Analyzing Gender Representation in News Media using Automated Methods

Sabina Hartnett

August 2022

MA Thesis for the Master's in Computational Social Science Program at the University of Chicago

Faculty Advisor: Jon Clindaniel

Preceptor: Sanja Miklin

Abstract

Through a multitude of platforms and sources, news media permeates online daily interactions. This reach affords news media significant social influence. Analyzing news articles at scale can reveal latent trends in news media, which ultimately have the potential to be norm-setting. In this study, we implement computational tools to reveal large scale trends in news reporting. Specifically, we integrate NER parsing, record linkage (in the form of gender prediction), topic modeling and word embeddings to reveal trends both in the corpus overall, as well as specific to gendered contexts. Named Entity Recognition and record linkage to isolate contexts in which an individual is reported on (and predict the individual's gender in order to make larger claims about gender representation in news media). These contexts are then used to train the word embeddings: illuminating differences in the semantic contexts and roles for women/men in news media contexts. This study contributes to an emerging field at the intersection between machine learning and quantitative social science by implementing advanced model architectures to answer questions based in cultural and media studies.

Keywords

Named entity recognition, LDA Topic Modeling, Word Embeddings, Machine Learning, Gender Stereotypes, Media and Communications Studies, Natural Language Processing, Computational Social Science.

Introduction

General Introduction / Project Motivation

News publications have long been the touchpoint between individuals and significant events, culture, and knowledge. Many users consume news media on a daily basis, relying on articles for political, social, cultural and entertainment updates (Flaxman et al., 2016). In the process of consuming news, an individual's perspective of the world and their respective role in it is influenced by that piece of information (or publisher). In this study, we explore gender representation in a diverse corpus of news media content in order to expose semantic contexts of gendered entities as they are referenced in the news.

This research is motivated the a growing availability of news media sources, a continuous increase in news consumption - with smartphones, push notifications, and the integration of many news outlets into social media feeds, internet and smartphone users are more exposed than ever before to news media content - and the sheer volume of shared news content that those two factors enable in combination (Ahlers et al., 2006). Computational tools, and the latest in natural language processing model architectures: including NER parsing, topic models, and word embeddings allow this study to make novel contributions to research on language, gender, and representation in news media.

Framing and Representation

News media is a central component in the cycle of large-scale ideology formation and norm-setting. There are many factors which influence the content and language that are ultimately included in news publications (*encoding* as coined by Hall (1973) and *framing* as conceptualized by Entman (2010)), which, in turn, influence a reader's conception of reality. External factors such as current events, governing bodies, socio-political contexts, profitability, and political affiliation can influence the content and rhetoric of a publication (Schudson, 1989). As articles are created and published by distinct entities, meaning is embedded within them as determined by the publishers. The text, as it is distributed in the form of an article, has thus been encoded by the producer. In mainstream news distribution, a producer will attempt to make their meaning the dominant one. This is achieved through widespread distribution and consistent messaging (encoding). Although this is often difficult to detect in qualitative content analysis, when trends in the organization of lexical items (language used in distributed media) are consistent, media producers *encode* articles with that trend.

Additionally, publications seek to construct a single, consistent *version* of reality for their readership by coherently encoding/framing (Entman, 2010; Hall, 1980). Messages in mainstream, distributed media have been stylized and conventionalized by a code of codes (‘dominant’ because they are conventional but subject to change and never fully deterministic) that has influence on readership and is detectable at scale. This consistent messaging shapes and informs a reader's interpretation of their own reality and can thus further propagate the goals which motivate frames/encoded messages (Gamson, 1992). Our analysis uses this theoretical framework to assume cultural content is transmitted through the semantics of news media as a cultural object, to reveal latent meanings in semantic consistencies as they relate to gender representation.

Of course, not all readers hold the same view of the world; news consumption and interpretation further refines readers' understanding of the world and their role in it. Readers actively engage with news media and interpret meaning from the text (in both content and semantics) as well as images that they consume. This *decoding*, in turn, influences the individual, their perspective of the world (and the news as a means of engaging with the world), and their actions within that context (Gamson, 1992; Hall, 2003). That is to say, the cycle of news production and consumption actively influences the parallel creation and maintenance of culture and society.

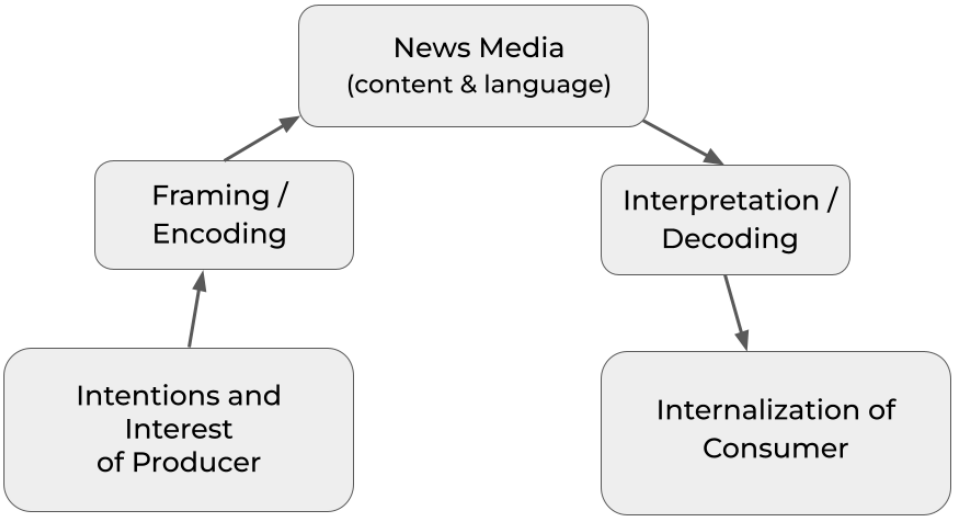


Figure 1. Framework of the cycle of production/
consumption used in this analysis, as inspired by Hall (1973).

Trends observed in news media are significant in their intentions (as *encoded* by the producers) as well as their potential impact downstream (in the reality framing and internalization by consumers). In this study, we specifically focus on the trends which emerge when investigating 'gendered - contexts' (text contexts in which individual entities of a single gender are referenced in news media). Specifically, is there a thematic or semantic difference in the way that men and women are referenced in news contexts?

Scholars have long used semantic analysis to discover cultural indicators and we will replicate some of the methods commonly used in qualitative content analysis using quantitative methods in this study (Collins, 2011; Gamson, 1992; Krippendorff, 2018). The central innovation of this study is both the scale at which we analyze news media, as well as the consideration and integration of part-of-speech tagging in our analysis. We reduce uncertainty in the word embedding space by isolating relevant contexts and training our own word embedding model.

Qualitative content analysis of news media has revealed distinct differences in gender representation in news contexts. In a 2021 study of nearly 10 years of Swiss news media articles, Vogler and Schwaiger revealed a discrepancy in the frequency with which women appear in news media (less frequently than men), as well as discrepancy in the context in which women and men appear. Women were found to be referenced in 'soft' news and culture contexts more often and men were more often referenced in politics, sports and economics (Vogler et al., 2021; Rao et al., 2021). Additionally, Vogler et al. found that women are less often cited as sources, and less frequently quoted as experts on topics. Similarly, Collins (2011) found that women are under-represented (compared to population size), are often portrayed in a subordinate manner and are highly sexualized in media contexts. Ultimately, qualitative studies of news media have concluded that women are portrayed as less significant to the dominant culture. Informed by the findings of these qualitative analyses, this study offers an approach for replicating qualitative semantic analysis with computational methods to identify differences in gender representation in news media at scale.

