GENDER BIAS IN THE NEWS: TOPIC MODELLING AND VISUALIZATION FRAMEWORK

Shokhistakhon Mamasodikova,
master, Uzbek State
World Language University, Tashkent, Uzbekistan

Guli Ergasheva,
doctor of philological sciences,Uzbek State
World Language University, Tashkent, Uzbekistan

Abstract: We present a topic modelling and data visualization methodology to examine genderbased disparities in news articles by topic. Existing research in topic modelling is largely focused on the text mining of closed corpora, i.e., those that include a fixed collection of composite texts. We showcase a methodology to discover topics via Latent Dirichlet Allocation, which can reliably produce human-interpretable topics over an open news corpus that continually grows with time. Our system generates topics, or distributions of keywords, for news articles on a monthly basis, to consistently detect key events and trends aligned with events in the real world.

Keywords: gender bias; news media; topic modelling; natural language processing.

Introduction. Gender equality is one of the UN's 17 Sustainable Development Goals. By all measures, most societies are far from such equality. Although progress has been made, women are not equally represented in positions of power are not equal in science, including in publication metrics and do not appear in the news as often as men. Worryingly, COVID-19 seems to have resulted in set-backs not only in areas such as participation in the workforce and unequal share of unpaid work, but also in media representation.

The underrepresentation of women in certain areas of the news such as politics, business, or sports is well documented. There is little large-scale data, however, about representation across entire news organizations, and even less so over a period of time [1].

In this paper, we tackle one measure of gender equality in the media: the number of times men and women are quoted in news articles. More specifically, we present a scalable topic modelling methodology to explore how many times men and women are quoted across news topics. Our unsupervised large-scale analyses are enhanced with corpus-based studies of the language used in male-prominent and female-prominent articles, thus revealing not only which topics are different, but exactly how they are different in their language.

Our analyses rely on statistics provided by the Gender Gap Tracker,1 a purpose-built tool that measures the proportion of male and female sources quoted in mainstream Canadian media in English. The Gender Gap Tracker is an automated software system that monitors men and women's voices on seven major Canadian news outlets in real time, using Natural Language Processing. The goals of the Tracker are to enhance awareness of women's portrayal in public discourse through hard evidence and to encourage news organizations to provide a more diverse set of voices in their reporting. The Tracker has been collecting and analyzing data since October 2018. In this article, we analyze 2 years worth of news articles, a total of 612,343 articles from seven Canadian English-language mainstream news organizations [2].

We apply topic modelling, an unsupervised machine learning technique (Blei et al., 2003), to this data, organizing it into the top 15 most representative topics each month, and then study the distribution of women and men quoted by topic. This method allows us to discover topics across a diverse range of outlets, since we do not rely on news categories provided by the news organizations. Topic modelling can, in principle, scale to unlimited amounts of data and be deployed without much customization.

Our results show that there are clear trends in the proportion of men and women quoted depending on the topic of the article they are quoted in. Women are consistently quoted more frequently than men in the topics "Lifestyle", "Arts and Entertainment", and "Healthcare", whereas men are quoted more frequently than

women in the topics “Business”, “Politics”, and “Sports”. This comes as no surprise, regrettably, as the frequency of quotes seems to reflect a gendered cultural division of duties, where women are placed into the role of caregiver and nurturer and are less represented in the spheres of politics and business. This gendered division of duties and representation in media is no doubt self-reinforcing; quoting female sources more frequently in a caregiving role and quoting male sources more frequently in political and business roles enshrines women’s status as caregivers and men’s status as leaders and breadwinners. Since the first step in any attempt at change is awareness, we contribute by gathering and analyzing the data in a systematic and digestible way to inform the public of these gender divides [3-4].

We would like to acknowledge that the terms we use in this paper are simplifications of a complex reality. We use the terms “women”, “men”, “female sources”, and “male sources”, implying a binary opposition that we know is far from simple. We know, at the same time, that lack of gender representation in many aspects of society is a reality. The aim of this study is to quantify that gender gap by analyzing language and the traditional associations of names with men and women.

