

Technical Report

Scalable Topic Modelling on a Large News Corpus: The Gender Gap Tracker

Prashanth Rao and Maite Taboada



Contact: mtaboada@sfu.ca

October 1, 2020

CONFIDENTIAL - PLEASE DO NOT DISTRIBUTE UNLESS AUTHORIZED BY ONE OF THE
AUTHORS

Table of Contents

| | |
|---|-----------|
| 1 Introduction | 3 |
| 2 What is topic modelling? | 5 |
| 2.1 Latent variables | 6 |
| 2.2 Sampling vs. variational inference | 7 |
| 2.3 Expectation maximization vs. online variational Bayes | 8 |
| 3 Related work | 8 |
| 3.1 Strengths and weaknesses of LDA | 8 |
| 3.2 Developing a reliable methodology | 9 |
| 4 Our methodology | 10 |
| 4.1 Tools and resources | 10 |
| 4.2 Preprocessing steps | 11 |
| 4.3 Stop words | 12 |
| 4.4 Custom lemma lookup table | 13 |
| 4.5 Limitations of our approach | 14 |
| 5 Experiments | 15 |
| 5.1 Number of topics | 15 |
| 5.2 Random seed | 17 |
| 5.3 Hyperparameter tuning | 19 |
| 5.3.1 Maximum iterations | 19 |
| 5.3.2 Relative pruning | 19 |
| 5.3.3 Vocabulary size | 20 |
| 5.4 Time span of data modelled | 21 |
| 5.5 Final fine-tuned hyperparameters | 23 |
| 6 Topic labelling guidelines | 23 |
| 6.1 Naming patterns | 23 |
| 6.2 Specificity | 24 |
| 6.3 Topic label reuse | 24 |
| 7 Topic dashboard | 25 |
| 7.1 Topic intensity | 26 |
| 7.2 Topic gender prominence | 27 |
| 8 Analysis and observations | 29 |
| 8.1 Monthly gender prominence for nine recurring topics | 29 |

| | |
|--|----|
| 8.2 Monthly gender prominence per outlet | 31 |
| 8.3 Corpus analysis | 33 |
| 8.3.1 Sports | 35 |
| 8.3.2 Business & market events | 37 |
| 8.3.3 Lifestyle | 38 |
| 8.3.4 Crimes & sexual assault | 40 |
| 8.3.5 Healthcare & medical research | 41 |
| 9 Conclusions | 43 |
| 10 References | 44 |

1 Introduction

The [Gender Gap Tracker](#) (GGT) is an automated software system that measures men and women's voices on seven major Canadian news outlets in real time. It analyzes the rich information in news articles using Natural Language Processing (NLP) and provides the means to quantify and measure the discrepancy in proportions of men and women quoted in the news. The larger goals of the project 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 Gender Gap Tracker is a collaboration between [Informed Opinions](#), a non-profit dedicated to amplifying women's voices in the 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 to identify who is mentioned and who is quoted by gender, and show the results on a public dashboard that is updated every 24 hours¹. The Tracker monitors seven English-language news sites from mainstream Canadian media (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. Figure 1 shows a screenshot of the live page.

