THE GENDER GAP TRACKER: USING LANGUAGE PROCESSING TO GENDER BIAS IN MEDIA

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 examine gender bias in media by tallying the number of men and women quoted in news text, using the Gender Gap Tracker, a software system we developed specifically for this purpose. The Gender Gap Tracker downloads and analyzes the online daily publication of seven English-language Canadian news outlets and enhances the data with multiple layers of linguistic information. We describe the Language Processing technology behind this system, the curation of off-the-shelf tools and resources that we used to build it, and the parts that we developed. We evaluate the system in each language processing task and report errors using real-world examples. Finally, by applying the Tracker to the data, we provide valuable insights about the proportion of people mentioned and quoted, by gender, news organization, and author gender.

Women's voices are disproportionately underrepresented in media stories. The Global News Monitoring Project has been tracking the percentage of women represented in mainstream media since 1995, when it was 17%. Twenty years later, in 2015, it had increased to only 24%, with a worrisome stalling in the previous decade [1]. At this rate, it would take more than 70 years to see 50% women in the media, a true reflection of their representation in society.

The underrepresentation of women is pervasive in most areas of society, from elected representatives [2–5] and executives [3, 6, 7] to presidents and faculty in universities [5, 8, 9]. Women are also underrepresented in political discussion groups. It is, therefore, not entirely surprising that news stories mostly discuss and quote men: Many news stories discuss politicians and business executives, drawing on expert opinion from university professors to do so. Perversely, in many stories where women are overrepresented, it is because they are portrayed as having little or no agency, as in the case of victims of violence [9] or politician's spouses. During international gatherings like G7/G8 or G20 summits, a set of stories often discuss the parallel meetings of spouses with a focus on their attire, and humorously commenting on cases when the lone man joins activities clearly planned for wives only. Countless studies have pointed out how the representation of women in media is different. In our project, we first tackle the question of how much of a difference there is in the representation of women.

Not a great deal of progress seems to have been made since Susan Miller found, in 1975, that photos of men outnumbered photos of women by three to one in the pages of the Washington Post, and by two to one in the Los Angeles Times. Among the more than 3,600 photos that Miller studied, women outnumbered men only in the lifestyle section of the two papers [6].

Most previous studies of gender representation in media have performed manual analyses to investigate the gap. Informed Opinions, our partner organization in this project, carried out a study in 2016, analyzing 1,467 stories and broadcast segments in Canadian media between October and December 2015, to find that women were quoted only 29% of the time [7]. The work was laborious and intensive. Similarly, the enormous effort of the Global News Monitoring Project is only possible thanks to countless volunteers in 110 countries and many professional associations and unions around the globe. Thus, it only takes place every five years. Shor et al.'s [4] study of a historical sample of names in 13 US newspapers from 1983 to 2008 found that the ratio went only from 5:1 in 1983 to 3:1 by the end of the period. It seems to be stubbornly stuck at that level. A recent analysis of news coverage of the COVID-19 pandemic [8] used a mix of manual and automatic

methods and found that men were quoted between three and five times more often than women in the news media of six different countries.

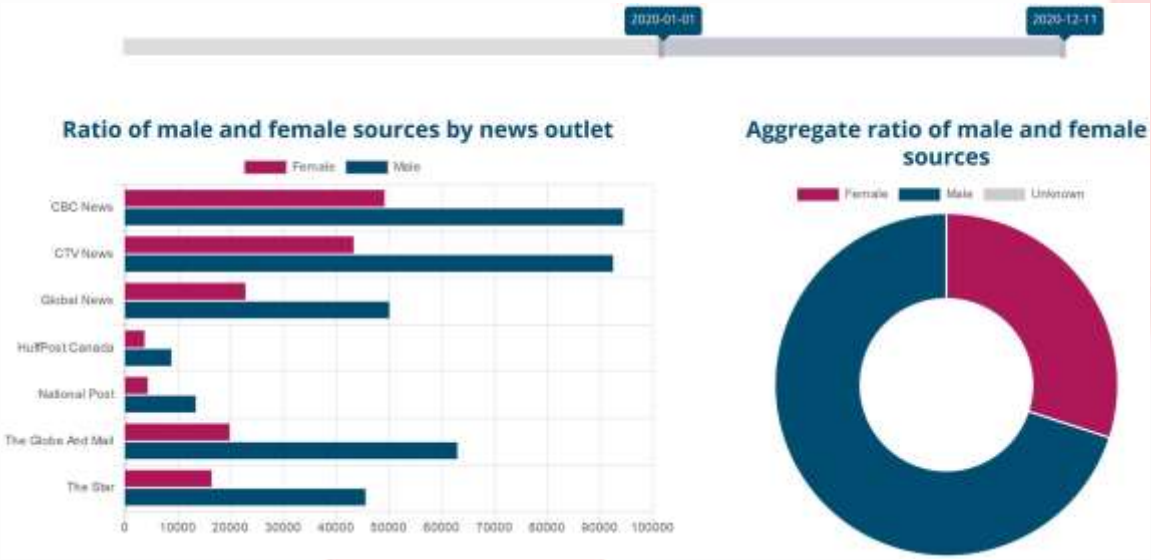

Fig 1. The Gender Gap Tracker online dashboard page.

