LDA) was one of the famous topic modeling methods (D. M. Blei et al., 2003). It was a generative statistical model that represented a topic as a distribution over terms and document as a distribution over topics. LDA was the most common topic model currently in use. LDA topic model had been used to examine the gender bias in magazines by performing word-frequency analysis (Kozlowski et al., 2020). It introduced sparse Dirichlet prior distributions over document-topic and topic-word distributions, encoding the intuition that documents over a small number of topics and that topics often use a small number of words (D. M. Blei et al., 2003). Extensions of LDA such as Pachinko Allocation Model (PAM) (Li et al., 2012) improved by modeling correlations between topics, and additionally the correlations between words that constitute topics. PAM provided more flexibility and greater expressive power than LDA. Hierarchical Latent Tree Analysis (HLTA) (Liu et al., 2014) used a tree of latent variables and the states of those variables to model word co-occurrence. The word co-occurrence corresponded to soft clusters of documents as topics.

Unlike clustering technique, topic modeling treated topics as an admixture (D. M. Blei et al., 2003), also known as mixed membership models. As shown in Figure 2.1, each topic was defined as a distribution of terms and each term had a probability to be within a topic, and each document was defined as a mixture over topics, meaning each document can have multiple topics.

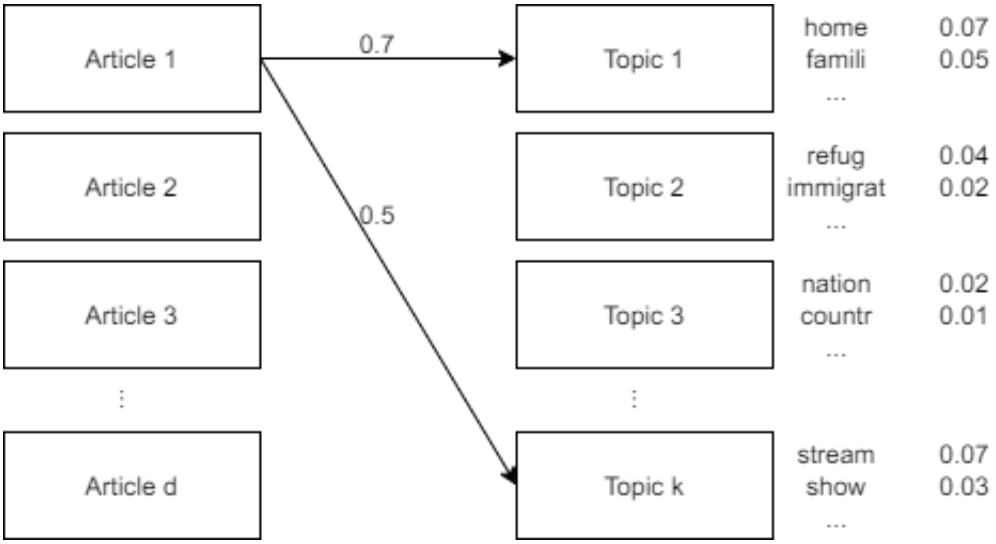

Figure 2.1. Mixed membership models

There were two primary matrices that topic models were concerned about. The first matrix was a $D \times K$ topic prevalence matrix, where D indicated the number of documents and K indicated the number of topics. As shown in Equation 2.1,

$$\theta = \begin{vmatrix} & Topic1 & Topic2 & ... & TopicK \\ Doc1 & 0.2 & 0.1 & ... & 0.05 \\ Doc2 & 0.2 & 0.1 & ... & 0.3 \\ ... & ... & ... & ... & ... \\ DocD & 0 & 0 & ... & 0.5 \end{vmatrix} \tag{2.1}$$

each row in topic prevalence matrix was a probability vector corresponding to each topic for a document. The second matrix was a $V \times K$ topic content matrix, where V was the size of vocabulary and K was the number of topics. Equation 2.2 shows that each row in topic content matrix is a probability vector of a term appearing in each topic.

$$\beta^T = \begin{vmatrix} & Topic1 & Topic2 & ... & TopicK \\ \text{“}text\text{”} & 0.02 & 0.001 & ... & 0.001 \\ \text{“}data\text{”} & 0.001 & 0.02 & ... & 0.001 \\ ... & ... & ... & ... & ... \\ \text{“}analysis\text{”} & 0.01 & 0.01 & ... & 0.0005 \end{vmatrix} \tag{2.2}$$

In LDA topic model, the topic prevalence matrix was drawn from uniform Dirichlet distributions with hyperparameters α and β. As shown in Figure 2.2, hyperparameters control the average shape and sparsity of the topic proportions, and a topic is drawn from $z_{d,n}$ and a term is drawn from $w_{d,n}$.

2.3.2 Semi-supervised Topic Model

The semi-supervised Topic Model allowed additional seed words into the estimation process. By feeding seed words into the model, they “forced” the model to assign a higher weight to the seeded terms if appeared in the documents. By increasing the weight to seeded terms, the model could discover the topics that are non-dominant or secondary in the documents. It was used

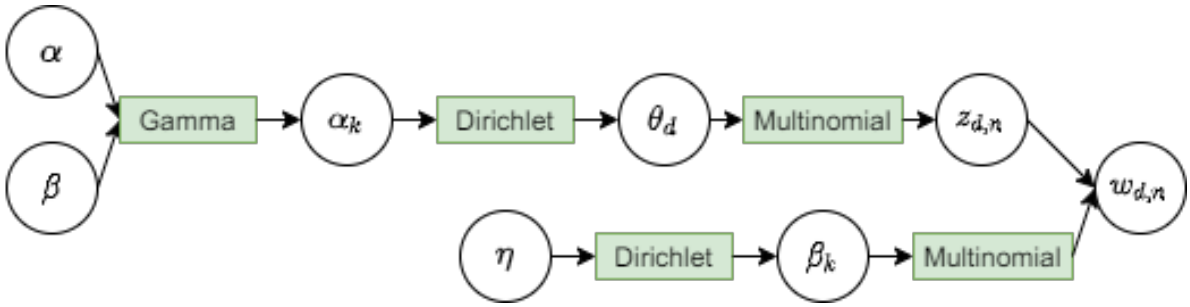

Figure 2.2. LDA generative process

with qualitative analysis to discover the gender bias in English and Swedish corpora (Devinney et al., 2020).

2.3.3 Correlated Topic Model

The correlated topic model (CTM) is a hierarchical model that explores the relationships between topics (D. Blei & Lafferty, 2006). It extends LDA by replacing the Dirichlet distribution with a logistic normal distribution and incorporating a covariance structure between topics (D. Blei & Lafferty, 2006). Such extension allows CTM to explore the topic correlations with topic proportions.

2.3.4 Structural Topic Model

There are several limitations of the traditional LDA topic modeling approach. Firstly, it restricts the topics to be independent. If a document has topic 1, then topic 1 does not provide additional information about other topics. This would result in missing some topics that are vital to the documents. Secondly, it only relies on the document text without considering other information sources, gender as in this thesis. Thirdly, as a result of the second limitation, it uses identical topic words for a topic among different documents.

When constructing a topic model from a corpus of documents, we wanted to see different angles for the same topic. For example, “health sequela” and “virus protection” could both describe COVID vaccination but from different angles. In this work, we would like to explore both female voices and male voices about the same topic. Prior methods primarily analyzed unstructured data and used words within the document to infer its subject (Roberts, Stewart, Airoldi, et al., 2014), causing information within documents to be leveraged. Additionally, contextual information about documents was lacking in prior methods. Structural topic model (STM) addressed such issues by combining LDA and contextual information. It included information about documents from metadata in two ways: topic prevalence and topic content. Topic prevalence could show that, for example, Democrats talk more about education than Republicans. Topic content would tell the word choices, like Democrats are less likely to use the

word "life" when talking about abortion than Republicans (Roberts, Stewart, Airoldi, et al., 2014). With this contextual information, the models could improve estimation accuracy and qualitative interpretability (Roberts, Stewart, Airoldi, et al., 2014).

The structural topic model (STM) defines a data-generating process for each document. These data are then used to estimate the most likely values for the model parameters. STM begins with document-topic and topic-term distributions to generate documents. In Roberts et al. (2019), for each document d and a vocabulary of size V with K topics, the document-level attention to each topic was drawn from a logistic-normal generalized linear model given a vector of documents X_d:

$$\overrightarrow{\theta_d}|X_{d\gamma},\Sigma \sim LogisticNormal(\mu = X_{d\gamma},\Sigma) \tag{2.3}$$

where X_d is a $1 \times p$ vector, γ is a $p \times K$ coefficients matrix and $\sum$ is a $K \times K$ covariance matrix. Then the document-topic distribution represents each topic (k) is generated using the baseline word distribution (m), the topic specific deviation κ_k^t, the covariate group deviation $\kappa_{y_d}^c$, and the interaction between the previous two $\kappa_{y_{d,}k}^i$:

$$\beta_{d,k} \propto exp(m + \kappa_k^t + \kappa_{y_d}^c + \kappa_{y_{d,}k}^i). \tag{2.4}$$

m, and each κ_k^t, $\kappa_{y_d}^c$, and $\kappa_{y_{d,}k}^i$ are vectors with length V, and each entry is a word in the vocabulary. To assign words to topics, document-topic distribution is used for each term in the document ($n \in \{1,...,N_d\}$):

$$z_{d,n}|\overrightarrow{\theta_d} \sim Multinomial(\overrightarrow{\theta_d}). \tag{2.5}$$

Given the chosen topic, a term is draw from the topic:

$$w_{d,n}|z_{d,n},\beta_{d,k=z_{d,n}} \sim Multinomial(\beta_{d,k=z_{d,n}}). \tag{2.6}$$

STM incorporated social science to allow the addition of document-level metadata and covariates into the estimation process, which altered the prior distributions used by the model (Davidson & Bhattacharya, 2020; Lucas et al., 2015; Roberts, Stewart, Tingley, et al., 2014). Metadata was defined as the original information about the document, for example, the content, author, and author gender. By incorporating metadata, the model was able to discover different