While there exist a number of computational text-based studies which investigate trends in language along cultural axes (Bolukbasi, 2016; Guo and Caliskan, 2017; Kozlowski et al., 2019), this study takes a unique approach to specifically isolate gender representation at a sentence level in news media content. By analyzing news media specifically, this study captures trends in language through a medium that is widely consumed, and can be interpreted through the encoding/decoding power structure of news distribution. Additionally, by isolating sentences in

which a gendered individual is the subject, this study specifically trains modern algorithms to 'close-read' texts for gender representation.

Ultimately, this study asks 3 primary questions:

1. Do men and women appear with different frequency in common news media?
2. In what contexts do men and women appear in common news media? Is there a discernible difference in the topics? How does this relate to preconceived stereotypes of each gender?
3. In the corpus of common news media, are there overall differences in semantic affiliations between men and women?

Data/Methods

Data Collection - News Articles

In order to study news media in a way that simulates public consumption, we consider a large and diverse dataset. Specifically, we sourced a dataset containing over 2.7Million articles, from 26 publication sources, with over 100 different sections (Thompson, *Components.* 2020). We then scaled our sampling to create a dataset representative of the distribution of the most commonly consumed news media sources, thus mimicking popular consumption (Statistica, June 2022). By using this sampling method, our findings and their interpretation are thus more robust and generalizable to the media-consuming public; simulating a *decoding* process that readers may experience (Adukia 2021; Hall 1980).

Corpus Exploration (EDA)

Publication	Monthly Readership (** statistica)	Frequency in Corpus	% represented	Altered frequency in corpus
Fox News	248M	20,144	18.4%	20,144
Washington Post	126M	40,882	9.33%	10,218
Business Insider	49M	57,953	3.6%	3,941
CNN	393M	127,602	29.11%	31,869
People	73M	136,488	5.4%	5,912
CNBC	109M	238,096	8.1%	8,868
The New York Times	352M	252,259	26.1%	28,573

Figure 2. Distribution of each publication in the dataset based on readership.

Shown above is the final, scaled representation for each publication within our dataset. For its readership level, Fox News was the relatively lowest sampled publication in the dataset. In order to build a representative corpus, we thus scaled all other frequencies to correlate with the relative frequency of Fox News. Our resulting corpus includes over 100,000 articles collected from 7 publications: diverse in political ideology, geographic location, and readership. The full dataset was collected between the years of 2016 and 2020.

NER Algorithm and Sentence/Context Selection

Recent advances in text (pre) processing include part of speech tagging, allowing users to identify, tag, and extract parts of speech within a sentence (within a corpus). These methods can

elevate text analysis far beyond traditional bag-of-words methods of analysis. For our study, we implemented part of speech (POS) tagging to identify the noun-subject and, if relevant, the direct object of each sentence in our corpus. These tags were then used, in conjunction with our gender prediction algorithm, to identify sentences in which individuals of each respective (predicted) gender are included as the subject and direct object (of the sentence), respectively.

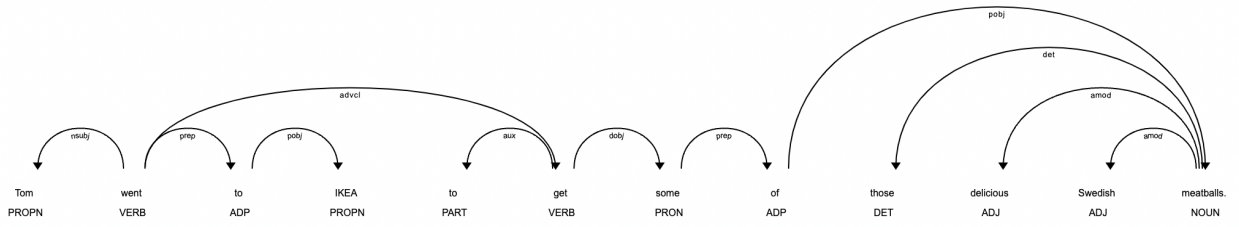


Figure 3. A visual representation of a sentence as POS tagged using SpaCy.

Gender Prediction Algorithm

News articles feature a number of different identity types: ranging from famous individuals, to niche experts, to local witnesses. Thus, in order to identify the gender of individuals (without assuming they have a public identity) identified using POS tagging, we created a gender prediction algorithm which takes a full name (of an individual referenced in text) and returns a predicted gender for that individual (male/female).

Our study classifies each sentence based on the gender of the identified subject (NER/POS). In order to complete this classification task, we collected a dataset of gendered names and built an algorithm to predict gender for each entity. In order to create a database of comprehensive names and respective gender, we collected the full database of the 1,000 most popular names (and

gender) for the years 1980-2021 from the Social Security Administration Website.[1] This dataset represents common names in the US in the last 40+ years, and should thus represent a large percentage of the names of individuals featured in American news media. Each name was annotated as the gender for which it had the highest 'ranking' over the 40 years (i.e. if 'Elizabeth' was the most popular female name in 1998 but the 100th most popular male name in 2010, it would be listed as female).[2] This approach includes all names from the entirety of the database - even those that are more gender ambiguous or infrequent.[3] This gender-prediction algorithm was then used to label the gender of entities as recognized by the NER algorithm (inspired by a method of gender prediction used by Adukia et al., 2021).

Label (predicted gender, tagged part of speech)	Number of Sentences Identified in Corpus
Female Subject	251,250
Male Subject	558,387
Female Direct Object	32,467
Male Direct Object	62,897

Figure 4. Count of each sentence-type label in the corpus..

The first and most straightforward goal of this study is to identify the frequency with which male and female entities appear in the corpus, respectively. Figure 4 displays the raw counts of male and female subjects and direct objects in the corpus. We can observe here that men are more than

[1] https://www.ssa.gov/oact/babynames/

[2] For simplicity of our study, with the acknowledgement that gender exists beyond the binary male/female divide, we will use 'male' and 'men' and 'female' and 'women' for grammatical consistency (as predicted by the names presented in the SSA dataset) since we cannot ever determine an individual's indetity or gender with full certainty (from afar) but the authors acknowledge that gender and sex are not interchangable.

[3] It is important to include infrequently occurring names because they might be member to minority populations and this study aims to include the representation of as many individuals as possible.

twice as likely as women to be the subject of a sentence - this correlates with larger trends in subject matter (that men are more likely to be featured as the subject of the article (Vogler et al., 2021)). Additionally, men are also approximately twice as likely to be featured as the direct object of a sentence, implicating that they are also integral to the action of the narrative. We are able to use the scale of our corpus to confirm that differences in representation persist in news media in a large and diverse set of articles (spanning many sections/topics and many publishers).