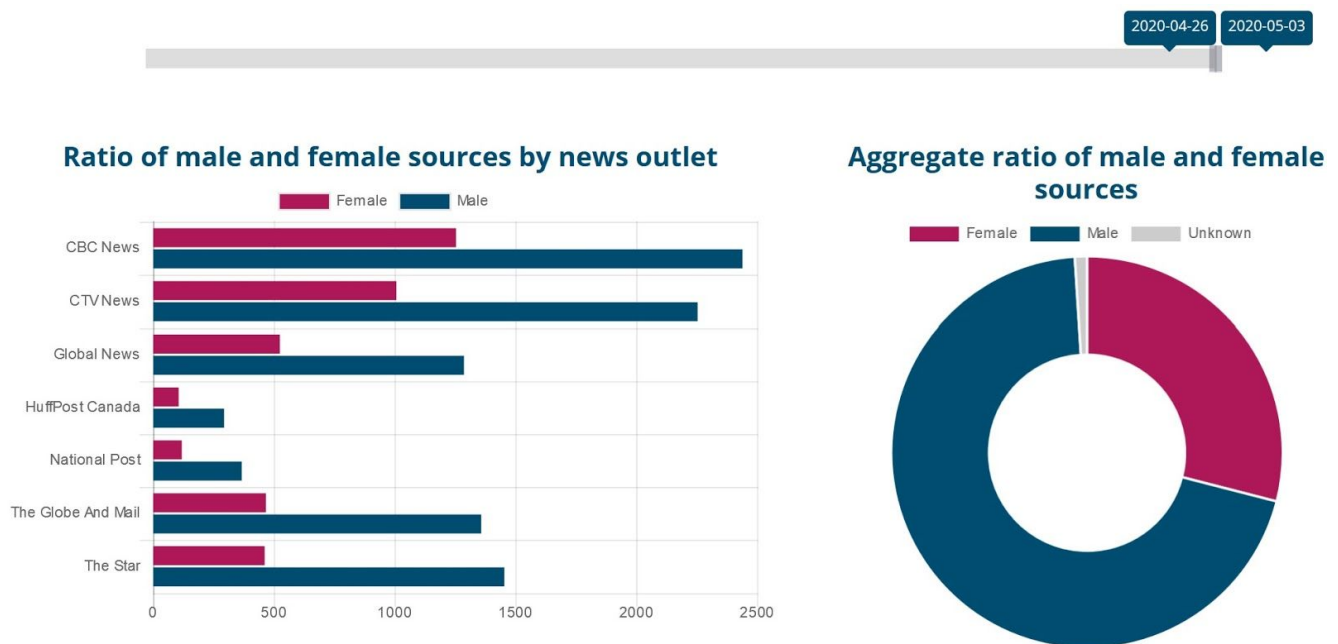


Figure 1: The Gender Gap Tracker online dashboard page (<https://gendergaptracker.informedopinions.org/>)

¹ <https://gendergaptracker.informedopinions.org/>

The Gender Gap Tracker displays the counts of men and women quoted (*sources*). The visualization captures a breakdown of the total number of articles per month that quote more men than women, and vice versa.

Figure 2 shows the number of articles per month with a majority of female and male sources over a two-year period. We can see that the number of articles that quote more men than women is roughly 3-4 times (on average) greater than the number of articles with a majority of female sources. In other words, news outlets on average quote men far more often than they do women, at a ratio of about **3:1**. Many of these sources being quoted are frequently repeated across all outlets, as well as across time periods—for example, our data tells us that Donald Trump and Justin Trudeau are the most quoted men by far, appearing as sources in a large proportion of all articles that feature politics and related topics. The fact that both curves in Figure 2 tend to rise and fall together in the same months indicates that the disparity between the raw numbers of men and women quoted is a constant feature of news coverage in the real world. In other words, the increase and decrease in Figure 2 is due to volume of articles and quotes, not to any significant changes in that 3:1 ratio over time.