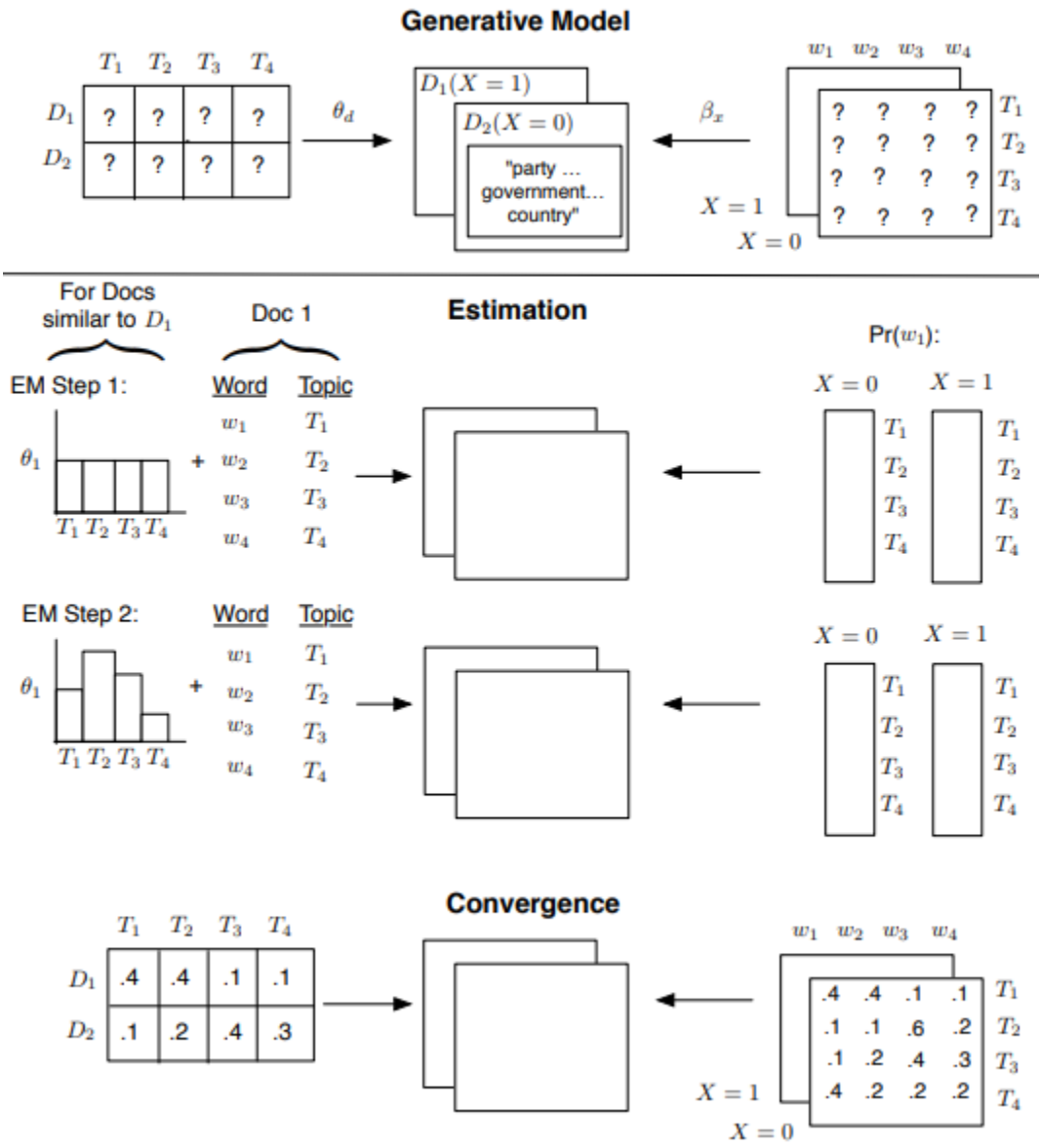

Figure 2.3. Generative process and estimation of the STM (Roberts et al., 2019)

angles about the same latent topic using different word choices. STM was used to estimate a probabilistic model that described the latent topics in a corpus of text with standard LDA (Davidson & Bhattacharya, 2020). Figure 2.3 provides a graphical representation of the STM.

There are three components used to estimate an STM, including documents, vocabulary, and metadata. Documents contain the distributions of topics for each document. Each topic is also a distribution of terms. Vocabulary records the index-term relationship. Metadata contains all text information within the original training data. It is incorporated into the STM estimation step

by two parts: topic prevalence estimation and topic content estimation. Both utilize the predefined covariates in topic prevalence estimation and topic content estimation. During topic prevalence estimation, covariates allow the observed metadata to affect how frequently a topic is discussed. During topic content estimation, covariates allow the observed metadata to affect how a topic is discussed. Topic prevalence (left-hand side of Figure 2.3) refers to the proportion of a document is linked with a topic and topic content (right-hand side of Figure 2.3) refers to the words that describe the topic.

CHAPTER 3. METHODOLOGY

This chapter will provide a brief review of the methodology employed to uncover associations between the latent topics and the gender of authors. We combined qualitative as well as quantitative evaluations to assess the difference between gendered topics. Text from the dataset was cleaned to make it more interpretable for topic modeling and analysis. The processed text was used to construct the corresponding structural topic model, exploring the latent topics within a corpus of documents. Top topic words were manually examined and aggregated into interpretable topics for human beings, and further split to examine the prevalence on different gender groups. Finally, gender-topic distribution and topic-gender distribution were extracted to provide more insights into the gender bias within new articles. Specifically, this chapter has the following order:

- Description of Corpus of documents
- Author gender classification
- Preprocessing of text
- Number of topics
- Training structural topic models
- Measurements
  - Quantitative results
  - Qualitative analysis

In the following sections, we will provide a detailed explanation of each step.

3.1 Description of Corpus of documents

The dataset used in this thesis was an online open-sourced dataset containing 2.7 million news articles and essays from 26 American publications. The American publications were specialized in various fields, including politics, business, technology, and so on. The data attributes include date, title, media outlet name, article text, year, month, and URL. For simplicity, we referred to American publications as media. Articles included in this dataset were published between 2013 and early 2020 in English. All the text was scraped by finding URLs in the sitemaps for each respective media and then scraped with Beautiful Soup version 4.8.2. Figure 3.1 shows the distribution of the number of articles published by each media and TABLE 3.1 shows the actual number of the documents.

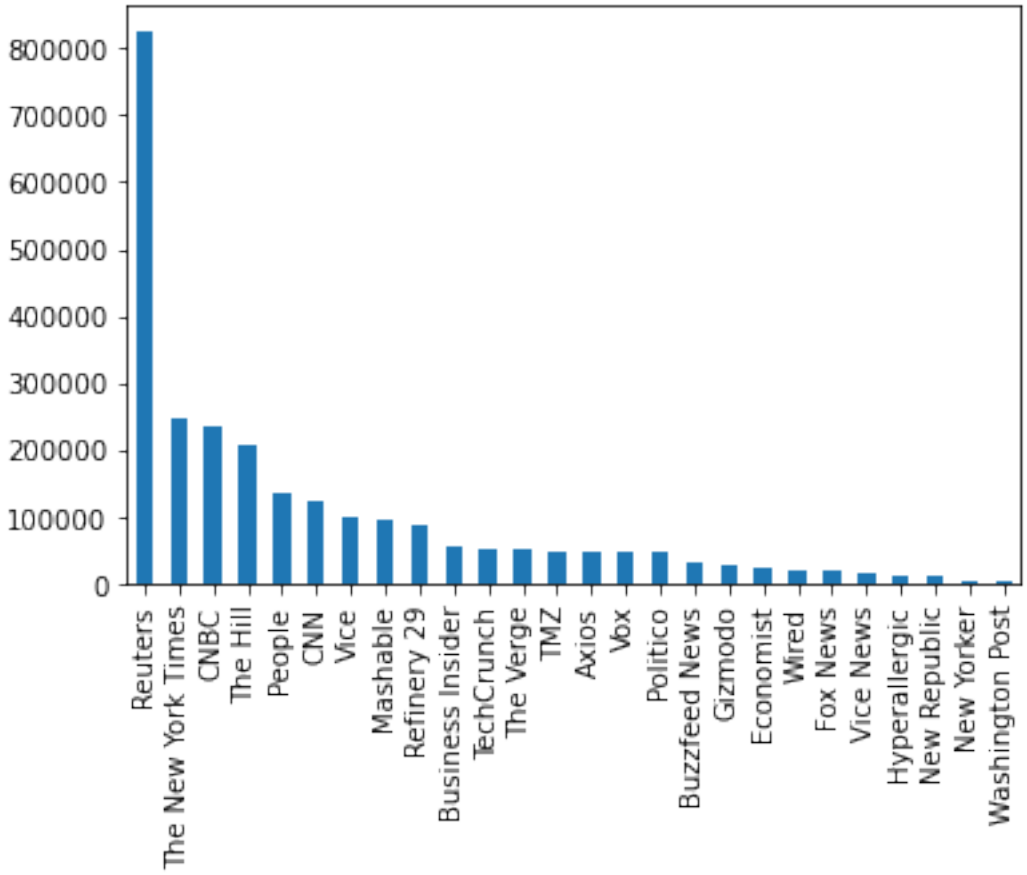

Figure 3.1. Distribution of articles published by each media

Table 3.1. *Actual number of articles published by each media*

Media	**# of Articles**
Reuters	825136
The New York Times	249086
CNBC	234141
The Hill	208411
People	135691
CNN	125300
Vice	100986
Mashable	94107
Refinery 29	87111
Business Insider	57934
TechCrunch	52051
The Verge	51556
TMZ	49504
Axios	47311
Vox	47265
Politico	46235
Buzzfeed News	32724
Gizmodo	27226
Economist	23200
Wired	20185
Fox News	20144
Vice News	15539
Hyperallergic	13539
New Republic	11807
New Yorker	4644
Washington Post	3332

Table 3.2. *Data Sample*

Date (str)	2016-12-09 18:31:00
Year (int)	2016
Month (int)	12
Day (int)	9
Author (str)	Lee Drutman
Title (str)	We should take concerns about the health of liberal democracy seriously
Article (str)	This post is part of Polyarchy, an independent blog produced by ...
URL (str)	https://www.vox.com/polyarchy/2016/...
Media (str)	Vox

3.1.1 Data Sample

A data sample can be found in TABLE 3.2. Each data entry contains a concatenated string format date of publication as well as the separate integer values corresponding to the year, month, and day. The author's full name is included and will be used for gender classification in the next step. The title and context of each article are included as text which needs to be cleaned. The rest attributes contain the media name and the URL.