Topic Modeling

In order to investigate our second question: *in what contexts do male and female individuals appear in common news media?* We first reveal the (overall) common topics in our news corpus, by implementing topic modeling. Topic models were first introduced by Blei, Ng, and Jordan in 2003 and have since become a common method for corpus exploration and investigation. Allowing researchers to cluster text in a generative, probabilistic way, topic models help users interpret prevalent themes of large document collections. Topic models represent text corpora as clusters of words where each cluster is defined as a 'topic'. In our study we implement a Latent Dirichlet Allocation (LDA) Topic Model using the gensim package (Blei, 2003). This creates a probabilistic representation of the corpus as terms clustered into a set number of topics (here, 32), and each document as a distribution of topics. In implementing our topic models, we make the following assumptions:

1. Each document can be composed of multiple topics (an article may include terms from a range of topics; it is not the case that each article is a single or distinct topic). And thus, documents are represented by a probability distribution across topics.

2. A topic has a set vocabulary (terms may exist with varying weight across multiple topics but the terms which comprise a topic exist prior to the models training).

Topic models have been used in a range of applications, including to track consistency in messaging across texts and to group users based on online behavior (text posts) (Rosen-Zvi et al., 2004; Cha and Cho 2012). In our implementation, we then look at topic distribution across the two sub corpora, built to represent the male and female contexts within news media respectively, to investigate differences in the thematic contexts in which male and female subjects are featured.

Word Embedding Algorithms

Trends in widespread language use (and consumption) capture cognitive affiliations of their socio-cultural context. Thus, measuring co-occurrence patterns, as represented by high-dimensional vector representations of semantic contexts, allows us to interpret these trends at scale, simulating human interpretation of semantic context (*decoding* of the corpus). Word embeddings are a machine-learning implementation which can be trained on a specific corpus to represent the relationship between the words within that corpus. Kozlowski et al., 2019 implement word embeddings as geometric abstractions of the semantic meaning and socio-cultural implementation of words. Abstracted word similarity in a vector space models the cognitive affiliations humans create when interpreting and judging the world around them (Landauer et al., 1997). This transformation of text reveals new insights previously unattainable with manual content analysis and allows for a quantitatively measurable interpretation of word affiliations (Guo and Caliskan, 2017).

Results

The first step in exploring our dataset involved investigating the themes that are present in the full corpus. In order to identify themes in this large corpus, we used LDA Topic Modeling which learns topics (sets of important words) from text corpora in an unsupervised way. LDA Topic Models 'learn' the number of topics assigned, so, in order to create a Topic Model, we need to provide a number of desired topics. Recent research has sought to automate topic coherence and estimate the desired number of topics, rather than leaving the output to human evaluation (Lau et al., 2014; Röder er al., 2015). In our study, we evaluate the number of topics using the C_V coherence measure (which was found to be the best performing coherence measure when evaluated against human-labeled data (Röder er al., 2015)). By evaluating topics within the range of 20 - 36 we were able to identify the optimal number of topics (32) to represent the complexity of our dataset without inviting overly redundant or niche topics.

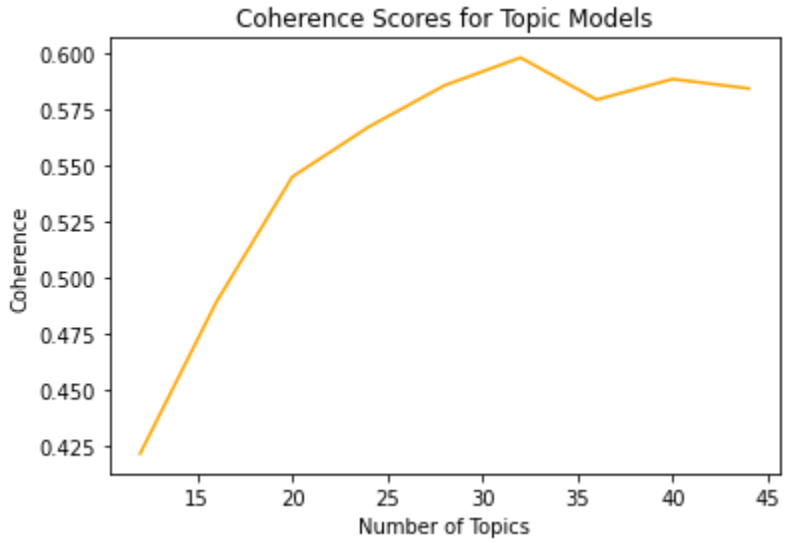


Figure 5. Coherence Score (using CV metric) for varying numbers of topics in gensim topic model (using mallet wrapper).

After determining the optimal number of topics using topic coherence metrics, we then input that number into our model to train on the full corpus of news articles. Training an LDA Topic Model on our corpus with 32 topics gave us a coherence metric of ~.6. We then used the package LDAvis to plot the topics on a 2-dimensional space (Sievert and Shirley, 2014).

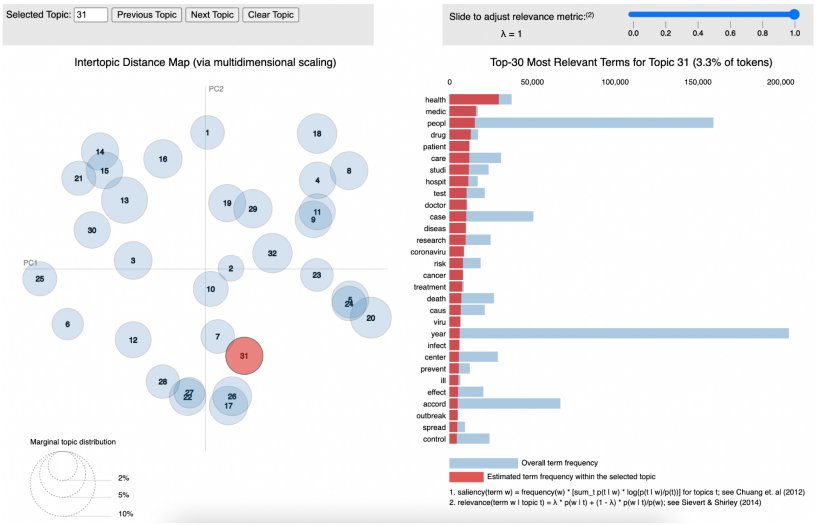


Figure 6. Image of a sample (31) topic as visualized in interactive LDAVis.

Each of the circles shown in Figure 6 represents one of the 32 distinct topics determined as a part of our model. The topics are mapped onto the plane using the first two components of Principal Component Analysis (their distance from each other in the space is significant in the first two principal components). LDAvis allows the user to interactively engage with the model (an html version of our interactive Topic Model can be found here) so we were able to visually interpret each of the topics and qualitatively confirm coherence.