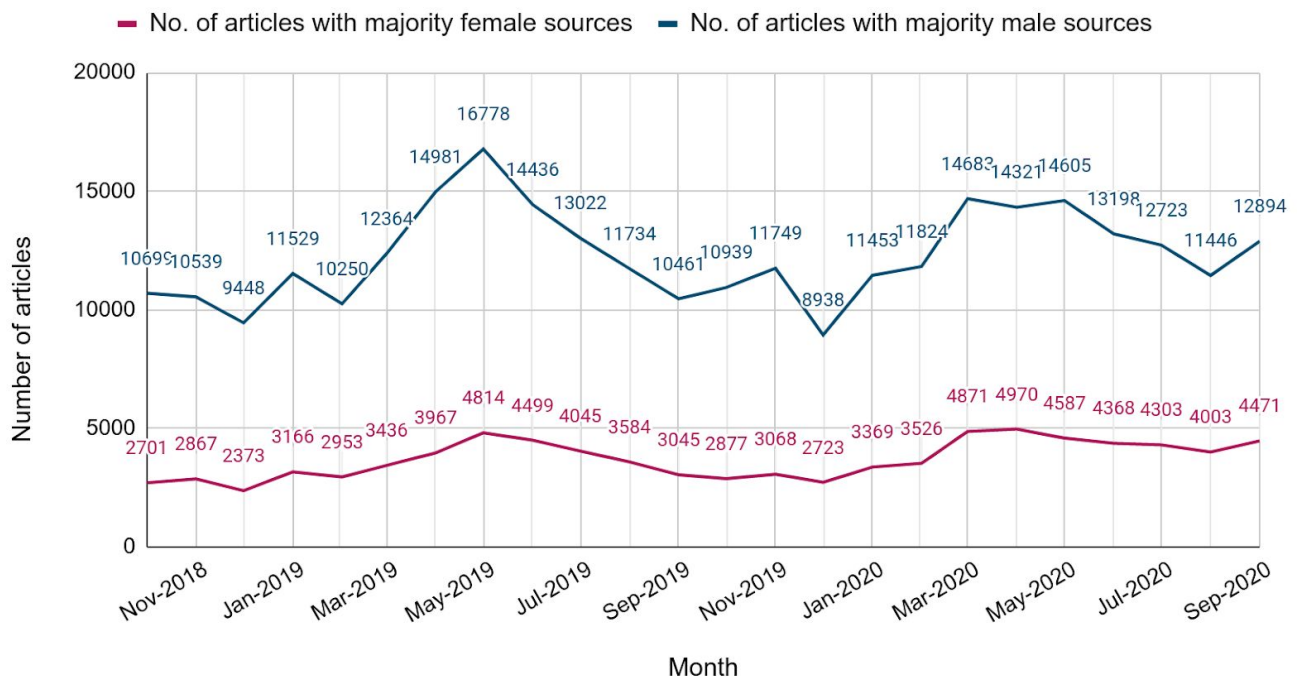


Figure 2. Monthly counts of articles that contain majority male/female sources

Numbers alone tell a compelling story about the lack of female voices in the media. We want, however, to probe those numbers further, studying whether the disparities are more or less marked in certain news topics. By performing large-scale discovery of topics, we aim to utilize our existing data from the Gender Gap Tracker system to analyze whether female and male sources are more likely to be associated with specific topics in the news.

2 What is 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 from the real world. Each document in a corpus can be imagined as consisting of multiple topics in different proportions all at once—for example, in an article about a major airline procuring new aircraft, it is reasonable to expect many words related to finance, geopolitics, travel policy, as well as passenger trends and market events that led to the deal taking place. A document can thus be composed of several topics, each consisting of specific words (that may or may not overlap between topics). 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.

The mathematical goal of topic modelling is to fit a model’s parameters to the given data using heuristic rules, such that there is a maximum likelihood that the data arose from the model. Such methods are known as parametric methods, among which *Latent Dirichlet Allocation* (LDA) is by far the most popular. For a more detailed survey of various parametric and non-parametric probabilistic topic models, see [Blei \(2012\)](#).