3.1.2 Author Gender Classification

In order to understand how topics were gendered, we considered the gender of author as a necessary attribute within the training data. Not all author names are real names in this dataset. Figure 3.2 shows some author names sampled from the dataset. There were articles that have more than one author, based on common sense that the first author is responsible for the context, we kept only the first author's name when there were multiple authors of one article. Some author names were anonymous, for example, "WIRED Staff" or "Associated Press" which is not a real person's name. Some names were gender-neutral, like "Alex" or "Kyle" which could be either male or female. We defined both anonymous and gender-neutral names as invalid names. We defined valid names when they were gender-specific, like "Dave" or "Maria". Since invalid author names can mislead or even bias the gender classification, we removed data entries that had invalid author names and kept the ones that had valid author names.

To validate gender-neutral and gender-specific names, we used gender-guesser version 0.4.0, a Python package from Pypi, for author gender classification. Gender-guesser takes first name and country name as inputs and then outputs the predicted gender. For the inputs, the first name is required and the country name is optional. The output has six potential values depending on the confidence: unknown, andy (androgynous), male, female, mostly_male, and mostly_female. Specifically, "andy" means the name has the same probability to be male than to be female. We only focused on the most gender-specific names, therefore data entries whose author names were classified as unknown, andy, mostly_male, or mostly_female were removed from the training data.

TABLE 3.3 shows the number of valid author names before and after processing. 58.2% of the original unique author names were anonymous. After author gender classification, we retained 76.2% unique author names with high confidence on their gender predictions.

WIRED Staff	15815
Field Level Media	7930
Associated Press	7428
John Bowden	7021
The Associated Press	6344
Rebecca Savransky	5822
Julia Manchester	5448
People Staff	5393
Axios	5275
Dave Quinn	5017
Jordain Carney	4422
Alexia Fernandez	4321
Stephanie Petit	4177
Karen Mizoguchi	4165
Max Greenwood	4133
Sarah Perez	3749
Jen Juneau	3729
Brett Samuels	3708
Mark Hensch	3673
Aurelie Corinthios	3670
Maria Pasquini	3636
Tal Axelrod	3546
Kathryn Lindsay	3512
Joe Concha	3488
Char Adams	3459
Kaitlin Reilly	3423
Brian Heater	3241
Sam Haysom	3217
The Editorial Board	3160

Figure 3.2. Sample author names with the number of publications

Table 3.3. *Comparison of the number of unique author names*

Original Corpus		**Processed Corpus**	
# of unique names	**# of anonymous names**	**# of gender-unspecific names**	**# of gender-specific names**
140155	81573	14869	47713

3.2 Preprocessing of Text

Both title and article content was scraped from the web, therefore we preprocessed the documents. Since the goal was uncovering the latent topics within the documents, we only kept the Article content, Author name, and Author gender and removed Date, Year, Month, Day, Title, and URL. To preprocess the text, we tokenized all documents and converted tokens to lowercase, then removed punctuations, stopwords, and numbers, and finally stemmed all tokens.

Filtering out infrequent terms is needed as they provide little information, but it is computationally heavy to include them in the model (Davidson & Bhattacharya, 2020). To mitigate the impact that infrequent words have on topic models, we filtered out terms that occurred in fewer than 5 documents. TABLE 3.4 shows the detail of the training data after preprocessing.

3.3 Number of topics

Before the estimation process of the structural topic models, we needed to find the appropriate number of topics for all models. We defined a parameter K corresponding to the appropriate number of topics expected from the topic model. Since the number of documents of each media varies significantly, we implemented multiple structural topic models and tested with different K values ranging from 10 to 30. Then we evaluated the results using quantitative diagnostics provided by the stm package, and qualitatively examine the results. Figure 3.3 shows the quantitative diagnostics.

When words have been removed from the document in the estimation step, the document-completion held-out likelihood estimates the probability of those words appearing

Table 3.4. *Dataset detail after preprocessing*

Media	**# of documents**	**# of terms**	**# of tokens**
The Hill	160009	58656	34758974
Reuters	158447	87653	31662861
The New York Times	116909	112488	36520394
People	111823	74911	19251720
CNN	86222	45448	19678446
Vice	76065	72220	20484865
Mashable	72987	41879	10881009
Refinery 29	72736	41735	11594691
TechCrunch	43965	41829	8663901
The Verge	40759	34652	7194069
Vox	37392	46603	11494434
Axios	34285	17753	3369618
Buzzfeed News	24562	33383	5809703
Gizmodo	23939	31846	4571036
Fox News	9278	17080	1653961
Hyperallergic	9194	27190	2830688
New Republic	8921	24089	2543373
Business Insider	8001	14104	1408769
New Yorker	3617	27978	2096902
Wired	1355	9158	468246

within a document (Asuncion et al., 2012; Hoffman et al., 2013; Roberts et al., 2019; Wallach et al., 2009). It helps assess the model's prediction performance when some of the data is removed from estimation and then later used for validation. Like cross-validation, the held-out likelihood helps assess the model's prediction performance by removing some data from the estimation step and then later using the removed data for validation.

Residuals check the overdispersion of the variance of the multinomial within the data generating process of STM (Roberts et al., 2019). In previous work, Taddy (2012) described that residuals overdispersion meant more topics were needed to resolve some of the extra variances.

As a criterion, semantic coherence (Mimno et al., 2011) is closely associated with pointwise mutual information (Newman et al., 2010). When the top probable words tend to frequently co-occur together in a topic, semantic coherence will have maximized value. It had been shown that the topic quality of the metric aligns well with human judgment (Mimno et al., 2011). Formal representation of semantic coherence for topic k with M most probable words is shown as

$$C_k = \sum_{i=2}^{M} \sum_{j=1}^{i-1} log\Big(\frac{D(v_i, v_j)+1}{D(v_j)}\Big), \tag{3.1}$$

where $D(v, v')$ is the number of times that words v and v' co-occur in a document.

The change in the approximate variational lower bound indicates the model's convergence. The model is considered converged once a small enough change value of the lower bound between iterations is identified.

When K = 20, the model would converge well with high semantic coherence, as well as lower residuals overdispersion and better prediction performance. Overall, we decided to use K = 20 after the evaluations.

3.4 Training Structural Topic Models

Structural Topic Model (STM) is a generative model of word counts by using document metadata. It defines a data generating process for each document and then uses the data to find the most likely values for the parameters within the model. The generative model begins at the top, with document-topic and topic-word distributions generating documents that have metadata

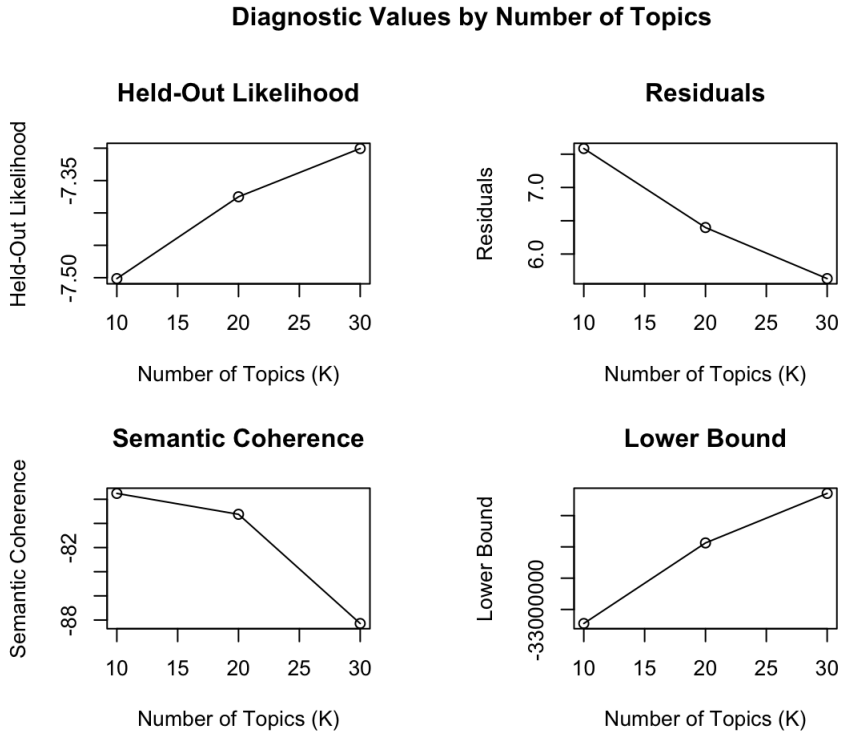

Figure 3.3. Choose the appropriate K value

associated with them. Within the framework, a topic is defined as a mixture of words where each word has a probability of belonging to a topic. And a document is a mixture of topics, meaning that a single document can be composed of multiple topics. The sum of the topic proportions across all topics for a document is one, and the sum of the word probabilities for a given topic is one. In this work, we used the stm package in R (Roberts et al., 2019) to estimate the STM. Further, to examine the topic prevalence on gender, we specified gender as a topic prevalence covariate when training the STMs.

The following content described the training step. A graphic representation of the whole pipeline can be found in Figure 3.4.

We used each media corpus of articles as training data for the corresponding topic model. Therefore there were 26 topic models corresponding to different media corpora. We separated each media corpus to avoid certain media dominating certain topics. Specifically, if one media was male prevalent on politics topic, then feeding all media corpora to train one topic model would minimize media that were female prevalent on politics topic, thus potentially falsify politics topic as stereotypical masculine.

We focused on the correlation between author gender and latent topics within articles. Therefore, during preprocessing of training data, we removed attributes that were less important like Date, Year, Month, Day, Title, and URL and only kept Article, Author, and Media. Then for each author, his or her gender was added into the training data. Specifically, we only added gender when the author name was gender-specific. Data entry whose author name was gender-neutral was removed from the training data. After the completion of gender classification, each article was tokenized and each token was converted to lowercase, then punctuations, stopwords, and numbers were removed from each token. During the stemming step, snowball stemmer was used to replace each token with the corresponding stem. Vocabulary terms could be dropped due to sparsity and stop word removal. For an article without any words after preprocessing, it was dropped from the training data and media with less than 1000 articles were dropped from the training data.

For each media corpus, a frequency threshold was identified by using the built-in function of the stm package (Roberts et al., 2019). The function was used to evaluate the minimum number of articles a term needed to appear in order to be kept within the vocabulary and the

optimal threshold was manually selected. If changes occurred during the preprocessing, the function re-indexed all metadata-article relationships. For example, an article that contained only rare terms was removed and its corresponding row in the metadata would be dropped as well (Roberts et al., 2019).