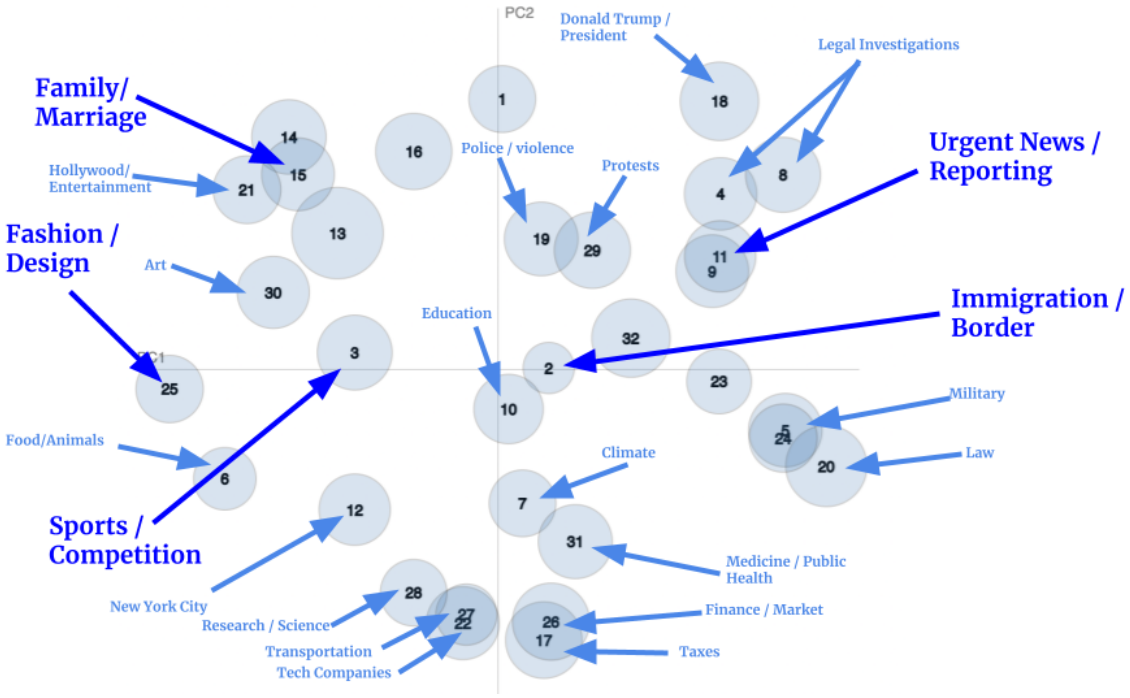


Figure 7. Labeled Topics (mapped by first two principal components).

After creating our Topic Model with the highest possible coherence score for this corpus, we qualitatively evaluated the topics. Figure 7 shows the manual annotations added after investigating the most significant terms for each topic (Blei, 2003). Even in the 'gender neutral' or less drastically affiliated topics, we can observe a natural clustering of similar themes in the model. The (human) interpretability of this model re-affirms that the model is both able to cluster terms which have a topic-affiliation as well as able to map them onto a reasonable semantic space in two dimensions.

After investigating the overall distribution of topics in the corpus (which is represented by the circle size in Figure 7), we wanted to explore the relevance of each topic in gender-specific contexts. Combining our gender prediction algorithm and part of speech tagging, we split the corpus into 2 sub-corpora: one which contains all sentences in which the subject of the sentence

is predicted to be female, and one in which the subject of the sentence is predicted to be male. The resulting sub-corpora thus contain all contexts in which individuals of each gender, respectively, are the subject.

To dive deeper into our driving question, *in what contexts do male/female subjects appear?* We investigate the topic distribution in each of the two sub-corpora. By isolating these two distinct categories, we can conduct an analysis on semantic representation of gendered individuals (in a tight context - full articles are too long and may contain too many contexts to derive meaningful interpretations about referenced individuals).

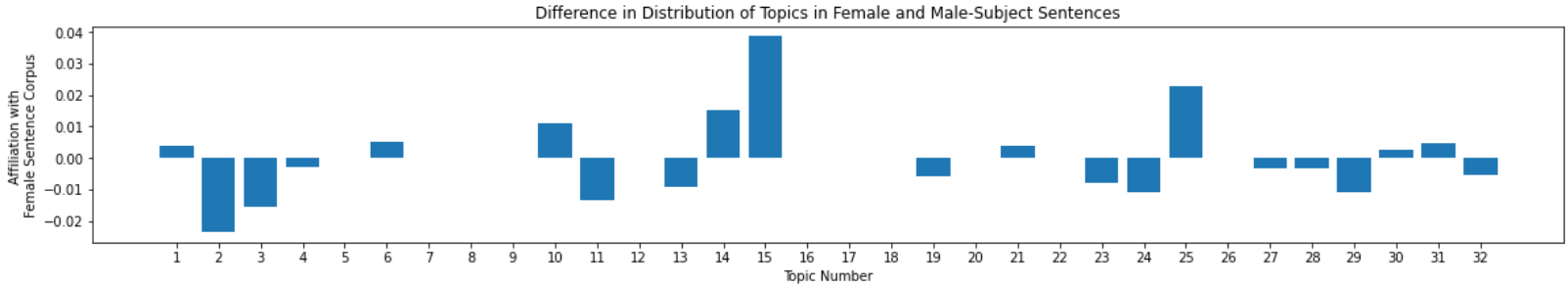


Figure 8. Distribution of Topics in Male and Female Subject Sentences.

Figure 8 plots the difference in the distribution of the topic between the female-subject and male-subject corpora. Here, topics with a more positive value have a higher distribution in the female-subject corpus, and topics with a more negative value have a higher distribution in the male-subject corpus. We can, thus, observe here which topics are most distinctly affiliated with gender-specific contexts. The most gender-polarized topics in this plot are: Topic 15 (female-subject-affiliated), Topic 25 (female-subject-affiliated), Topic 2 (male-subject-affiliated), Topic 3 (male-subject-affiliated), and Topic 11 (male-subject-affiliated). Shown below:

Topic Number	Gender Affiliation	Difference in Topic Frequency (frequency of topic appearance in female corpus - frequency of topic appearance in male corpus)	Topic Terms	Topic Label
15	Female	0.038724	Family, year, love, mother, father, life, children, son, daughter, friend, told, child, parent, couple, husband, home, wife, live, girl, kid, marry, time, sister, brother	Family/Marriage
2	Male	-0.023624	Immigrant, border, mexico, children, family, separate, migrant, illegal, mexican, cross, country, deport, asylum, unit, wall, agent, brazil, texas, undocu, detention, parent, secure, refuge, homeland	Immigration / Border
25	Female	0.022658	Design, wear, fashion, dress, color, hair, hand, style, cloth, black, beautiful, collect, body, model, brand, photo, watch, pair, shoe, face	Fashion/Design
3	Male	-0.015296	Game, team, play, player, season, year, win, sport, run, coach, point, time, league, score, final, won, start, world, fan, athlete, field, hit, week, football, lead, match, ball, open, cup, goal	Sports / Competition
11	Male	-0.013181	Report, cnn, told, according, statement, official, week, announce, release, [days of the week], contribute, ad, source, month, work, time, continue, member, comment, confirm, depart, march	Urgent News / Reporting

Figure 9. A closer look at the most (gender) polarized topics from Figure 8.

Figure 9 shows the details of the most gender-polarized topics in our dataset. Specifically, the topic number, the gender with which it most frequently appears (has a higher distribution), the magnitude of this difference (column 3), the most significant terms for that topic and the manually determined label for that topic. Here, we observe that the findings from Vogler's 2021 study are almost exactly replicated by our model. Our probabilistic model is trained using a randomized starting state and an unsupervised method and produces topics which almost exactly replicate the findings from Vogler's study which combined both manual and automated methods of analysis. We find that the topics Family/Marriage (15) and Fashion/Design (25) appear with

significantly more frequency in sentences with a female subject than those with a male subject. This finding signifies that in news media reporting, women are portrayed with significantly more frequency in contexts of family, love, appearance and home. Whereas, men are portrayed with significantly more frequency in contexts of urgency, significance, immigration, sports and competition; we find that the topics Immigation (2), Sports (3) and Urgent News (11) appear with significantly more frequency in sentences with a male subject than those with a female subject. In order to ensure that our corpora is not overly influenced by the disproportionate appearance of a single topic (or reporting on a single event), we analyzed the topic significance in each of the sub corpora over time.