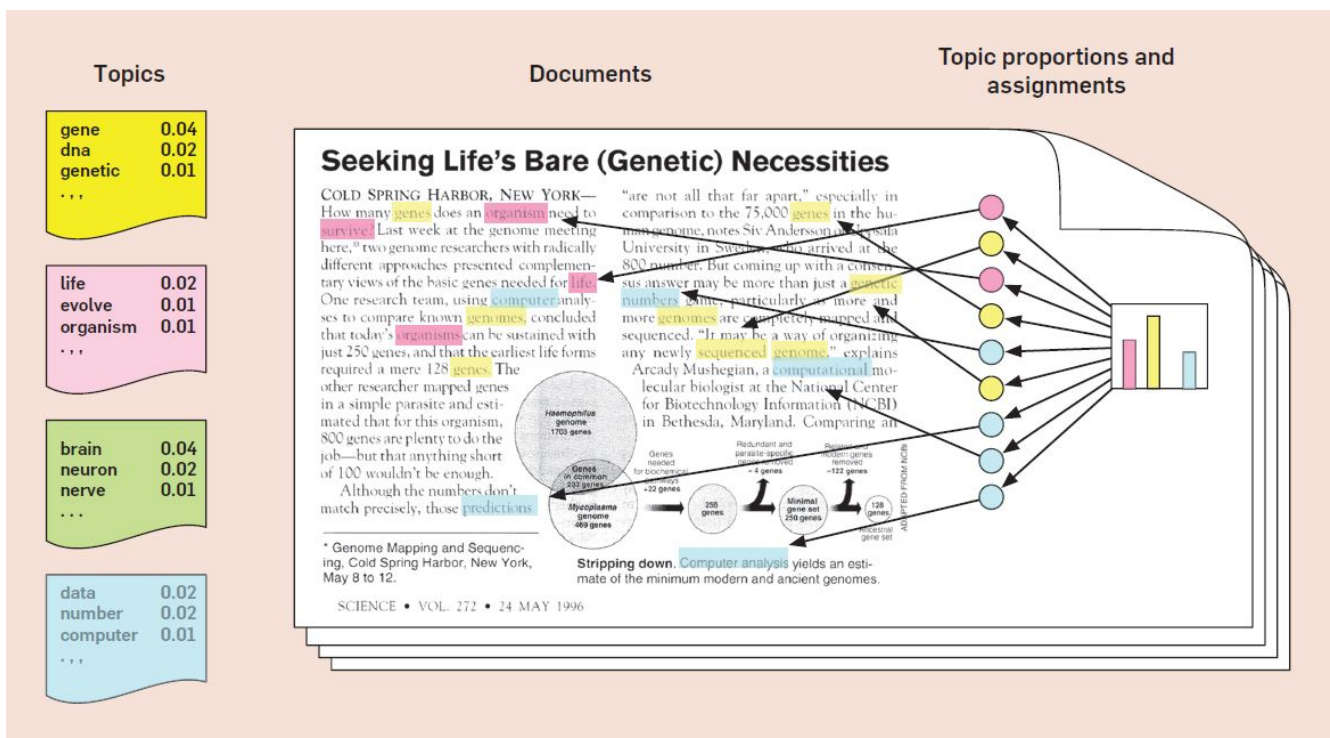


Figure 3. Intuition behind LDA ([Blei, 2012](#))

Figure 3 shows the intuition behind how an LDA model is applied in the real world. A fixed number of topics exist for the whole corpus, with each topic viewed as a distribution over words. Each document is viewed as a distribution over a fixed number of topics, which are themselves composed of words from a particular topic.

The below key qualities of LDA are relevant to the topic modelling methodology used in this study:

- It is a **generative probabilistic** model: The data in a corpus, i.e., the observed variables, are treated as though they arise from an imaginary random process that includes hidden (latent) variables.
- It is a **Bayesian** model: The generative process defines a joint probability distribution over both the observed and hidden random variables. This joint distribution is used during data analysis to compute the *posterior* distribution of hidden variables given the observed variables.
- It is a **mixed-membership** model: Each document exhibits multiple topics in different proportions, and each topic can exhibit words that also occur in other topics.
- It is a **bag-of-words** model: LDA does not consider word order in its distributions. This is quite sufficient for a coarse-grained semantic understanding of topic content over a large corpus.

2.1 Latent variables

The joint probability distribution of a topic mixture, a set of N topics and a set of N words is formulated as follows ([Blei, Ng, and Jordan, 2003](#)).

$$p(\theta, \mathbf{z}, \mathbf{w} | \alpha, \beta) = p(\theta | \alpha) \prod_{n=1}^N p(z_n | \theta) p(w_n | z_n, \beta) \quad (1)$$

The parameter α from eq. (1) governs the distribution of topics for the documents in the corpus, and is drawn from a *Dirichlet* distribution. β is a random matrix designed to parameterize the word probabilities, also modelled as a Dirichlet distribution. The dimensionality of β is given by $\beta_{(i,j)}$ which represents the probability of the i^{th} topic containing the j^{th} word. A simplified way to state the expression from eq. (1) would be:

$$\theta \sim \text{Dirichlet}(\alpha) \quad (2)$$

A major assumption in LDA is that the dimensionality k of the Dirichlet distribution (and thus the number of topics) is known and fixed beforehand by the user. In practice, determining the number of topics is a heuristic exercise ([Zhao et al. 2015](#)).

Because LDA is framed as a Bayesian problem, the key issue that needs to be resolved is one of inference, i.e., computing the posterior distribution of the hidden variables. Unfortunately, this distribution is intractable due to implicit coupling between θ and β in the summation over latent topics. The solution to this intractable problem was proposed by [Blei, Ng, and Jordan \(2003\)](#)—an approximation called *variational inference* that closely matches the true posterior is used instead.

$$q(\theta, \mathbf{z} | \gamma, \phi) = q(\theta | \gamma) \prod_{n=1}^N q(z_n | \phi_n) \quad (3)$$

According to this formulation, new “free” variational parameters are introduced (the Dirichlet parameter γ and the multinomial parameter ϕ). This reframes the inference problem as an optimization problem that seeks to minimize the KL-divergence² between the variational distribution and the true posterior ([Blei, 2012](#)).