Materials and methods. Data Data was obtained for the Gender Gap Tracker and analyzed for this related project. The Gender Gap Tracker data is scraped daily from the websites of seven major Canadian news outlets in English. We built scrapers from scratch for each outlet and maintain them regularly, as the sites change in structure frequently. The data is obtained through the “fair dealing” provision in Canada’sCopyright Act, which allows us to use it for research purposes. It can be made available upon request and upon signing a license agreement [5].

We analyzed 24 months worth of data, a total of 612,343 articles. As we see in Table 1, outlets vary in the volume of articles they publish, with CTV News being the most prolific, and Huffington Post Canada the least so. Articles range in length, roughly between 100 and 1,500 words per article, with a median length of 461 words.2 We provide numbers for people mentioned and people quoted. To highlight the lack of gender equality, the percentage columns in Table 1 indicate the percentage of women mentioned or quoted for each outlet and the average across outlets at the bottom.

Using the Gender Gap Tracker’sexistingNatural Language Processing pipeline, the data is enriched with information about the people mentioned and the people quoted, whom we refer to as sources. For the latter, we first identify quotes in the text and then associate a speaker with them, using coreference resolution. Finally, we look up the gender of the speakers (male, female, or other) using a mix of sources: an in-house cache of commonly quoted public figures and an API that stores the gender of people based on their first or full name. Thanks to this NLP pipeline, we obtain gender statistics (people named, people quoted, and their gender) for each article, whose aggregated form is shown in Table 1.Weuse thesegender statistics for the topic modelling and language analyses described in this paper.

In this work, people mentioned refers to all unique individuals named in an article, regardless of whether or not they were quoted. These names are extracted and identified by our NLP pipeline via Named Entity Recognition. People quoted refers to the subset of people mentioned that were quoted one or more times in an article, identified via coreference resolution. The counts of people mentioned and people quoted are per article, as we do not keep track of the same individual across articles, that is, our unit of analysis is the article.

Numbers alone tell a compelling story about the lack of female voices in the media, as we show in more detailed analyses of the Gender Gap Tracker data [6].

Our goal in this paper is to explore how certain news topics contribute disproportionately to those percentages, and whether we can identify any systemic patterns in the language used in articles where each gender dominates in articles with majority male or female sources.

TABLE 1 | Data for the study, October 1, 2018–September 30, 2020. Note: "people mentioned" is the number of all persons named in all the articles per outlet. "People quoted" is the number of mentioned persons who were quoted one or more times in each article. Each "quote" is only counted once per person per article.

Outlet	Number of articles	People mentioned	Percentage women mentioned (%)	People quoted	Percentage women quoted (%)
CBC News	151,288	749,351	30.2	286,770	32.6
CTV News	158,249	565,990	28.3	240,215	29.7
Global News	86,386	353,733	25.3	133,021	30.1
HuffPost Canada	15,765	86,429	29.2	25,975	30.8
National Post	27,925	166,080	21.4	45,663	23.9
The Globe and Mail	87,121	496,142	22.1	153,881	23.2
The Toronto Star	85,609	509,851	23.1	163,255	25.1
Overall	612,343	2,927,576	25.7	1,048,780	27.9

Topic Modelling Topic modelling is an unsupervised machine learning technique to discover the main topics, or themes, in a collection of unstructured documents. A topic here refers to a cluster of words that represents a larger concept in the real world. Each document in a corpus can be imagined as consisting of multiple topics in different proportions all at once—for example, an article about a major airline procuring new aircraft may contain many words related to finance, geopolitics, or travel policies, as well as passenger trends or market events. A document can thus be composed of several topics, each represented by specific words. Topic modelling encapsulates these ideas into a mathematical framework that discovers clusters of word distributions representing overall themes within the corpus, making it a useful technique to analyze very large datasets for their content [7].

Topic modelling has been successfully deployed to study changes, relations, and impact in the scientific literature or different points of view in the news. It is particularly well-suited for organizing and classifying large numbers of documents and has been deployed in many areas of the social sciences, e.g., for policy analysis discourse studies news media studies or social media data. It is a highly customizable method and results can vary depending on how parameters are set. We present a summary of the main parameters and experiments on how to optimally set them.