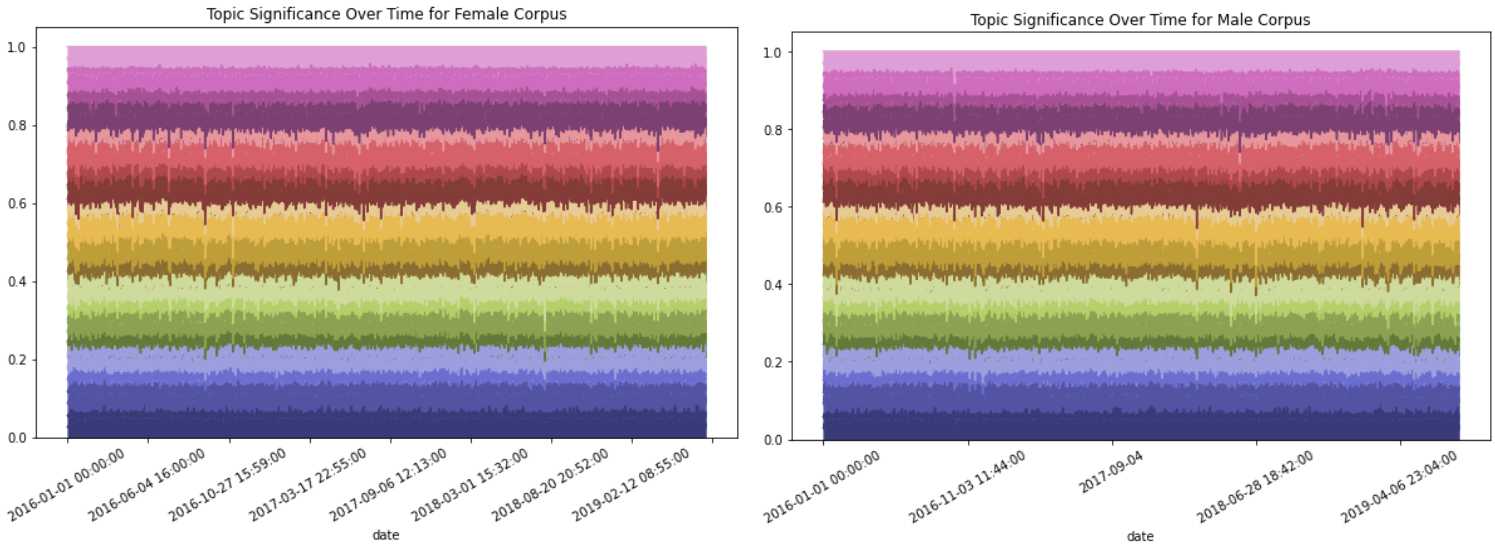


Figure 10. Topic distribution (significance) over time for subcorpora.

Figure 10 is an area plot of the 32 topics as they are distributed across the two sub corpora (Female and Male subject sentences respectively). Overall, the topics are relatively evenly distributed across the corpora and are robust over time. We can observe small spikes, which arise

from single events and change the contents/language of reporting temporarily (by changing the frequency of a topic's appearance), but the topics largely stabilize as time goes on. If a single event were to trigger a large change in topic representation (for example: if an event were to change the way news publications talk about (language used or frequency of topic use) of Topic 1, we would see the light purple at the top of the plot permanently altered from that point onwards), we would see an increase in the width of that topic's band in Figure 10.

Investigating topic clusters and their distribution across our two sub corpora gives a deeper understanding of the thematic trends that emerge in contexts with female and male subjects respectively. We then sought to investigate the direct semantic relationship between words in the corpus to reveal more latent trends. In order to further extrapolate semantic differences between the corpora, we then trained word embedding models on the full corpus and each of the two sentence-level corpora. Word embeddings vectorize a semantic space and are then able to calculate similarities between words based on distance within that space. Thus, biases and relationships between words in the training corpus can be revealed through distance metrics. Word embeddings have been used to reveal bias in a number of sources (including children's books and literature) both in snapshots and over time (Adukia et al., 2021; Bolukbasi, 2016; Guo and Caliskan, 2017). To look more closely at the semantic affiliations between gendered terms within our corpus, we used the similarity function provided by gensim's Word2Vec (which relies on cosine similarity and vector calculations) to find the nearest vector (term) based on positive and negative vector (term) affiliations (Guo and Caliskan, 2017; Kozlowski et al., 2019).

Corpus (Gender Subject)	Positive Affiliation	Negative Affiliation	Calculated Embeddings
Full Corpus	‘woman’, ‘she’	‘man’, ‘he’	**Couple**, **pregnancy**, human, **mother**, relapse, lovato, twin, **birth**, **sister**, **supermodel**, **pregnant**, **breastfeed**
Female	‘woman’, ‘she’	‘man’, ‘he’	Launch, **women**, earn, malaria, **athlete**, project, digit, **model**, **active**, **nonprofit**
Male	‘woman’, ‘she’	‘man’, ‘he’	**Girlfriend**, **wive**, **pregnant**, recal, quaalud, me, **cheat**, **breast**, **husband**, **mother**
Full Corpus	‘man’, ‘he’	‘woman’, ‘she’	**Loyalist**, playbook, paul, buck, **predecessor**, guilty, **loyalist**, **unpredictable**, **enemy**, **neutral**, **command**, foe, **leftist**
Female	‘man’, ‘he’	‘woman’, ‘she’	again, **trump**, upset, unforgiving, yell, anybody, remember, insane, knew
Male	‘man’, ‘he’	‘woman’, ‘she’	**Capital**, **country**, alliance, **loyalist**, strategies, partnership, increasingly, central, populist, robust, establish

Figure 11. Words most affiliated with gendered terms.

The terms included in the *Calculated Embeddings* column of Figure 11 are those which exist most closely to the location in the semantic space which results from combining the positive affiliation terms and subtracting the negative affiliation terms (we would expect, for example, that a positive affiliation with ‘actress’ minus a negative affiliation with ‘woman’ would equal ‘actor’ (the male equivalent)) (Bolukbasi, 2016). Thus, our calculations here are meant to reveal terms that are most unique to men and women conceptually.

For the word embedding model trained on the full corpus, we see terms similar to those in the family/marriage topic from our topic model appear. This indicates that not only do those terms appear frequently together, but they appear uniquely in contexts which do not feature men or

masculine-terms. We also find that the terms most similar to ‘man’ and ‘he’ (and negatively affiliated with ‘woman’ / ‘she’) are thematically consistent. This suggests that word embeddings are able to map terms, into thematic clusters, which reveal trends in gender affiliations.

Additionally, when investigating models trained on each of the two subcorpora, word embeddings reveal a clear distinction between the semantic affiliation of ‘woman’ and female pronouns between the two. In the corpus of female-subject sentences, women are affiliated with a range of professions and activities (*athlete, nonprofit, launch/project*) whereas, in the corpus of male-subject sentences, women are most closely semantically affiliated with their role in respect to men/family/reproduction (*girlfriend, wive, pregnant, husband, mother*). Whereas, the distinction between the two corpora is less clear for the semantic affiliation of ‘man’ and male pronouns. This is likely due to the smaller number of documents in each of the sub corpora and small size of each document (smaller and less consistent semantic space to train on).