Due to further mathematical complexities imposed by the mixture model setting and the presence of sparsity (a new document is very likely to contain words that did not appear in any documents in the training corpus), a Dirichlet *smoothing* step is applied. This introduces yet another parameter, η , which is used to smooth the free variational parameters. This ultimately influences the words distributed over topics through the β parameter.

$$\phi \sim \text{Dirichlet}(\eta) \quad (4)$$

In summary, an LDA algorithm learns the below hidden (latent) variables:

- α : Parameter that governs the topic distribution for each document.
- η : Parameter that governs the word distribution for each topic.
- θ : Random matrix $\theta_{i,j}$ representing the probability of the i^{th} document containing the j^{th} topic.
- β : Random matrix $\beta_{i,j}$ representing the probability of the i^{th} topic containing the j^{th} word.

Note that the α and η parameters are not necessarily scalars or symmetric vectors—in practice, they are modelled as asymmetric vectors to improve the stability and fitting accuracy of the algorithm.

2.2 Sampling vs. variational inference

Some implementations of LDA use a sampling-based approach to compute the approximation of the true posterior. The most common sampling method used is *Gibbs sampling*, in which a Markov chain of random variables is constructed with each variable dependent on the previous ones—the limiting value of this distribution equals the true posterior. The algorithm is run on the Markov chain defined on the hidden variables for a particular corpus and a number of samples are drawn using a *Markov Chain Monte Carlo* algorithm, following which the approximate distribution is constructed from the collected samples. While sampling-based methods are guaranteed to be identical to the true posterior under limiting conditions and can produce less biased results overall, they are quite computationally expensive and do not scale as well as variational Bayes methods do, as the corpus grows in size.

Due to the volume of data in the Gender Gap Tracker, we chose to avoid working with Gibbs sampling altogether. Instead, we used a *variational Bayes* inference model as implemented in Apache Spark³ for all our experiments.

2.3 Expectation maximization vs. online variational Bayes

The variational method proposed by [Blei, Ng, and Jordan \(2003\)](#) transforms a Bayesian inference problem to an optimization problem, i.e., an *expectation maximization* (EM) procedure that maximizes a

² In information theory, KL-divergence is the measure of the distance between two probability distributions.

³ [Large scale topic modeling: Improvements to LDA on Apache Spark](#)

lower bound with respect to the model parameters. However, due to the complex latent spaces typically seen in real-world corpora, it is quite common for variational EM methods to get stuck in local optima, resulting in the algorithm converging too slowly toward a global optimum or not at all.

An improved algorithm called *online variational Bayes* was developed to address this problem ([Hoffman, Blei and Bach, 2010](#)). The creators of this method note that, although variational Bayes (VB) inference methods are significantly faster than Gibbs sampling, they also suffer from computational difficulties for very large datasets. This is primarily because a standard VB algorithm must regularly switch between analyzing each observed batch and updating the dataset-wide variational parameters, which does not scale very well with enormous datasets consisting of millions of documents. The online VB algorithm achieves faster convergence to a global optimum through stochastic optimization. Moreover, online VB does not locally store the documents—each document arrives in a stream and can be discarded after use, greatly improving its speed and performance.

Due to the massive efficiency gains seen with online VB, we opted to use this method as implemented in Apache Spark for all our topic modelling experiments.

3 Related work

To further inform our work, we studied existing literature on topic modelling as applied to media and communication research.

3.1 Strengths and weaknesses of LDA

The key strengths of LDA are its scalability to extremely large corpora, and its ability to identify per-document topic proportions much more effectively than what human judgment might allow. However, it cannot qualitatively do better than a domain expert, due to its lack of human competence in performing comparative or contrastive analyses ([Puschmann & Scheffler, 2016](#)). A researcher who is intimately familiar with a small dataset would be able to perform a far more exhaustive and descriptive analysis than an algorithm could. The choices made by the user in removing particular words (i.e., stop word removal) and elimination of sparse terms that appear relatively infrequently can greatly alter the topics discovered. For analyses that span longer periods in time (particularly relevant to news corpora), word meanings might change based on events in the real world. In LDA, however, it is assumed that word meaning remains constant, so such words are treated as one and the same over the entire period. Modelling the data over longer periods can also risk missing signals that are very strong at a particular point in time. As a result, there is an implicit temporal component in a topic distribution that can significantly affect its human interpretation.