The output training data contained documents, vocabulary, and metadata. All documents, vocabulary, and metadata were used to train STMs. Author gender was used as a covariate in topic prevalence estimation and topic content estimation. During topic prevalence estimation, gender allowed the observed metadata to affect how frequently a topic was discussed. During topic content estimation, gender allowed the observed metadata to affect how a topic was discussed. Each model was set to run at most 100 iterations, but it could terminate early if converged.

3.5 Measurements

After preprocessing and frequency filter, we removed the corpora of articles from CNBC, TMZ, Politico, Economist, Vice News, and Washington Post, and there were 20 media corpora of articles left for training, including The Hill, Reuters, The New York Times, People, CNN, Vice, Mashable, Refinery 29, TechCrunch, The Verge, Vox, Axios, Buzzfeed News, Gizmodo, Fox News, Hyperallergic, New Republic, Business Insider, New Yorker, and Wired. Therefore the final analysis was based on STMs corresponding to 20 media corpora of articles. Since there are no appropriate quantitative metrics to measure the quality of topic words, we used the quantitative results generated by the model to assist qualitative analysis.

3.5.1 Quantitative Results

There were three quantitative results used for analysis. Estimated Topic Proportions result shows the proportion of documents across the corpus composed of each topic, ranking topic with higher proportion to the top and the one with lower proportion to the bottom. The second quantitative result is Topic Prevalence, which maps 20 topics with prevalence by gender. It is useful as gender has binary values. The third quantitative result, Topic Correlation, shows the correlation between each topic. Correlations were useful in qualitative analysis to examine the

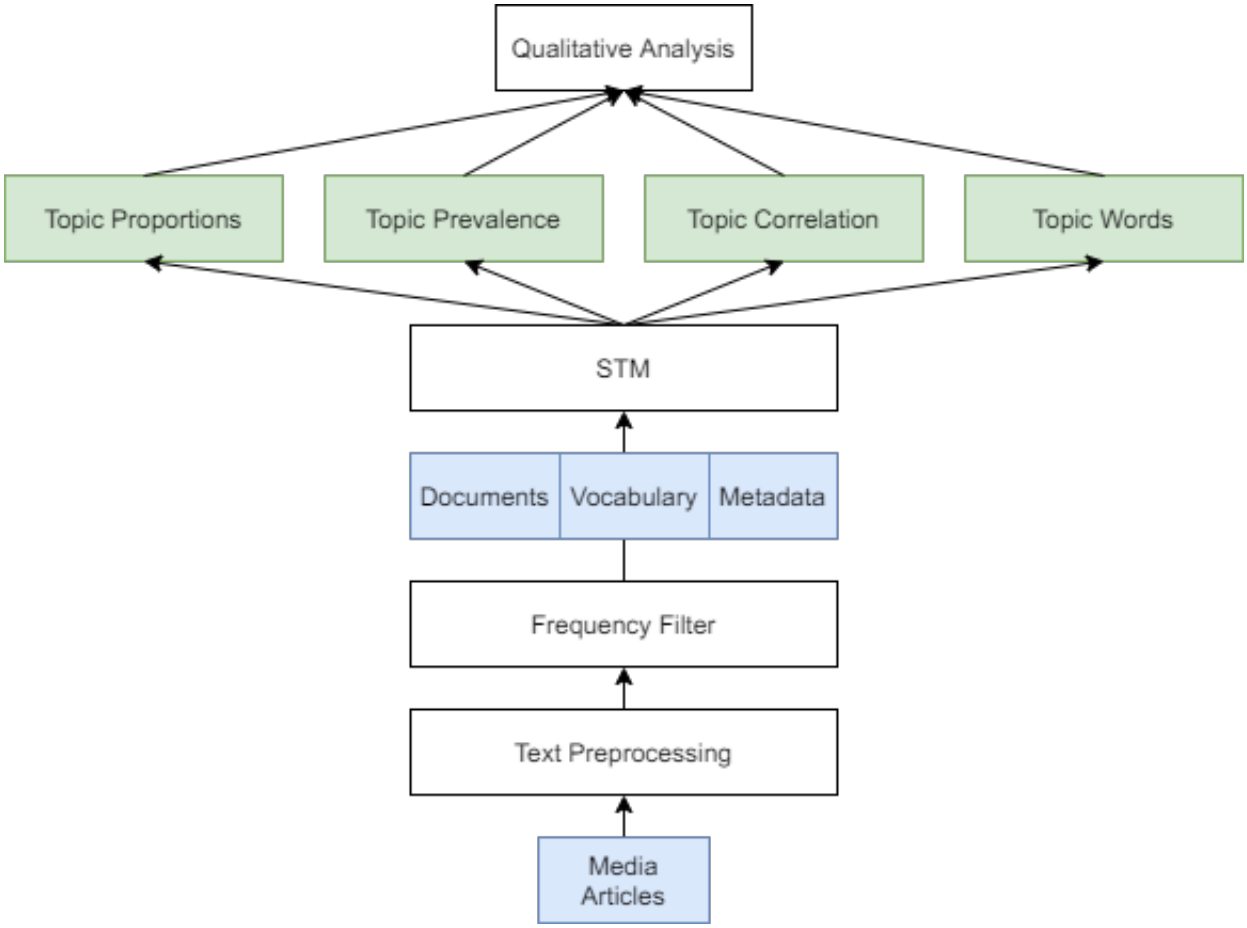

Figure 3.4. Pipeline

connection between two topics. Additionally, Topic Words were used to explore the words associated with each topic. There were four different types of word profiles used in Topic Words, including highest probability, FREX, Lift, and Score. FREX metric (Airoldi & Bischof, 2016; Bischof & Airoldi, 2012) weights words by overall frequency and the level of exclusivity to the topic.

$$FREX_{k,v} = \left(\frac{\omega}{ECDF(\beta_{k,v}/\sum_{j=1}^{K}\beta_{j,v})} + \frac{1-\omega}{ECDF(\beta_{k,v})}\right)^{-1} \tag{3.2}$$

ECDF is the empirical cumulative distribution function and ω is the weight which was set to 0.7 to favor the exclusivity (Roberts et al., 2019).

Lift weights words by dividing the frequency of the words in other topics, and giving higher weight to words that are less frequent in other topics (Roberts et al., 2019; Taddy, 2012). Similarly, for a word, Score divides the log frequency in the topic by the log frequency in other topics (Roberts et al., 2019). In actual experiments, Lift often provided less interpretable topic words than the other three, therefore we decided to only use the topic words generated by highest probability, FREX, and Score.

3.5.2 Qualitative Analysis

In order to answer whether there is gender bias in the news articles, as well as the factors that would impact the mitigation of gender bias, we set up several research questions for qualitative analysis, reflecting the gender inequalities and stereotypes.

1. What are the gender differences?
    - Are female prevalent topics associated with the feminine sphere?
    - Are male prevalent topics associated with the masculine sphere?
2. Do media have a balanced topic/gender distribution?

To determine the appropriate number of topic words for each topic, five experiments were performed by setting the number of topic words as 5, 10, 15, 20, 25. After the completion of training for each topic model, for each of the 20 estimated topics, we printed out the top 5, 10, 15, 20, or 25 topic words and compared how well they represented the topics. We found that more

topic words described the topic well but also contained synonyms, whereas fewer topic words did not contain synonyms but constructed the topic poorly. After manual examination, 10 topic words provided the optimal tradeoff. Therefore, we annotated each topic by looking at the top 10 words from each word profile, which resulted in a total of 30 topic words for each topic. Then based on the Topic Prevalence result and Topic Proportions result, we examined media/gender distribution and the topic/gender distribution. The Topic Prevalence result divided the topics into three groups by their corresponding weights.

In Figure 3.5, for each document X_d, the topic prevalence $\mu_{d,k}$ on k^{th} topic was computed by $X_d\gamma_k$ where γ_k followed a normal distribution $N(0,\sigma_k^2)$. In Roberts et al. (2019), σ_k^2 followed a Half-Cauchy distribution $Half-Cauchy(1,1)$, making the variance be close to zero as the mean of the normal distribution is zero. Such implementation made only the metadata that is highly correlated with a topic to be influential whereas non-influential coefficients would be close to zero (Lebryk, 2021).

Topic proportions θ_d for document d could then be correlated and the topic prevalence μ_d could be impacted by the covariate X, in this work the gender, through a regression model, $\theta_d \sim LogisticNormal(\mu_d,\Sigma)$. Then each document would have a prior distribution over topics instead of using a global mean (Hill, 2020).

Figures 3.6 and 3.7 showed two representative topic prevalences from CNN and New Republic, where the topic prevalence for CNN had little variance and the one for New Republic had higher variance. the X-axis of the topic prevalence represents the binary prevalence, in this word more male prevalent or more female prevalent and the Y-axis has no specific meaning. The numbers on the plot next to the data items represent the topics. Each of the 20 topics was plotted based on its corresponding prevalence on the gender covariate. Topics appeared on the left area meaning they appeared mostly in male author articles and on the right area meaning female authors mostly talked about them. Topics appeared on the middle line meaning they were covered equally by both gender groups. The topic prevalence for each topic was shown as a horizontal line with the middle point marked as a dot. The dot of each topic represents the mean prevalence value of each topic and the two endpoints of the horizontal line represent the maximum prevalences being covered by either gender group. The values of the dot and two endpoints could be found on the X-axis. If the horizontal line is short and does not cross the middle line, meaning the topic is

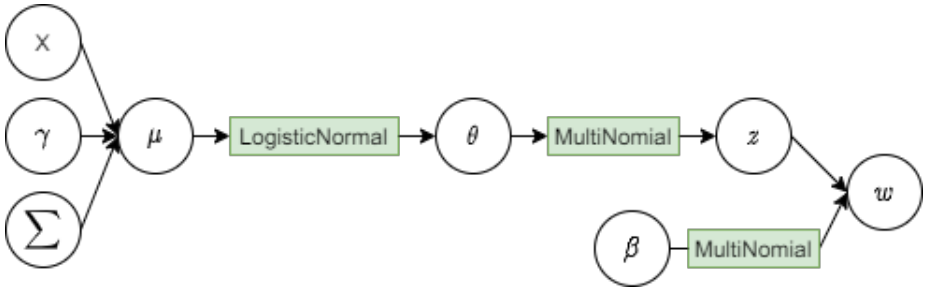

Figure 3.5. Structural topic model generative process

dominated by the gender group like topic 12 was dominated by female authors in CNN. Otherwise, if the horizontal line crosses the middle line, then the topic is mostly covered by one gender group, but the other gender group also has outputs on this topic, like topic 18 was covered almost evenly by both gender groups, but slightly more from the female side. Topics in CNN had little variances, whereas the variances of topics in New Republic were higher, as shown in Figure 3.7. The topics varied between each media outlet depending on the media corpus, therefore topic 9 from CNN could be different from topic 9 from New Republic.