Conclusion

In this study we used computational methods of analysis to replicate common methods in news media studies and distill gendered trends in a corpus of news articles. Specifically, we investigated the quantitative differences in appearances between men and women in the corpus and found that men appear as the subject of a sentence more than twice as often as women do in our corpus. We then approached a more qualitative analysis, exploring the topics which are more prevalent in contexts where women vs. men are the subject of a sentence. For this analysis we trained an unsupervised LDA model on the full dataset to determine relevant topics, then calculated the topic distribution of these overall topics in each of the sub corpora. We found, as

suggested by the literature, that women were more likely to be the subject of a sentence in which 'soft' news and culture were the primary topic and that men were more likely to be the subject of a sentence in which politics, sports, and news were the primary topic (Vogler et al., 2021; Rao et al., 2021). By producing these topics in an unsupervised (and randomized (Blei, 2003)) way, and then investigating their relationship to gendered contexts, we reveal that the qualitative findings of prior news media scholars persist in modern media and at scale. Finally, we investigated the full text corpus as a vectorized space using word embeddings and found similar relationships between gendered words and ideas to persist. Overall this study makes significant contributions to the field of media studies and the tools possible for recognizing differences in gender representation at scale.

Next Steps

The findings of this study suggest that there are a number of avenues in news media analysis that would benefit from large scale, reproducable analysis. Especially continuation off of our first question of investigation: *do men and women appear in news media with different frequencies?* Although our study looked at part of speech as a component of analysis, considering a larger context and the form in which the individual is referenced would continue with Vogler et al.'s 2021 finding that women were less often cited as experts or sources in news media. While we were able to identify that women are less often the *subject* and *direct object* of a sentence, future steps in this analysis could investigate the ways in which women are cited/referenced in the corpus to reveal if these observations persist in our dataset.

Additionally, the implications of gender bias and consistent gendered messaging are harmful to individual consumers and society at large. Understanding the impact of the messaging

propagated by news media would further contextualize the significance of these findings and continued research.

Acknowledgements

Thank you to Jon Clindaniel and Sanja Miklin for their patience and advising on this project!

Bibliography

Adukia, Anjali, Alex Eble, Emileigh Harrison, Hakizumwami Birali Runesha, and Teodora Szasz. What we teach about race and gender: Representation in images and text of children's books. No. w29123. National Bureau of Economic Research, 2021.

Ahlers, Douglas. "News consumption and the new electronic media." *Harvard International Journal of Press/Politics* 11, no. 1 (2006): 29-52.

Andrew Thompson. 2.7 Million news articles and essays. April 4, 2020. Distributed by Components. https://components.one/datasets/all-the-news-2-news-articles-dataset.

Blei, David M., Andrew Y. Ng, and Michael I. Jordan. "Latent dirichlet allocation." *Journal of machine Learning research* 3, no. Jan (2003): 993-1022.

Blei, David, and John Lafferty. "Correlated topic models." *Advances in neural information processing systems* 18 (2006): 147.

Bolukbasi, Tolga, Kai-Wei Chang, James Y. Zou, Venkatesh Saligrama, and Adam T. Kalai. "Man is to computer programmer as woman is to homemaker? debiasing word embeddings." *Advances in neural information processing systems* 29 (2016).

Cha, Youngchul, and Junghoo Cho. "Social-network analysis using topic models." In *Proceedings of the 35th international ACM SIGIR conference on Research and development in information retrieval*, pp. 565-574. 2012.

Chaney, Allison, and David Blei. "Visualizing topic models." In *Proceedings of the International AAAI Conference on Web and Social Media*, vol. 6, no. 1, pp. 419-422. 2012.

Chen, Qiuxing, Lixiu Yao, and Jie Yang. "Short text classification based on LDA topic model." In 2016 International Conference on Audio, Language and Image Processing (ICALIP), pp. 749-753. IEEE, 2016.

Collins, R. L. (2011). Content analysis of gender roles in media: Where are we now and where should we go?. *Sex roles*, *64*(3-4), 290-298.

Danescu-Niculescu-Mizil, C., Lee, L., Pang, B., & Kleinberg, J. (2012, April). Echoes of power: Language effects and power differences in social interaction. In Proceedings of the 21st international conference on World Wide Web (pp. 699-708).

Du Gay, P., Hall, S., Janes, L., Madsen, A. K., Mackay, H., & Negus, K. (2013). Doing cultural studies: The story of the Sony Walkman. Newbury Park, CA: Sage.

Entman, Robert M. "Framing bias: Media in the distribution of power." *Journal of communication* 57, no. 1 (2007): 163-173.

Entman, Robert M. "Media framing biases and political power: Explaining slant in news of Campaign 2008." Journalism 11, no. 4 (2010): 389-408.

Flaxman, Seth, Sharad Goel, and Justin M. Rao. "Filter bubbles, echo chambers, and online news consumption." *Public opinion quarterly* 80, no. S1 (2016): 298-320.

Gamson, W. A., Croteau, D., Hoynes, W., & Sasson, T. (1992). Media images and the social construction of reality. Annual review of sociology, 18(1), 373-393.

Gerbner, George. "Ideological perspectives and political tendencies in news reporting." Journalism Quarterly 41, no. 4 (1964): 495-516.

Guo, Wei, and Aylin Caliskan. "Detecting emergent intersectional biases: Contextualized word embeddings contain a distribution of human-like biases." In *Proceedings of the 2021 AAAI/ACM Conference on AI, Ethics, and Society*, pp. 122-133. 2021.

Hall, S. (1980). Encoding/decoding. In S. Hall, D. Hobson, A. Lowe, & P. Willis (Eds.), Culture, media, language: Working papers in cultural studies (pp. 128–138). London: Hutchinson.

Hall, Stuart. 1973. Encoding and decoding in the television discourse. Birmingham: Centre for contemporary cultural studies.

Krippendorff, K. (2018). Content analysis: An introduction to its methodology. Sage publications.

Kozlowski, Austin C., Matt Taddy, and James A. Evans. "The geometry of culture: Analyzing the meanings of class through word embeddings." *American Sociological Review* 84, no. 5 (2019): 905-949.

Lau, Jey Han, David Newman, and Timothy Baldwin. "Machine reading tea leaves: Automatically evaluating topic coherence and topic model quality." In *Proceedings of the 14th Conference of the European Chapter of the Association for Computational Linguistics*, pp. 530-539. 2014.

Mikolov, Tomas, Kai Chen, Greg Corrado, and Jeffrey Dean. "Efficient estimation of word representations in vector space." arXiv preprint arXiv:1301.3781 (2013).

Peirce, Charles Sanders. "Logic as Semiotic: TheTheory of Signs." The philo-sophical writings of Peirce. New York: DoverPress (1955).

Rao, P., & Taboada, M. (2021). Gender bias in the news: A scalable topic modeling and visualization framework. *Frontiers in Artificial Intelligence*, *4*.

Rehurek, R., & Sojka, P. (2011). Gensim–python framework for vector space modelling. *NLP Centre, Faculty of Informatics, Masaryk University, Brno, Czech Republic*, *3*(2).

Rosen-Zvi, Michal, Thomas Griffiths, Mark Steyvers, and Padhraic Smyth. "The author-topic model for authors and documents." *arXiv preprint arXiv:1207.4169* (2012).