Results. Monthly Gender Prominence for Recurring Topics We examined nine topics that feature regularly in most months between October 2018 and September 2020. The topics cover a broad range of domains typically seen in the news, as seen on the vertical axis of Figure 1. From our dashboard we obtain per-month, per-outlet mean topic gender prominence measures for each of these topics over the 2-year period. We define gender prominence as the difference in mean topic weights between the female and male corpus or dataset for a given topic. A topic is categorized as having male prominence if the mean topic weights from the male corpus are greater than those from the female corpus. See the last section in the Supplementary Material for more details on how we define male and female prominence and how we separate topics into male and female corpora.

The values for prominence range from positive (red in Figure 1), indicating female prominence for that topic, to negative (blue), indicating male prominence. We tabulate these numbers per outlet and topic for the entire duration, and then calculate the mean gender prominence for each topic over all outlets. The resulting table is plotted as a time-series heat map, as shown in Figure 1. The white (neutral) squares in the heat map indicate that a given topic did not appear at all in that month.

Figure 1 shows that there are topics that clearly exhibit male or female prominence over extended periods of time. This indicates that, for specific topics, news outlets tend to consistently feature either men's or women's voices more frequently, resulting in majority male or female sources quoted in a large fraction of the articles for that topic, regardless of the outlet. For example, the topic "Arts and entertainment", which primarily discusses news about public events, art exhibitions, films, television, and celebrities, tends to display a high topic intensity in the female corpus. We see the topic is not as widely covered in February–May 2020, likely because, in the first few months of the COVID-19 pandemic, most live arts events were cancelled [8-9].

Other topics that display strong female prominence over time include "Lifestyle", "Healthcare and medical research", and "Legal and court cases". The primary reason for the "Lifestyle" topic being female-prominent is that it regularly features mothers involved in childcare, women reliving their past experiences, and female experts offering personal care advice. The "Legal and court cases" topic tends to feature more women mainly

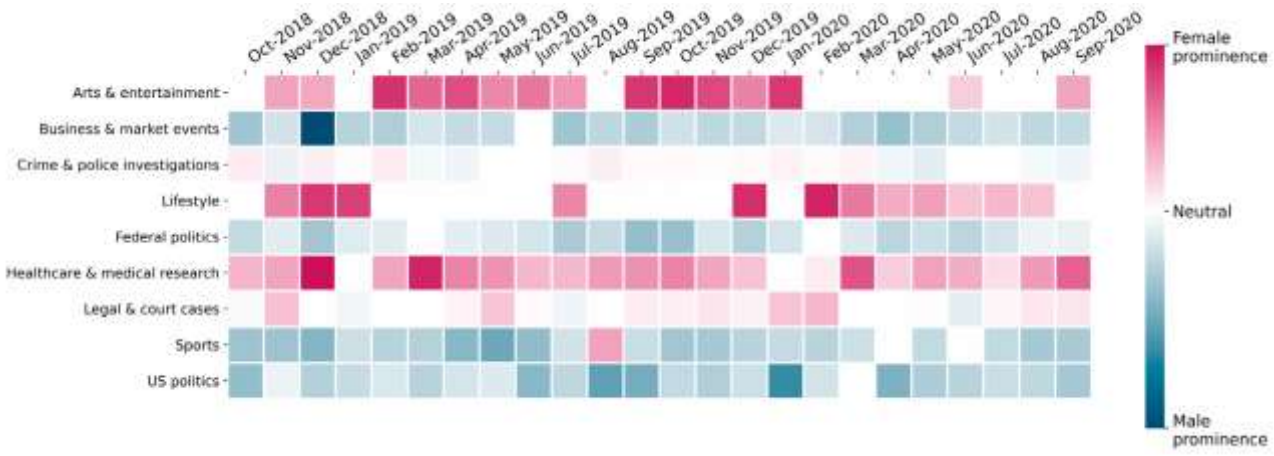

Figure 1. Monthly topic gender prominence for nine recurring topics (average over all outlets).

due to cases that pertain to sexual assault or women's rights, featuring both victims and experts who are women. The "Healthcare and medical research" topic is particularly interesting, because, in Canada, a sizable proportion of healthcare experts and Public Health Officers happen to be women. This leads to the healthcare topic regularly featuring quite strongly in the female corpus, with a range of women's health issues being covered and prominent expert women being quoted [10].