For each media outlet, we ranked the 20 topics by the topic proportions and labeled each topic by its gender prevalence. Within the top articles from each of the last 10 topics, because the topic proportions were mostly less than 0.05, the last 10 topics were less influential on the article contents. Therefore, we decided to use the top 10 topics to represent each media. Sample topic proportions from CNN and New Republic are shown in Figures 3.8 and 3.9, as the horizontal bars show the proportions of CNN corpus of articles in each topic.

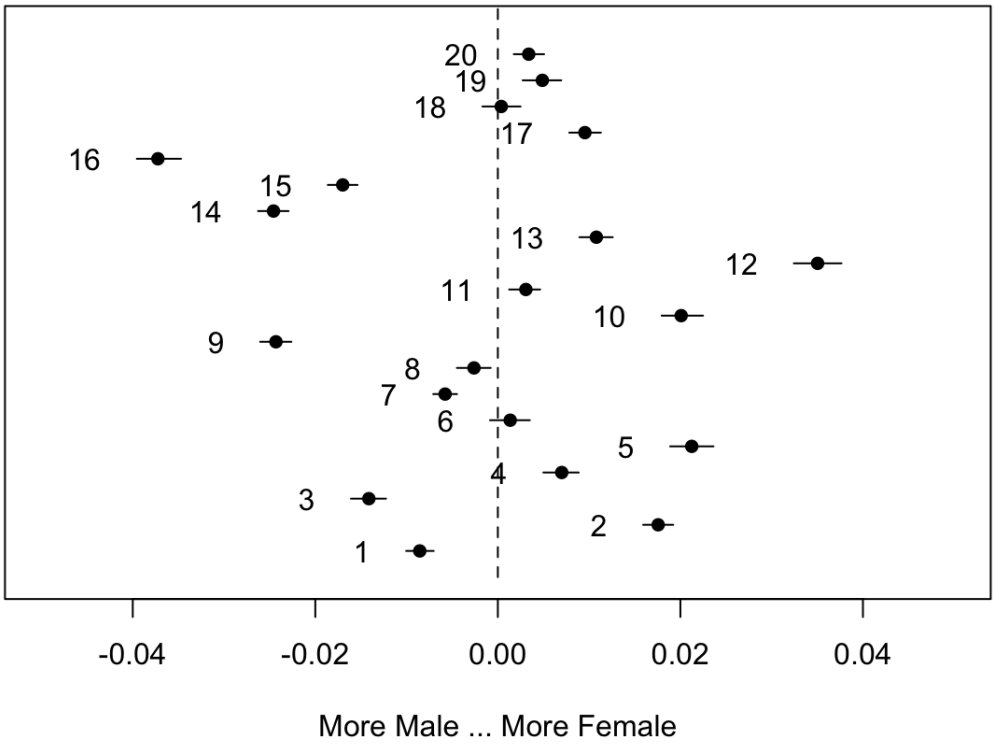

Figure 3.6. Topic Prevalence for CNN

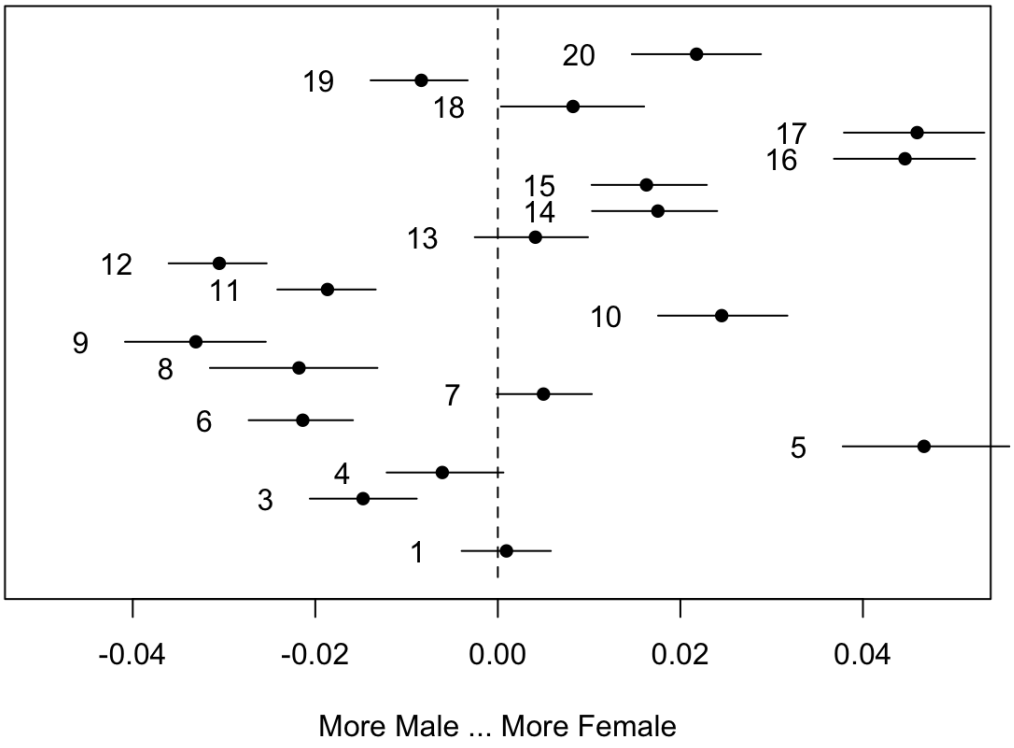

Figure 3.7. Topic Prevalence for New Republic

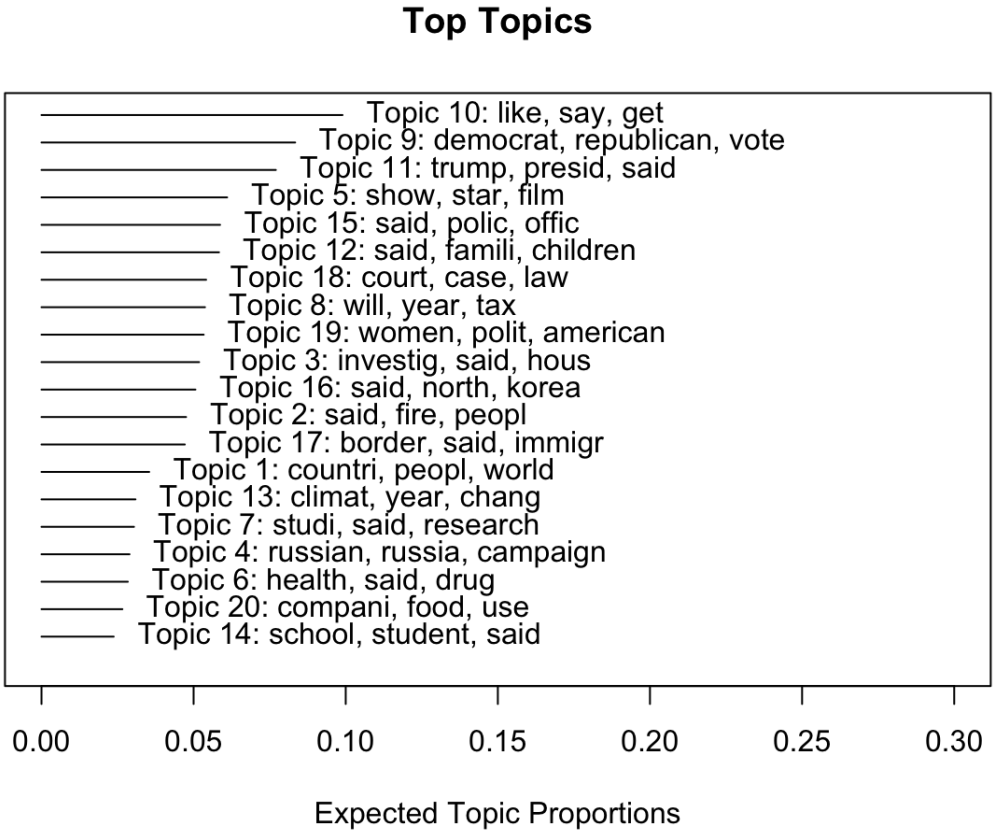

Figure 3.8. Topic Proportions for CNN

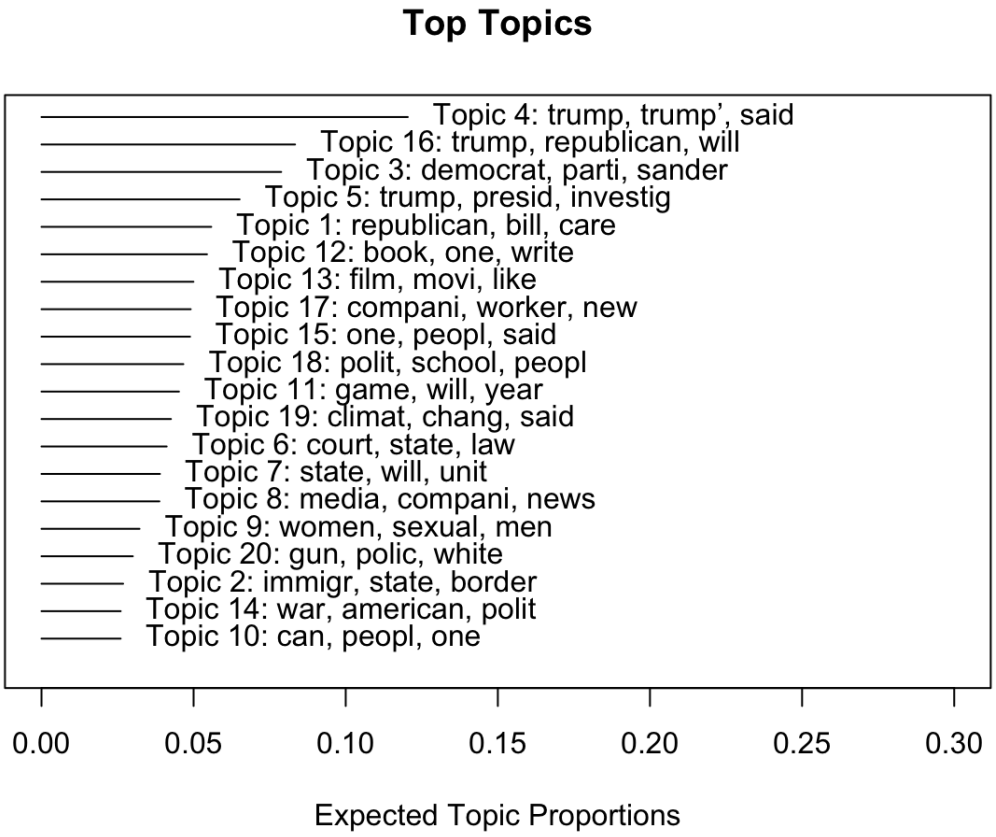

Figure 3.9. Topic Proportions for New Republic

CHAPTER 4. RESULTS

The analysis discovered both explicitly- and implicitly-gendered topics. Not all topics were aligned with the expected gender stereotypes, but we still found that topics mostly mentioned by male authors differed significantly from the ones mostly mentioned by female authors. The further examination went into exploring what factors would impact the mitigation of gender bias or stereotypes by exploring the topic/gender distribution of each media.