Röder, Michael, Andreas Both, and Alexander Hinneburg. "Exploring the space of topic coherence measures." In Proceedings of the eighth ACM international conference on Web search and data mining, pp. 399-408. 2015.

Schudson, M. (1989). The sociology of news production. Media, culture & society, 11(3), 263-282.

Sender, Katherine, and Peter Decherney. "Stuart Hall lives: cultural studies in an age of digital media." Critical Studies in Media Communication 33, no. 5 (2016): 381-384.

Sievert, Carson, and Kenneth Shirley. "LDAvis: A method for visualizing and interpreting topics." In *Proceedings of the workshop on interactive language learning, visualization, and interfaces*, pp. 63-70. 2014.

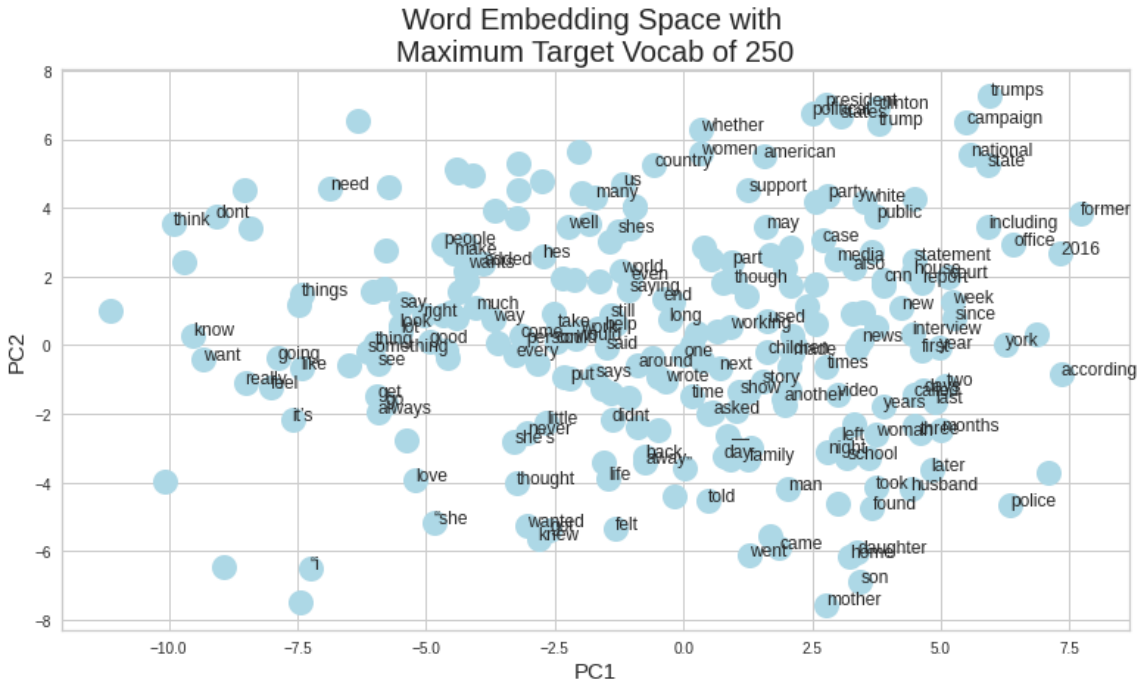
Vogler, D., & Schwaiger, L. (2021). Situational effects of journalistic resources on gender imbalances in the coverage of Swiss news media: A longitudinal analysis from 2011 to 2019. *Journalism*, 14648849211036309.

Watson, Amy. "Leading News Websites U.S. by Monthly Visits 2022." Statista, June 30, 2022. https://www.statista.com/statistics/381569/leading-news-and-media-sites-usa-by-share-of-visits/.