The bag-of-words assumption results in an LDA model considering only raw unigram frequencies, so any semantic relationship between the terms is lost ([Puschmann & Scheffler, 2016](#)). Existing work on addressing the bag-of-words assumption focuses on incorporating n-grams to form more meaningful associations between words. [Wallach \(2006\)](#) proposes a hierarchical Bayesian approach that combines a hierarchical Dirichlet bigram language model ([MacKay & Peto, 1995](#)) with a traditional unigram-based topic modelling approach that makes inferences using a Gibbs sampling procedure. It is important to note that, to make topic inferences effectively, word order can be very important — for example, the phrases “the department chair couches offers” and “the chair department offers couches” are identical in

terms of their unigram statistics. However, considering the word order in the first sentence makes it much more likely that it was generated by a topic that assigns high probability to words related to university administration ([Wallach, 2006](#)).

[Wang et. al \(2007\)](#) propose a topical n-gram (TNG) model for information retrieval that generates topics and topical phrases while considering word order. This model automatically determines phrases based on context through a Gibbs sampling inference procedure, following which it assigns a mixture of topics to both individual words and n-gram phrases. Although adding phrases increases the model's complexity, considering unigrams and bigrams based on context can significantly improve performance in text mining tasks. For example, the phrase "white house" carries a special meaning in relation to politics, whereas "yellow house" might simply appear in a document about real estate. The TNG model produces topics that are much more interpretable than LDA ([Wang et. al., 2007](#)), while also providing a probabilistic framework that helps linguists discover more meaningful phrases with the right context.

3.2 Developing a reliable methodology

Because of LDA's inherently non-deterministic and statistical qualities, methodological decisions involving data preprocessing, choosing the number of topics, tuning the algorithm parameters and evaluating the model's results become all the more important ([Maier et al., 2018](#)). In their article, Maier *et al.* describe best practices and lay out a reliable methodology for topic modelling on a corpus of documents pertaining to food safety. Because their data was extracted from the web, a significant amount of text cleaning was performed, followed by filtering the documents down to just the relevant ones for the task and de-duplicating them based on their similarity to one another. The authors stress the importance of a rigorous data cleaning step upstream of topic modelling.

The data cleaning step is followed by tokenization, i.e., the process of breaking down the text content and punctuation of each article into individual units called tokens. This is followed by lowercasing, which helps with term unification and reduces the overall vocabulary size by avoiding duplicate token counts. Punctuation (with the exception of hashtags, or other informative tokens depending on the scenario) is considered unimportant for the purposes of topic modelling, since the model does not seek to gain a semantic understanding of each and every term. The next step is stop-word removal, where commonly used prepositions, verbs or otherwise unimportant tokens are removed from the vocabulary. The dictionary of stop-words to be removed is chosen specifically for the corpus at hand, so that only those words which sufficiently represent a document's content are considered during topic modelling.

Two common techniques used in retrieval systems to make inflected words comparable to each other are *stemming* and *lemmatization*. Stemming removes both derivational suffixes as well as inflections that change the form of words to conflate word variants to the same roots or stems. On the other hand, lemmatization uses lexical and morphological analysis of a word to remove inflections, reducing it to its dictionary form or lexeme ([Balakrishnan and Lloyd-Yemoh, 2014](#)). For example, the word 'organized' is reduced to 'organ' when stemmed, but to 'organize' when lemmatized. Both stemming and lemmatization are found to trade off precision and recall in information retrieval systems, due to issues of pragmatics of word use rather than issues of linguistic morphology ([Manning et al., 2009](#)). However, from a topic interpretability standpoint, lemmatization is much more preferable to stemming.

Relative pruning is the final text cleaning step applied in the methodology of [Maier et al.](#) In this step, both very frequent and extremely rare words are stripped from the vocabulary prior to training. This is based on Zipf’s law, which, when applied to a dataset, states that a large fraction of words in a vocabulary occur extremely infrequently ([Manning & Schütze, 2003](#)). Performing relative pruning not only reduces the size of the corpus that the model has to work with, but also helps stabilize the LDA algorithm’s stochastic inference.

To quantify a model’s reliability and interpretability, a commonly used metric is *topic coherence*, which measures how frequently the top words of a topic co-occur ([Mimno et al., 2011](#)). Perplexity is another useful statistical measure to quantify the model’s goodness (or likelihood) of fit—a lower perplexity indicates better generalization and that there is a greater likelihood that the data arose from the given model ([Blei, 2012](#); [Asmussen & Møller, 2019](#)). However, it is emphasized that a lower perplexity measure does not necessarily translate to a higher topic interpretability. A two-step approach can be followed to evaluate a topic model’s word distributions—the initial step uses quantitative diagnostic metrics, such as topic coherence or perplexity, which is followed by a qualitative step that aims to match the model’s results with the theoretical concept in question ([Maier et al., 2018](#)).