The corpora contained a wide range of different topics, which was expected given a large number of documents. Of the 36 generated topics, we were able to agree upon names for 20 topics by checking if the topics align with the article contents: International Politics; U.S. Politics; Law and Crime; Immigration; Market and Finance; Economy; Education; Technology; Health and Medication; Science; Social Media; Sports; Art; Movie; Music; Literature; Environment; Family and Relationships; Entertainment and Beauty. The remaining 16 topics were less intelligible, therefore we did not consider them in this analysis.

Across all corpora, male prevalent topics were mostly associated with the stereotypical masculine sphere, including International Politics (*state, russian, taliban, iran, nuclear*) , U.S. Politics (*presid, republican, democrat, senat, elect*) , Law and Crime (*polic, kill, shoot, death, court*) , and Sports (*team, game, player, season, playoff*) . Among the first three male prevalent topics, although male authors outputted the most articles, female authors also contributed a lot of views. But in Sports, female authors had fewer outputs compared to male authors, and this reflected the stereotype that Sports is more masculine.

On the other hand, female prevalent topics were mostly linked with the stereotypical feminine sphere: Music (*song, album, record, band, lyric*) , Literature (*literari, poem, book, harri, potter*) , Family and Relationships (*mom, dad, home, kid, home*) , Entertainment (*show, episod, podcast, disney, comedi*) , and Social Media (*youtub, snapchat, instagram, stream, post*) . Similarly, although Music and Family and Relationships mostly appeared in articles published by female authors, some male authors also had outputs about this topic, but such cases were rare and did not constitute enough quantifiable impact.

We found that topics with less than 0.05 proportions were less influential on the article content. In the best case, if the topic is neutral on gender prevalence, its gender prevalence value

on each group would be near 0.02. To understand the gender differences, we defined $K \times M$ matrix A to represent the topic association in each model by dividing the weights from Topic Prevalence result into three categories:

$$A_{k,m} = \begin{cases} weak & \text{if } 0 <= \frac{|\mu_{k,m}|}{0.02} < 1 \\ moderate & \text{if } 1 <= \frac{|\mu_{k,m}|}{0.02} < 2 \\ strong & \text{if } 2 <= \frac{|\mu_{k,m}|}{0.02} \end{cases} \tag{4.1}$$

where topics with weights from 0 to $|0.02|$ was defined as weak association, meaning the group did not outnumber the other group on the number of articles about the topic. From $|0.02|$ to $|0.04|$ was defined as moderate association, meaning the group published more articles about the topic than the other group. Topics with weights larger than $|0.04|$ was defined as strong association, meaning the group dominated the articles about the topic.

Another $K \times M$ matrix E was created to represent the topic-gender prevalence in each model:

$$E_{k,m} = \begin{cases} male & \text{if } \mu_{k,m} < 0 \\ female & \text{if } \mu_{k,m} > 0 \end{cases} \tag{4.2}$$

where $\mu_{k,m}$ indicated the topic prevalence weight of topic k in topic model m. Negative $\mu_{k,m}$ values indicated the topic was male prevalent and positive values indicated the topic was female prevalent. The co-occurrence of $A_{k,m}$ and $E_{k,m}$ was counted and the table in Figure 4.1 shows the results we used for interpretation. From the topic prevalence results, each cell indicated the times an association occur among all media corpora. Female authors were shown to have fewer articles on political topics. In International Politics, even though there were many associations linked with female authors, but most of them were weak associations. Even though there were some articles published by female authors that mentioned International Politics, but the same topic also appeared in many articles published by male authors. Contrast with female authors, male authors had less weak associations with International Politics, but their moderate and strong associations were more than that of female authors. The outnumbered associations indicated that male authors were more sensitive to International Politics and had more distinct points of view on International Politics than female authors. Such differences also appeared in the analysis of many media

corpora. In U.S. Politics, male authors had less strong associations than female authors, but their moderate and weak associations outnumbered those of female authors. We interpreted this as female authors were able to publish distinct views on U.S. Politics, but the publications were not consistent. On the other hand, male authors tended to have more outputs than female authors, but the differences were not significant. Similarly, of Law and Crime, female authors had some moderate associations but lack of distinct views as there was no strong association. Overall, U.S. Politics, International Politics, and Law and Crime aligned with stereotypical masculine representations.

There were topics that we expected to have gender differences but actually did not. Immigration was a topic closely related to politics and therefore we expected to see similar patterns from International Politics and U.S. Politics. However, male authors had little outputs while female authors produced more distinct outputs on Immigration. Among the articles with male prevalent Immigration topic, the contents were all about U.S.-Mexico border control, whereas the contents of articles with female prevalent Immigration topic talked about not only the border control, but also VISA, refuge, and immigration regulations. We looked at the top words from Immigration and found terms like *famli, child, home*. This provided a potential explanation that Immigration correlated with Family and Relationships, therefore female authors had more outputs than male authors. Both gender groups had relatively similar views of Market and Finance, but female authors had more outputs about the related topic Economy. Compared with Economy, Market and Finance are concerned more with companies and regulations, whereas Economy mentioned events like the employment rate. Considering both politics and law topics were male prevalent and Market and Finance mentioned regulations frequently, male authors would have more outputs on Market and Finance than on Economy, which was reflected in the results. Female authors, on the other hand, were concerned more about social events, thus had more outputs on Economy than on Market and Finance. Technology topic was a male prevalent topic, but we found that female authors also had the same amount of outputs, even had more distinct views than male authors. Within the articles published by male authors, most technology topics were about electronics, health devices, robots, self-driving techniques, and quantum. Female authors not only talked about previous content, but also discussed security, blockchain, smart devices, and applications of technology in psychology and education. Figure 4.2 shows the

	Strong		Moderate		Weak	
	Male	Female	Male	Female	Male	Female
International Politics	3	2	2	0	4	9
U.S. Politics	1	2	6	4	5	3
Law and Crime	1	0	4	5	5	3
Immigration	0	0	0	1	2	3
Market and Finance	1	1	3	2	2	3
Economy	0	1	0	1	2	1
Education	1	0	0	0	1	1
Technology	0	1	5	4	5	6
Health and Medication	0	0	3	3	7	8
Science	0	1	0	0	1	0
Social Media	0	1	0	3	3	0
Sports	1	1	3	0	1	1
Art	1	0	0	2	3	0
Movie	0	0	2	2	1	2
Music	0	0	0	0	3	3
Literature	0	0	1	1	0	4
Environment	1	0	3	0	4	2
Family and Relationships	0	2	0	1	2	1
Entertainment	0	0	0	4	5	3
Beauty	0	0	0	1	0	1

Figure 4.1. Topic association with each gender group

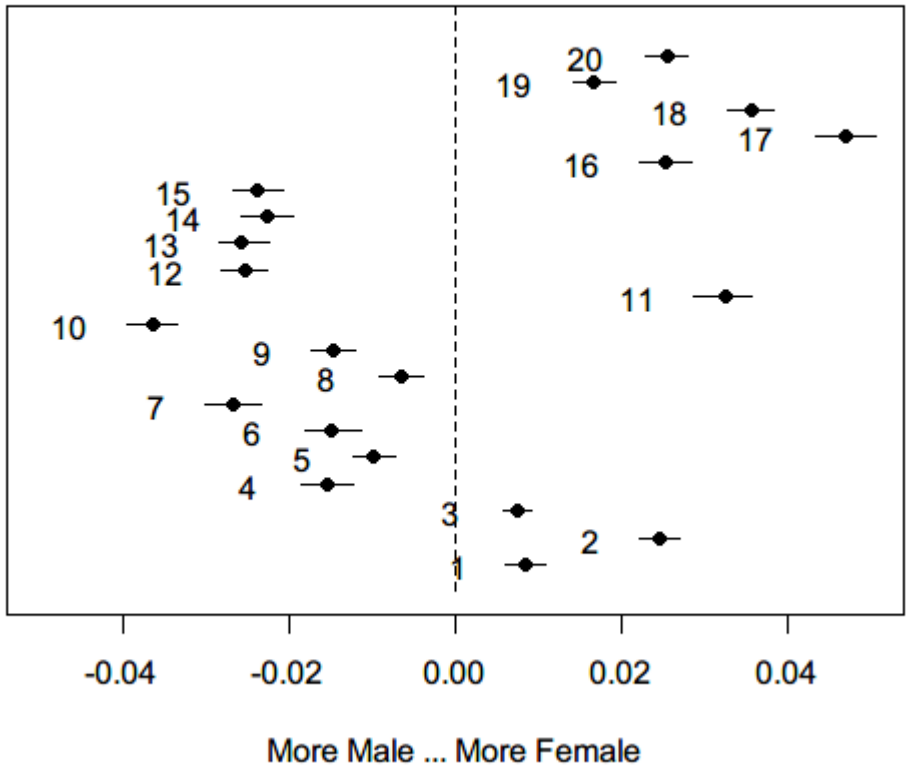

Figure 4.2. Topic prevalence for TechCrunch

topic prevalence for TechCrunch, a media focuses on technology reports. Topics 4, 9, 10, 12, and 16-20 were all about Technology, and female prevalent topics 16-20 were distinct from male prevalent topics 4, 9, 10, and 12 based on the context. Such a pattern could also be verified by the prevalence in Social Media. As a byproduct of Technology, female authors produced the majority of outputs in this area.