The causes and solutions to the underrepresentation of women in society in general and in news articles in particular are too complex to discuss in this paper (but see [29–33]). We focus here on the first step in any attempt at change: an accurate characterization of the current situation. Just like a step tracker can motivate users to increase their physical activity, we believe that the Gender Gap Tracker can motivate news organizations to bring about change in areas they have control over. It is obvious that, if a news story requires a quote from the Prime Minister or a company's president, the journalist does not have a choice about the gender of those quoted. Journalists, however, do have control over other types of sources, such as experts, witnesses, or individuals with contrasting viewpoints.

Indeed, when journalists keep track of their sources and strive to be more inclusive, both anecdotal and large-scale evidence show that parity is, in fact, possible. Ed Yong, staff writer for The Atlantic who covers science news, reported that keeping track of his sources was the simple solution to ensure gender parity in his articles [4]. Ben Bartenstein, who covers financial news for Bloomberg, improved the gender ratio in his sources by keeping lists of qualified women and tracking the sources in his stories [5]. The BBC's 50:50 project also uses strategic data collection and measurement to achieve 50% women contributing to BBC programs and content.

It is with this goal in mind—of motivating news organizations to improve the ratio of people they quote—that the Gender Gap Tracker was born. The Gender Gap Tracker is a collaboration between Informed Opinions, a non-profit organization dedicated to amplifying women's voices in media, and Simon Fraser University, through the Discourse Processing Lab and the Big Data Initiative.

We harness the power of large-scale text processing and big data storage to collect news stories daily, perform Natural Language Processing (NLP) to identify who is mentioned and who is quoted by gender, and show the results on a public dashboard that is updated every 24 hours (https://gendergaptracker.informedopinions.org). The Tracker monitors mainstream Canadian media, seven English-language news sites (a French Tracker is in development), motivating them to improve the current disparity. By openly displaying ratios and raw numbers for each outlet, we can monitor the progress of each news organization towards gender parity in their sources. Fig 1 shows a screenshot of the live page. In addition to the bar charts for each organization and the doughnut chart for aggregate values, the web page also displays a line graph, charting change over time (see Fig 2 below).