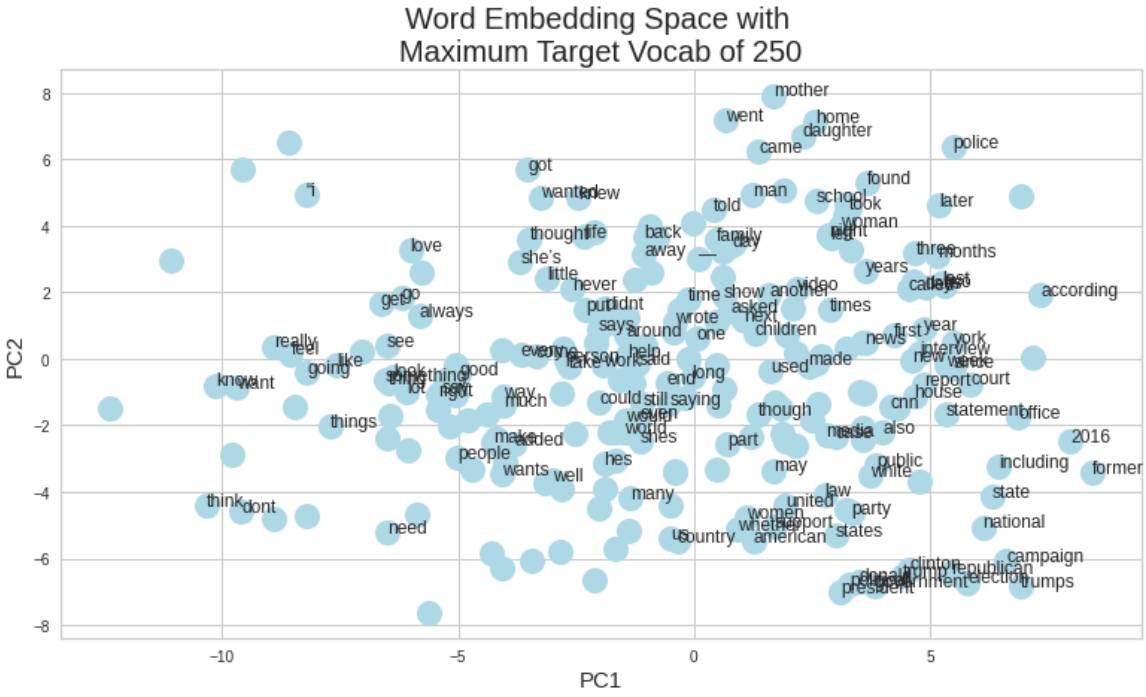
Appendix

Additional Figures

Word Embedding Visualizations



Female-Subject Sentences

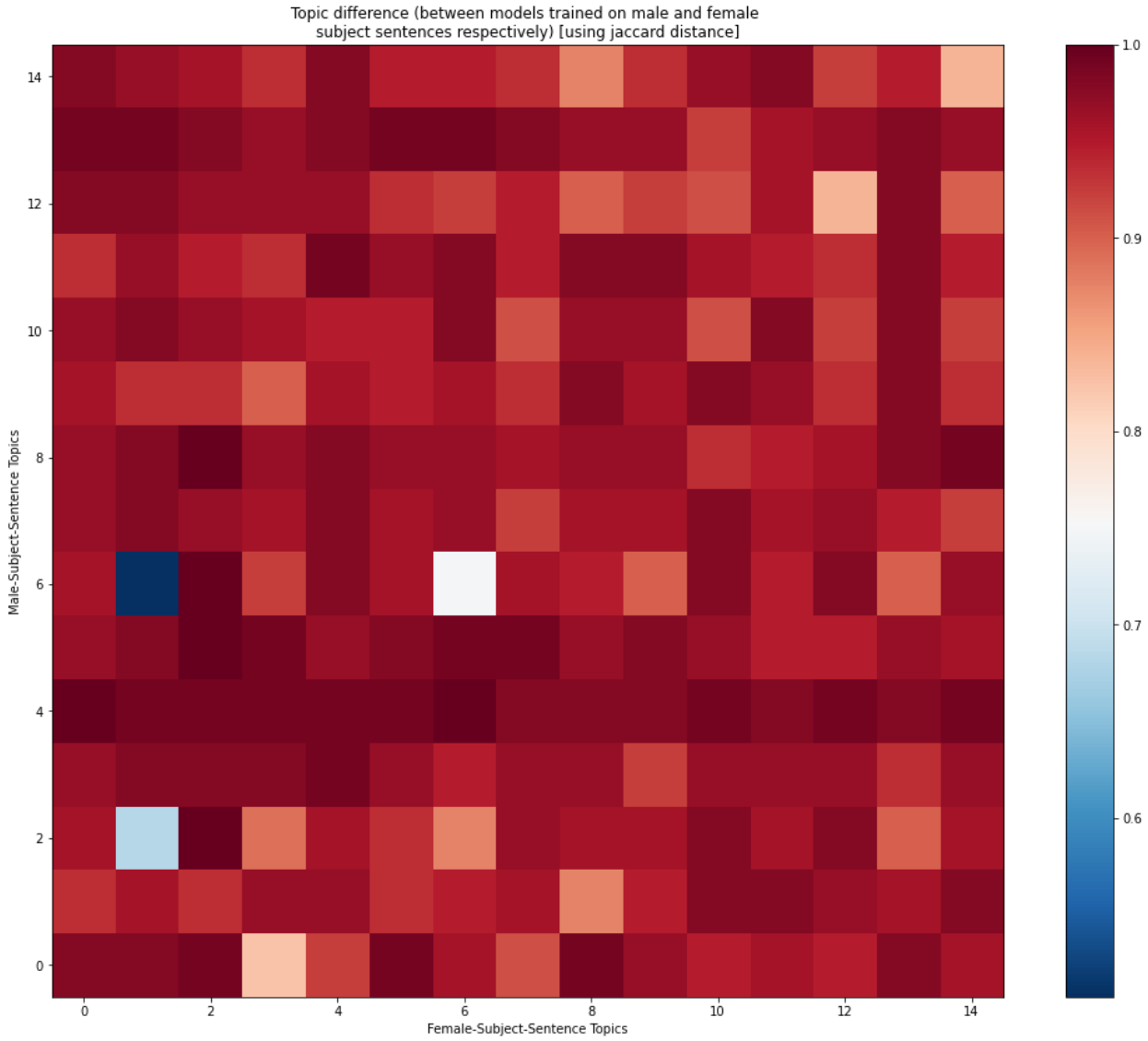


Male-Subject Sentences

Word Embedding Equivalents
(mirroring the methods proposed by Bolukbasi, 2016)

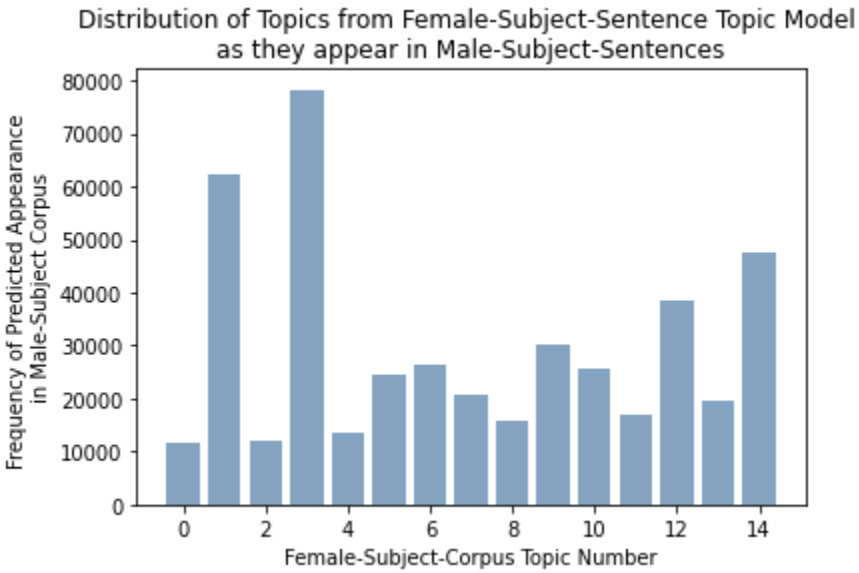
she	he
woman	man
mother	father
sister	brother
mom	dad
girl	boy
spokeswoman	spokesman
daughter	son
chairwoman	chairman
actress	actor
melania	donald
princess	emperor
pregnant	marriage
billboard	espn
empower	promote
shame	embolden
mtv	showtime
bathtub	grenade
rape	murder
cornell	georgetown
doll	pony
amazing	great
highlight	outline

beautiful	nicest
bathroom	room
necklace	tie
documentary	fiction
fertility	cognition
poverty	economics
instagram	twitter
bare	tight
lgbtq	racism
cry	yell
stare	sit
email	memo
adolescent	mature
royal	crown
swimmer	player
feminine	ego
muscle	flex
idol	hero
jazz	chess
rihanna	eminem
study	expert
pornography	serial

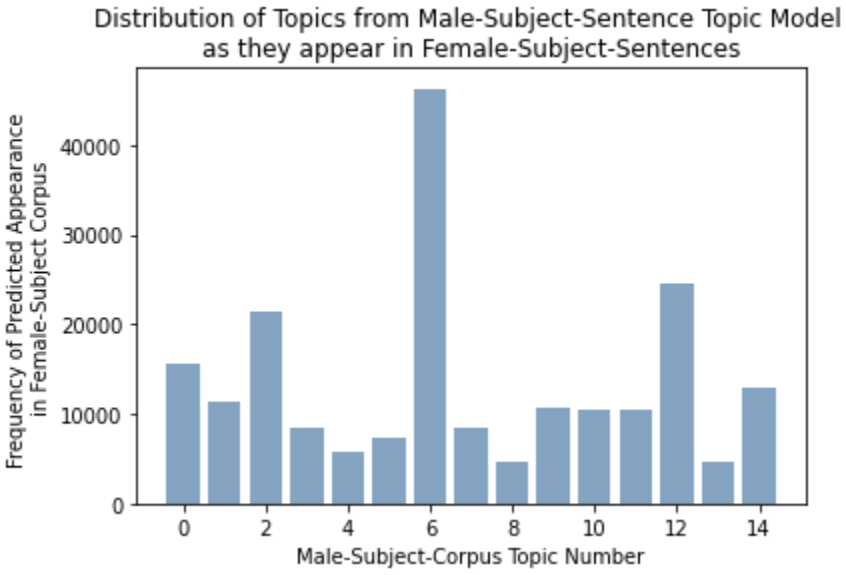


Topic similarity between male and female-subject sentence level corpora.

Topic Occurrence Across Subcorpora
(For Topic Models Trained on Subcorpora)



The distribution of Female-Topics in the Male Corpus.



The distribution of Male-Topics in the Female Corpus.