To further understand the gender bias in media corpora, we examined the topic/gender distributions from matrix E of each media by limiting K to the top 5 topics. Results were shown in TABLE 4.1.

Of all media, we first examined the topic/gender distribution of Technology to verify the previous result. Previous results showed that U.S. Politics, International Politics, and Law and Crime were male prevalent topics, but there were 7 out of 17 media had more female authors

writing about U.S. Politics than male authors, 4 out of 11 media had more female authors talked about International Politics than male authors and 5 out of 10 media that more female authors published articles about Law and Crime than male authors. Of Social Media, there were more articles from male authors than from female authors in 3 out of 7 media. There were 6 out of 8 media where male authors dominated the publications on Market and Finance, which is contrary to previous results that both gender groups provided the same views on this topic. We hypothesized the skewed gender distribution of each media could affect the topic/gender distributions. Skewed gender distributions would favor one gender group over the other gender group on topic prevalence.

In order to mitigate the gender imbalance impact on topic/gender distribution, we looked at the gender prevalence of each media to assist further interpretation on topic/gender distribution. TABLE 4.2 shows the gender prevalence by media.

After examining gender prevalence, we found a correlation between topic-gender prevalence and media-gender prevalence: 15 out of 20 media showed that male prevalent media would have male prevalence on the selected topics whereas female prevalent media would have female prevalence or equal prevalence on the selected topics. Specifically, in male prevalent media, articles about U.S. Politics, International Politics, Law and Crime were mainly published by male authors, and female prevalent media would have more articles about Entertainment and Social Media published by female authors. Technology was used as a baseline because articles were published equally by both gender groups.

4.1 Discussion

This work explored the relationship between author gender and the gender differences in a large corpus of unlabelled news articles using structural topic modeling. This section will discuss our interpretation of the results and answer the proposed research questions.

Table 4.1. *Media-gender prevalence on selected topics. "B" stands for Both gender groups. "-" means the topic did not appear in the media top 10 topics*

Media	**U.S. Politics**	**International Politics**	**Law and Crime**	**Entertainment**	**Social Media**
The Hill	M	M	-	M	-
Reuters	M	B	F	-	-
The New York Times	F	F	F	-	-
People	M	-	M	-	F
CNN	M	M	M	F	-
Vice	M	-	-	-	M
Mashable	F	-	M	M	-
Refinery 29	F	-	-	B	-
TechCrunch	-	-	M	-	M
The Verge	F	-	-	B	F
Vox	F	B	-	-	-
Axios	F	M	F	-	-
Buzzfeed News	B	M	F	-	F
Fox News	M	B	M	M	F
New Republic	M	F	-	-	-
Business Insider	M	F	F	F	-
New Yorker	F	F	-	F	-
Wired	M	-	-	-	M

Table 4.2. *Gender prevalence by media*

Media	Male	Female
The Hill	10	9
Reuters	11	8
The New York Times	7	8
People	10	8
CNN	8	10
Vice	7	5
Mashable	8	9
Refinery 29	10	5
TechCrunch	11	8
The Verge	5	12
Vox	9	9
Axios	7	11
Buzzfeed News	7	9
Fox News	9	10
New Republic	8	7
Business Insider	9	8
New Yorker	6	9
Wired	11	8

4.1.1 What are the gender differences?

The results showed distinct gender differences between female prevalent topics and male prevalent topics, as female authors preferred to talk about stereotypical feminine topics and stereotypical masculine topics were mostly covered by male authors. Among the 20 topics, female prevalent topics were Music, Literature, Family and Relationships, Entertainment, and Social Media which correspond to stereotypical feminine sphere, whereas male prevalent topics were found to be U.S. Politics, International Politics, Law and Crime, Sports, which can be linked with the stereotypical masculine sphere. Such gender differences among topic choices align with stereotypical gender bias. In this case, topics that are strongly associated with the stereotypical masculine sphere should appear to be male prevalent. However, topics like Immigration, Economy, and Technology did not align with prior gender differences. Immigration was slightly more female prevalent because it correlated with Family and Relationships: top words from Family and Relationships also appeared as top words in Immigration. The results showed that Market and Finance was neutral on gender prevalence. Economy was closely related to Market

and Finance, but because its content was concerned more about social events, it appeared to be female prevalent. Unlike other topics that were related to the stereotypical sphere, Technology did not have gender prevalence on either gender group. Both female and male authors provided an almost equal amount of opinions in Technology. There were signals that female authors had different views from male authors on International Politics, U.S. Politics, Economy, and Social Media, and male authors also had outputs on Art, Family and Relationships, and Entertainment. These signals were not consistent enough to mitigate the gender bias or stereotypes within news articles.

4.1.2 Do media have a balanced topic/gender distribution?

To understand if skewed gender distribution affects topic prevalence, we examined the topic/gender distribution in each media. The topic/gender distribution of Technology aligned with prior results, as there was the same number of media outlets where female publications outnumbered male publications or vice versa. Therefore, we set this distribution as a baseline and compared it with other topic/gender distributions. We found that almost 50% of the media had female authors dominated U.S. Politics, International Politics, and Law and Crime; almost 50% of the media had male authors dominated Social Media; 75% of the media had male authors dominated Market and Finance. These results did not align with the gender differences we had discovered.

Among the 7 media outlets that had more female authors talked about U.S. Politics, 6 media outlets were female prevalent; 50% of the media with more female authors discussed International Politics were female prevalent; 60% of the media with more female authors covered Law and Crime were female prevalent. Similarly, all media with more male authors wrote about Social Media were male prevalent. In media that were male prevalent in U.S. Politics, 70% were male prevalent. 50% of the media that were male prevalent in International Politics were male prevalent. 66.7% of the media that were female prevalent in Entertainment were female prevalent. 75% of the media that were female prevalent in Social Media were female prevalent.

4.2 Future Work

This work is a step towards understanding the formation and reproduction of bias or stereotypes within news article corpora from mainstream media. We used structural topic model to extract the intelligible topics from raw text. Future work can extend to combining embeddings and topic modeling technique which could preserve more semantic relationships between words. It is also important to apply this methodology to news articles prior to 2013 to learn how bias or stereotypes are in an early period. This will help to improve our understanding of the development of human bias or stereotypes.

CHAPTER 5. CONCLUSION

This work used the structural topic model to study the gender bias within news articles, for articles ranging from 2013 to early 2020 of 20 media. For each article, we retained the gender-specific author names for gender classification and discarded anonymous names as well as gender-neutral names. For each media, raw text was tokenized and stemmed in preprocessing procedure and a structural topic model was trained given the documents, vocabulary, and metadata. Three quantitative results were generated after the model training completion and the results were Topic Proportions, Topic Correlations, and Topic Prevalence. Along with the three quantitative results, 30 topic words from three word profiles were generated by the model and the three word profiles were: highest probability, FREX, and Score. Finally, the quantitative results and topic words were used for qualitative analysis to reveal any gender bias within the news articles.

Given 20 media corpora, 20 intelligible topics were manually named and selected for analysis. There were multiple gender differences observed in different topics. The notable ones were US Politics, International Politics, Immigration, Law and Crime, Market and Finance, Economy, Technology, and Social Media. 18 media were selected to analyze the notable topics. Gender differences appeared in US Politics, International Politics, and Law and Crime aligned with gender stereotypes as male authors have more published articles than female authors. Topics closely related to politics like Immigration, however, did not show similar gender bias. Technology did not show distinct gender differences among all media. It was perceived as a stereotypical male prevalent area, but the results indicated a balanced topic/gender distribution, meaning there is less gender bias. Similar results were expected for topics derived from Technology, like Social Media, but the topic/gender distribution showed a female prevalence signal. To further understand the author's gender's effects on gender bias, 5 most controversial topics were selected based on the topic association of each gender group, including US Politics, International Politics, Law and Crime, Entertainment, and Social Media. Gender prevalence and topic dominance of each media were analyzed. We found that male prevalent media tended to have more male authors talking about selected topics, while female prevalent media tended to more female authors discussing the selected topics. Combined with the previous analysis, gender

bias within the articles was the product of gender imbalance in media, as the male prevalent media had more articles published by male authors on controversial topics and so does the female group. But the reproduction of gender bias came from the author's choices. Female authors were found to lack distinct views on controversial topics, and they were more productive on feminine topics. But one notable topic, Technology, was found to be gender-balanced as there were equal views from both gender groups.

The methods of understanding bias in text were diverse, but many were lack of explainable procedure. This work presented a procedure to analyze bias using the structural topic model and also presented explainable results for understanding gender bias within news articles.

REFERENCES

Airoldi, E. M., & Bischof, J. M. (2016). Improving and evaluating topic models and other models of text. *Journal of the American Statistical Association*, *111*(516), 1381–1403.

Asuncion, A., Welling, M., Smyth, P., & Teh, Y. W. (2012). On smoothing and inference for topic models. *arXiv preprint arXiv:1205.2662*.

Bandura, A., Ross, D., & Ross, S. A. (1963). Imitation of film-mediated aggressive models. *The Journal of Abnormal and Social Psychology*, *66*(1), 3.

Bischof, J., & Airoldi, E. M. (2012). Summarizing topical content with word frequency and exclusivity. In *Proceedings of the 29th international conference on machine learning (icml-12)* (pp. 201–208).

Blei, D., & Lafferty, J. (2006). Correlated topic models. *Advances in neural information processing systems*, *18*, 147.

Blei, D. M., Ng, A. Y., & Jordan, M. I. (2003). Latent dirichlet allocation. *the Journal of machine Learning research*, *3*, 993–1022.

Blodgett, S. L., Barocas, S., Daumé III, H., & Wallach, H. (2020). Language (technology) is power: A critical survey of" bias" in nlp. *arXiv preprint arXiv:2005.14050*.

Bolukbasi, T., Chang, K.-W., Zou, J., Saligrama, V., & Kalai, A. (2016). Man is to computer programmer as woman is to homemaker? debiasing word embeddings. *arXiv preprint arXiv:1607.06520*.