4 Our methodology

4.1 Tools and resources

Our topic modelling pipeline uses Apache Spark’s⁴ parallelized LDA implementation⁵ via its Python interface, PySpark. The primary reason we chose Spark is related to performance and *horizontal scalability*. Our news data in the Gender Gap Tracker is continually growing, with new data being added daily through the use of automated scrapers, so we do not know upfront the number of articles that may appear in any given week or month. Our database statistics tell us that on average, we add roughly 800 to 1,500 English news articles every day, amounting to anywhere between 20,000 and 35,000 articles in any given month. In the 2-year period between October 2018 and September 2020, we scraped and stored the content of approximately 613,000 English articles in total.

Because our needs for topic modelling might vary with time, our primary focus is to build a robust, CPU and memory-efficient data and analysis pipeline that scales well. Horizontal scaling means that we are able to scale our computation by simply adding more distributed machines to our resource pool to handle the larger workloads. For this project, we utilize the extensive computational resources available on Simon Fraser University’s *Cedar* supercomputer⁶.

Spark’s DataFrame API

Spark comes with a robust and scalable LDA model available in one of two APIs—the RDD and the DataFrame API. It is important to note that the functionality in Spark’s DataFrame API is more modern and in many cases, offers better run time performance than the RDD API, due to internal optimizations on columnar aggregations and grouped operations via Spark DataFrames. In addition, for Spark versions 2.4 and above, there are significant differences in the available functionality between the two APIs. In

⁴ <https://spark.apache.org/>

⁵ [Databricks: Topic modeling with Latent Dirichlet Allocation](#)

⁶ [Supercomputer Cedar](#)

the recent past, Spark's DataFrame API has been more actively maintained and updated, so we work only with this API, with DataFrames being the primary data structure used.

Spark NLP

The LDA model in Spark is implemented as part of its machine learning library (Spark MLlib⁷). However, lemmatization, which is an important functionality we require for our pipeline, is absent in native Spark. Luckily, a third-party open-source library, Spark NLP⁸, is available that provides lemmatization as well as a host of other NLP utilities for use on top of Spark ML. We utilize Spark NLP in our pipeline to convert text into its lemma form prior to running LDA in Spark.

Learnable priors in Spark

It is important to note that in Spark's online variational Bayes LDA implementation, the parameter α that governs the document distribution over topics is *learned* during optimization—it does not need to be specified by the user. The same is true for the η parameter that governs the topic distribution over terms. By automatically learning these parameters during training, we avoid having to perform expensive grid search operations to find their optimum values. To compare results across runs, a constant random seed value (of 1) is specified for the pseudo-random number generator used during model fitting.

4.2 Preprocessing steps

By and large, we follow topic modelling best practices in the ordering of each preprocessing step as per [Maier et al. \(2018\)](#). Tokenization and normalization (i.e., removing all unnecessary symbols and artifacts) are performed simultaneously in Spark through the use of regular expressions that catch only relevant text while ignoring symbols and punctuation. This step is immediately followed by lowercasing. Stop-word removal is done prior to (and not after) lemmatization, primarily to reduce the vocabulary size upfront so that fewer lemma lookups are performed, improving performance.

A useful side-effect of removing stop-words before lemmatization is that we can selectively filter out specific singular forms of words while keeping their plurals that refer to a collective. For example, we include 'woman' as a stop-word, so in analyzing singular mentions like "a woman fell to her death from an apartment balcony", removing the singular form of the word 'woman' can help better interpret the larger theme (in this case, death). On the other hand, the plural 'women' is excluded from the stop-word list (which is retained and later lemmatized to its singular form). This allows us to capture the context of such collective terms in sentences like "the issue of women's equality is a significant one". On average, and with careful selection of stop-words, we believe that this approach helps us better disambiguate topics that should otherwise have no relationship to each other.

Figure 4 shows the sequence of preprocessing and feature transformation steps applied to each news article's text prior to training a topic model. The boxes are coloured according to the Spark module used for each task.