Conclusion. We analyzed a large news corpus to understand the relationship between topics in the news and the gender of those quoted. We also perform corpus-based language analyses using each article's dominant topic distribution, confirming our hypothesis that articles that quote more men than women, on average, tend to use different words and verb-object combinations.

Our results, which consistently show that women are quoted more frequently in topics related to lifestyle, healthcare, and crimes and sexual assault, are, unfortunately, not unexpected. Multiple studies have found that women's voices tend to be relegated to the domestic sphere and traditional female-dominated areas. Specifically within the news domain and focusing on women as experts, women's expertise is more likely to be found in "the sphere of the private, emotional and subjective". Kassova also found that women are more central in crime and celebrity stories and less prominent in political and financial news. Women tend to speak as citizens rather than experts and, when they speak as experts, they do so more often about health, a caregiving profession, than about politics. When women are represented in political news, the lens is often gendered. This is a form of framing effect, which has an influence on how the public views gender in general.

Balanced gender representation in the media is within our reach, if enough effort is devoted to this goal and if we incorporate accountability into the effort. The Gender Gap Tracker, together with the topic-based analyses presented here, paint a clear picture of inequality in gender representation in the media. Our hope with the Gender Gap Tracker's dashboard is that it be used as an accountability tool to encourage and facilitate gender parity in sources. We have seen from other initiatives, such as the BBC's 50:50 campaign, that accountability leads to better balance in sources. In our view, providing organizations with a visual means to narrow down

which topics exhibit the strongest disparities can have a tangible impact on improving the gender balance of sources.

REFERENCES

1. Asr, F. T., Mazraeh, M., Lopes, A., Gautam, V., Gonzales, J., Rao, P., et al. (2021). The Gender Gap Tracker: Using Natural Language Processing to Measure Gender Bias in media. PLoS ONE 16 (1), e0245533. doi:10.1371/journal.pone. 0245533
2. Baker, P. (2014). Using Corpora to Analyze Gender. London: Bloomsbury.
3. Berenbaum, M. R. (2019). Speaking of Gender Bias. Proc. Natl. Acad. Sci. USA 116 (17), 8086–8088. doi:10.1073/pnas.1904750116 Blei, D., Carin, L., and Dunson, D. (2010). Probabilistic Topic Models: A Focus on Graphical Model Design and Applications to Document and Image Analysis. IEEE Signal. Process. Mag. 27 (6), 55–65. doi:10.1109/MSP.2010. 938079
4. Blei, D. M., Ng, A. Y., and Jordan, M. I. (2003). Latent Dirichlet Allocation.
5. J. Machine Learn. Res. 3, 993–1022. doi:10.5555/944919.944937 Blei, D. M. (2012). Probabilistic Topic Models. Commun. ACM 55 (4), 77–84. doi:10.1145/2133806.2133826
6. Blodgett, S. L., Barocas, S., Daumé, H., III, and Wallach, H. (2020). “Language (Technology) Is Power: A Critical Survey of “Bias” in NLP,” in Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics, 5454–5476. https://www.aclweb.org/anthology/2020.acl-main.485.
7. British Broadcasting Corporation (2020). 50:50 the Equality Project—50:50. London: BBC. https://www.bbc.co.uk/5050.
8. Brookes, G., and McEnery, T. (2019). The Utility of Topic Modelling for Discourse Studies: A Critical Evaluation. Discourse Stud. 21 (1), 3–21. doi:10.1177/ 1461445618814032
9. Butt, M. (2010). “The Light Verb Jungle: Still Hacking Away,” in Complex Predicates: Cross-Linguistic Perspectives on Event Structure,48–78.
10. Caldas-Coulthard, C. R., and Moon, R. (2010). ‘Curvy, Hunky, Kinky’: Using Corpora as Tools for Critical Analysis. Discourse Soc. 21 (2), 99–133. doi:10. 1177/0957926509353843