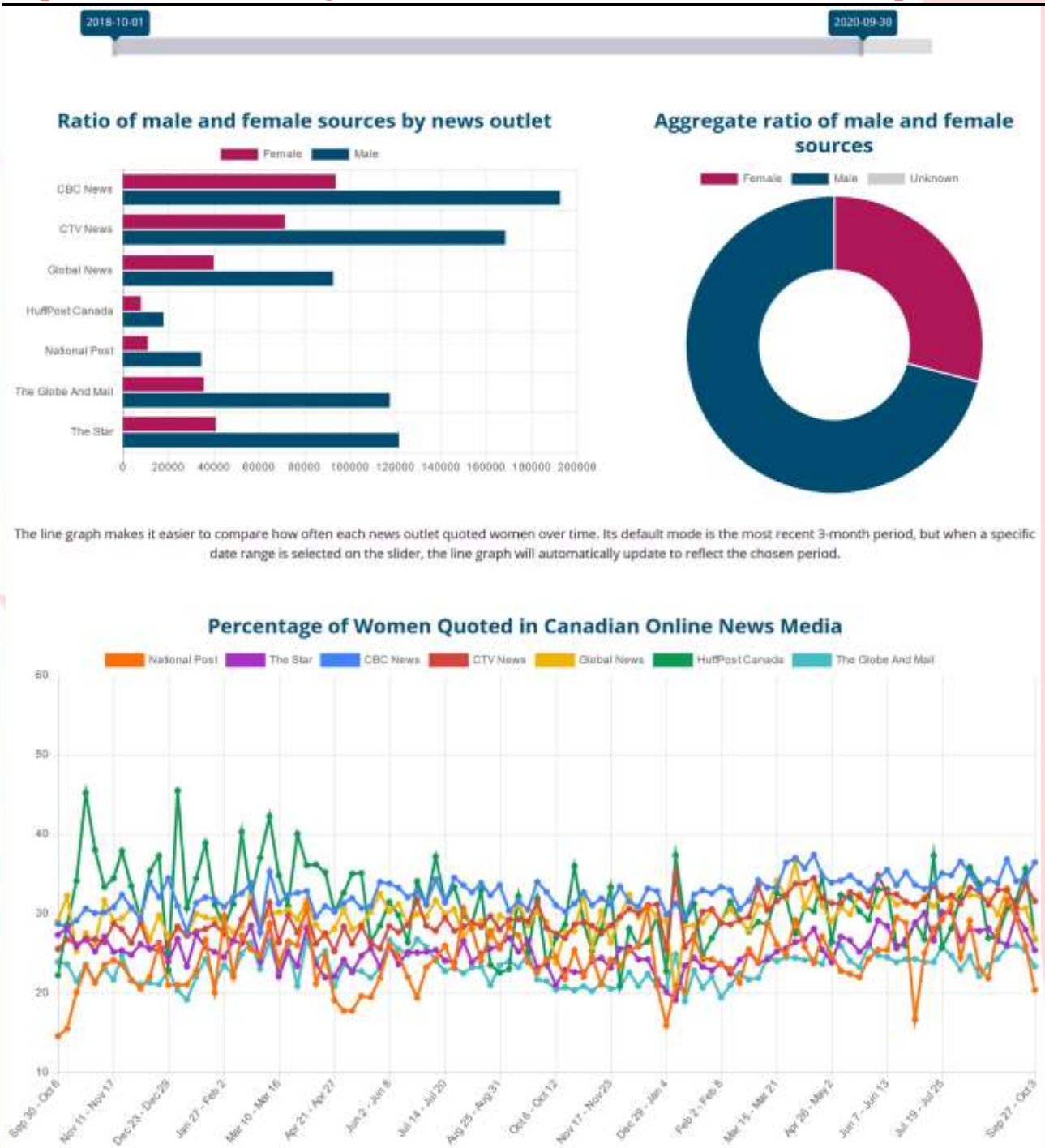

Fig 2. Counts and percentages of male vs. female sources of opinion across seven news outlets

For the two years since data collection started on October 1, 2018 until September 30, 2020, the average across the seven news outlets is 29% women quoted, versus 71% men, with a negligible number of unknown or other sources. We have, however, observed an increase in the number of women quoted between the first and the last month in that period, from 27% in October 2018 to 31% in September 2020. Some of that increase can be directly attributed to an increase in the quotes by public health officers during the COVID-19 crisis. It just so happens that a large number of those public health officers across Canada are women [6]. We report some of the analyses and insights we are gathering from the data in the section Analysis and observations.

In this paper, we describe the data collection and analysis process, provide evaluation results and a summary of our analysis and observations from the data. We also outline other potential uses of the tool, from quantifying gender representation by news topic to uncovering emerging news topics and their protagonists. We start, in Related work, with a review of existing literature on quotation patterns, extracting information

from parsed text, and potential biases in assigning gender to named entities. We then provide, in Data acquisition and NLP processing pipeline, a high-level description of the data acquisition process and how we deploy NLP to extract quotes, identify people, and predict their gender. More detail for each of those steps is provided in the S1 Appendix. Throughout the development of the Gender Gap Tracker, we were mindful of the need for accuracy, in both precision and recall of quotes, but also in terms of any potential bias towards one of the genders (e.g., disproportionately attributing names or quotes to one gender). In order to ensure that the Gender Gap Tracker provides as accurate a picture as possible, we have performed continuous evaluations. We describe that process in the section on Evaluation. The section Analysis and observations answers the most important questions that we posed at the beginning of the project: Who is quoted, in what pro-portions? We add more nuanced analyses about the relationship between author gender and the gender breakdown of the people those authors quote. Finally, Conclusion offers some reflections on the use of the Gender Gap Tracker as a tool for change, also discussing future improvements and feature additions [9].

Before delving into the technical aspects of the Gender Gap Tracker and the insights it provides about the gender gap in media, we would like to acknowledge that the language we choose to describe people matters and that the terms we use are simplifications of a complex reality. We use ‘women’ and ‘men’ and ‘female sources’ and ‘male sources’, implying a binary opposition that we know is far from simple. Gender is more nuanced than that. We know, at the same time, that lack of gender representation in many aspects of society is a reality. Our goal is to quantify that lack of representation by using language and the traditional associations of names and pronouns with men and women. We discuss this issue in more detail in the section on on Gender prediction and gender bias in Natural Language Processing.

REFERENCES

1. Macharia S. Who Makes the News? Global Media Monitoring Project; 2015.
2. Chesser SG. Women in National Governments Around the Globe: Fact Sheet. Washington, DC: Congressional Research Service; 2019.
3. World Economic Forum. The Global Gender Gap Report 2020. Geneva, Switzerland: World Economic Forum; 2019.
4. Pew Research Center. The Data on Women Leaders. Pew Research Center; 2019.
5. Jalalzai F. Shattered, Cracked, or Firmly Intact?: Women and the executive glass ceiling worldwide. Oxford: Oxford University Press; 2013.
6. Zillman C. The Fortune 500 has more female CEOs than ever before. Fortune. 2019; May 16, 2019. Available from: https://fortune.com/2019/05/16/fortune-500-female-ceos/.
7. Johnson GF, Howsam R, Smith MS, Bray N. Leadership Pipelines at Five Canadian Universities. University of Alberta; 2020. Available from: https://uofaawa.wordpress.com/awa-diversity-gap-campaign/ the-diversity-gap-in-2020-leadership-pipelines-at-five-canadian-universities.
8. Johnson GF, Howsam R. Whiteness, power and the politics of demographics in the governance of the Canadian academy. Canadian Journal of Political Science. 2020; 53(3): 676–694. https://doi.org/10.1017/S0008423920000207.
9. Beauvais E. The gender gap in political discussion group attendance. Politics & Gender. 2019; 16: 1– 24.