Bordia, S., & Bowman, S. R. (2019). Identifying and reducing gender bias in word-level language models. *arXiv preprint arXiv:1904.03035*.

Butler, J. (2011). *Gender trouble: Feminism and the subversion of identity*. routledge.

Butler, J., & Lourties, M. (1998). Actos performativos y constitución del género: un ensayo sobre fenomenología y teoría feminista. *Debate feminista*, *18*, 296–314.

Caliskan, A., Bryson, J. J., & Narayanan, A. (2017). Semantics derived automatically from language corpora contain human-like biases. *Science*, *356*(6334), 183–186.

Crawford, K. (2017). *The trouble with bias.* Retrieved from `https://www.youtube.com/watch?v=fMym_BKWQzk&ab_channel=TheArtificialIntelligenceChannel`

Dahllöf, M., & Berglund, K. (2019). Faces, fights, and families: topic modeling and gendered themes in two corpora of swedish prose fiction. In *Dhn 2019, 4th digital humanities in the nordic countries 2019, university of copenhagen, copenhagen, denmark, march 6–8, 2019* (pp. 92–111).

Dastin, J. (2018). *Amazon scraps secret ai recruiting tool that showed bias against women.* Retrieved 2018-10-11, from `https://www.reuters.com/article/us-amazon-com-jobs-automation-insight-idUSKCN1MK08G`

Davidson, T., & Bhattacharya, D. (2020). Examining racial bias in an online abuse corpus with structural topic modeling. *arXiv preprint arXiv:2005.13041*.

del Teso-Craviotto, M. (2006). Words that matter: Lexical choice and gender ideologies in women's magazines. *Journal of pragmatics*, *38*(11), 2003–2021.

Devinney, H., Björklund, J., & Björklund, H. (2020). Semi-supervised topic modeling for gender bias discovery in english and swedish. In *Gebnlp2020, coling'2020–the 28th international conference on computational linguistics, december 8-13, 2020, online* (pp. 79–92).

Dixon, L., Li, J., Sorensen, J., Thain, N., & Vasserman, L. (2018). Measuring and mitigating unintended bias in text classification. In *Proceedings of the 2018 aaai/acm conference on ai, ethics, and society* (pp. 67–73).

Ethayarajh, K. (2020). Is your classifier actually biased? measuring fairness under uncertainty with bernstein bounds. *arXiv preprint arXiv:2004.12332*.

Foucault, M. (1990). The history of sexuality: An introduction, volume i. *Trans. Robert Hurley. New York: Vintage*, *95*.

Frye, M. (1983). *The politics of reality: Essays in feminist theory*. Crossing Press.

Gonen, H., & Goldberg, Y. (2019). Lipstick on a pig: Debiasing methods cover up systematic gender biases in word embeddings but do not remove them. *arXiv preprint arXiv:1903.03862*.

Greenwald, A. G., McGhee, D. E., & Schwartz, J. L. (1998). Measuring individual differences in implicit cognition: the implicit association test. *Journal of personality and social psychology*, *74*(6), 1464.

Hall, S. (1997). The work of representation. *Representation: Cultural representations and signifying practices*, *2*, 13–74.

Hill, C. (2020). *Topic models.* Retrieved 2020-04-17, from https://rpubs.com/chelseyhill/672546

Hoffman, M. D., Blei, D. M., Wang, C., & Paisley, J. (2013). Stochastic variational inference. *Journal of Machine Learning Research*, *14*(5).

Hofmann, T. (1999). Probabilistic latent semantic indexing. In *Proceedings of the 22nd annual international acm sigir conference on research and development in information retrieval* (pp. 50–57).

Huang, H. (2018). *Dominated by men.* Retrieved from https://fingfx.thomsonreuters.com/gfx/rngs/AMAZON.COM-JOBS-AUTOMATION/010080Q91F6/index.html

Kiritchenko, S., & Mohammad, S. M. (2018). Examining gender and race bias in two hundred sentiment analysis systems. *arXiv preprint arXiv:1805.04508*.

Kozlowski, D., Lozano, G., Felcher, C. M., Gonzalez, F., & Altszyler, E. (2020). Gender bias in magazines oriented to men and women: a computational approach. *arXiv preprint arXiv:2011.12096*.

Lamas, M. (2015). El género: la construcción cultural de la diferencia sexual. *El género*, 1–348.

Lauscher, A., & Glavaš, G. (2019). Are we consistently biased? multidimensional analysis of biases in distributional word vectors. *arXiv preprint arXiv:1904.11783*.

Lebryk, T. (2021). *Introduction to the structural topic model (stm).* Retrieved 2021-04-17, from https://towardsdatascience.com/introduction-to-the-structural-topic-model-stm-34ec4bd5383

Levesque, H., Davis, E., & Morgenstern, L. (2012). The winograd schema challenge. In *Thirteenth international conference on the principles of knowledge representation and reasoning.*

Li, W., Blei, D., & McCallum, A. (2012). Nonparametric bayes pachinko allocation. *arXiv preprint arXiv:1206.5270*.

Liu, T., Zhang, N. L., & Chen, P. (2014). Hierarchical latent tree analysis for topic detection. In *Joint european conference on machine learning and knowledge discovery in databases* (pp. 256–272).

Lucas, C., Nielsen, R. A., Roberts, M. E., Stewart, B. M., Storer, A., & Tingley, D. (2015). Computer-assisted text analysis for comparative politics. *Political Analysis*, *23*(2), 254–277.

Maass, A., & Arcuri, L. (1996). Language and stereotyping. *Stereotypes and stereotyping*, 193–226.

Manzini, T., Lim, Y. C., Tsvetkov, Y., & Black, A. W. (2019). Black is to criminal as caucasian is to police: Detecting and removing multiclass bias in word embeddings. *arXiv preprint arXiv:1904.04047*.

May, C., Wang, A., Bordia, S., Bowman, S. R., & Rudinger, R. (2019). On measuring social biases in sentence encoders. *arXiv preprint arXiv:1903.10561*.

McCurdy, K., & Serbetci, O. (2020). Grammatical gender associations outweigh topical gender bias in crosslinguistic word embeddings. *arXiv preprint arXiv:2005.08864*.

Mehrabi, N., Morstatter, F., Saxena, N., Lerman, K., & Galstyan, A. (2019). A survey on bias and fairness in machine learning. *arXiv preprint arXiv:1908.09635*.

Mimno, D., Wallach, H., Talley, E., Leenders, M., & McCallum, A. (2011). Optimizing semantic coherence in topic models. In *Proceedings of the 2011 conference on empirical methods in natural language processing* (pp. 262–272).

Nadeem, M., Bethke, A., & Reddy, S. (2020). Stereoset: Measuring stereotypical bias in pretrained language models. *arXiv preprint arXiv:2004.09456*.

Newman, D., Lau, J. H., Grieser, K., & Baldwin, T. (2010). Automatic evaluation of topic coherence. In *Human language technologies: The 2010 annual conference of the north american chapter of the association for computational linguistics* (pp. 100–108).

Nissim, M., van Noord, R., & van der Goot, R. (2020). Fair is better than sensational: Man is to doctor as woman is to doctor. *Computational Linguistics*, *46*(2), 487–497.

Nussbaum, M. C. (1999). *Sex and social justice*. Oxford University Press.

Olson, P. (2018). The algorithm that helped google translate become sexist. *Last visited*, 3–12.

Papadimitriou, C. H., Raghavan, P., Tamaki, H., & Vempala, S. (2000). Latent semantic indexing: A probabilistic analysis. *Journal of Computer and System Sciences*, *61*(2), 217–235.

Park, J. H., Shin, J., & Fung, P. (2018). Reducing gender bias in abusive language detection. *arXiv preprint arXiv:1808.07231*.

Pryzant, R., Martinez, R. D., Dass, N., Kurohashi, S., Jurafsky, D., & Yang, D. (2020). Automatically neutralizing subjective bias in text. In *Proceedings of the aaai conference on artificial intelligence* (Vol. 34, pp. 480–489).

Roberts, M. E., Stewart, B. M., Airoldi, E. M., Benoit, K., Blei, D., Brandt, P., & Spirling, A. (2014). Structural topic models. *Retrieved May*, *30*, 2014.

Roberts, M. E., Stewart, B. M., & Tingley, D. (2019). Stm: An r package for structural topic models. *Journal of Statistical Software*, *91*(1), 1–40.

Roberts, M. E., Stewart, B. M., Tingley, D., Lucas, C., Leder-Luis, J., Gadarian, S. K., ... Rand, D. G. (2014). Structural topic models for open-ended survey responses. *American Journal of Political Science*, *58*(4), 1064–1082.

Rudinger, R., Naradowsky, J., Leonard, B., & Van Durme, B. (2018). Gender bias in coreference resolution. *arXiv preprint arXiv:1804.09301*.

Scott, J. W. (2015). El género: una categoría útil para el análisis histórico. *El género: una categoría útil para el análisis histórico*, 251–290.

Serret Bravo, E. (2006). *El género y lo simbólicola constitución imaginaria de la identidad femenina* (No. 305.420972 S4).

Shah, D., Schwartz, H. A., & Hovy, D. (2019). Predictive biases in natural language processing models: A conceptual framework and overview. *arXiv preprint arXiv:1912.11078*.

Stanovsky, G., Smith, N. A., & Zettlemoyer, L. (2019). Evaluating gender bias in machine translation. *arXiv preprint arXiv:1906.00591*.

Taddy, M. (2012). On estimation and selection for topic models. In *Artificial intelligence and statistics* (pp. 1184–1193).

Tan, Y. C., & Celis, L. E. (2019). Assessing social and intersectional biases in contextualized word representations. *arXiv preprint arXiv:1911.01485*.

Wallach, H. M., Murray, I., Salakhutdinov, R., & Mimno, D. (2009). Evaluation methods for topic models. In *Proceedings of the 26th annual international conference on machine learning* (pp. 1105–1112).

Zhao, J., Wang, T., Yatskar, M., Ordonez, V., & Chang, K.-W. (2018). Gender bias in coreference resolution: Evaluation and debiasing methods. *arXiv preprint arXiv:1804.06876*.

Zhou, X., Sap, M., Swayamdipta, S., Smith, N. A., & Choi, Y. (2021). Challenges in automated debiasing for toxic language detection. *arXiv preprint arXiv:2102.00086*.