⁷ [MLlib | Apache Spark](#)

⁸ [Spark NLP: State of the art Natural Language Processing](#)

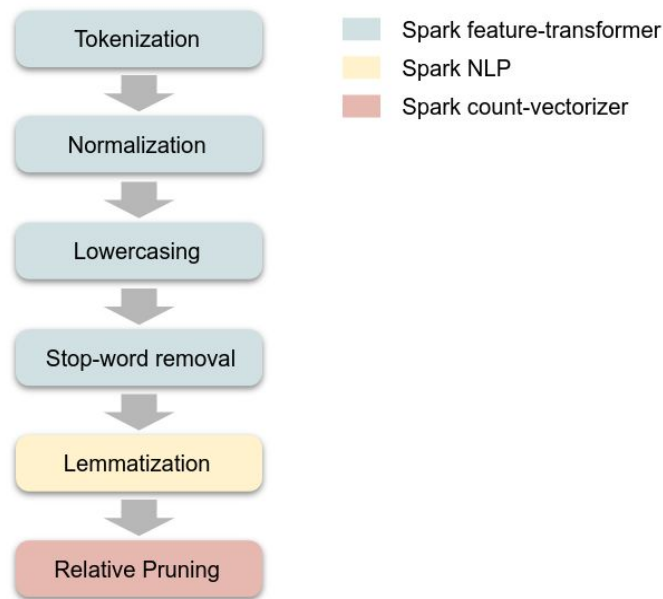


Figure 4. GGT topic model preprocessing and transformation steps

4.3 Stop words

We formulated our list of stop-words iteratively, by beginning with a list of standard stop words from the literature, including articles, pronouns and prepositions. We studied different formulations of *light verbs*, i.e., verbs that have little semantic content of their own, but instead form predicates with other expressions. Some examples are shown below.

have a rest, a read, a cry, a think
take a sneak, a drive, a walk, a plunge
give a sigh, a shout, a shiver, a pull, a ring
 ([Jespersen, 1965](#))

These verbs do not predicate fully, that is, one does not actually “take” a “plunge” but rather one “plunges”. However, they are not completely devoid of semantic content either: There is a clear difference between *take a bath* and *give a bath* ([Butt, 2003](#)). This makes such verbs semantically *light*, in a way that they are neither full nor empty in their semantic content.

From a topic modelling standpoint, light verbs do not add any value to topic interpretability. Because our news article data in the Gender Gap Tracker is unbounded in size and scope, we knew that it might take multiple iterations of topic modelling to isolate specific groups of light verbs relevant to our corpus. As a result, our light verb list was continually extended through the course of our experiments until we achieved satisfactory levels of topic interpretability.

Unlike in the case with verbs, we chose to retain a large portion of *nouns* during text preprocessing, due to the importance of nouns in topic interpretation. It has been shown that a combination of

lemmatization and limiting the corpus to just nouns provides a coherence advantage and lower word intrusion in topic modelling (Martin and Johnson, 2015). While we do still include some very common nouns in our stop-word list, these were chosen by carefully inspecting topic keywords for nouns that might cloud a human labeller's judgment when, later on, we assign labels to topics based on representative words for the topic (e.g., *people, man, woman, report, page, story*).

Table 1. Curated stop-words by category

| Category | Example words |
|-----------------------|--|
| Social-media related | <i>post, sign, like, love, tag, star, call, group, video, photo, pic, inbox</i> |
| URL and embeds | <i>http, https, href, ref, com, cbc, ctv, src, twsrc, 5etfw</i> |
| Frequent common nouns | <i>people, man, woman, family, friend, news, report, press, page, story</i> |
| Light verbs | <i>call, comment, continue, do, feel, give, get, take, like, make, tell, think</i> |
| Time of the day/week | <i>morning, afternoon, evening, today, yesterday, tomorrow</i> |
| Time periods | <i>day, week, month, year</i> |
| Time zones | <i>edt, est, pdt, pst</i> |
| Days of the week | <i>monday, tuesday, wednesday, thursday, friday, saturday, sunday</i> |
| Months of the year | <i>january, february, march, ..., october, november, december</i> |
| Year | <i>2018, 2019, 2020</i> |

Our final stop-word list, a sample of which is shown in Table 1, includes news media related artifacts from the web. It also includes URL terms, social media embed terms, days of the week and other nouns indicating time spans. Our reasoning in removing these terms is that news articles tend to describe real-world events with numerous time and date markers that do not in any way aid topic interpretability.

4.4 Custom lemma lookup table

Term lemmatization is performed using a lookup process during run time, where each token is checked for presence in a dictionary of lemmas provided by the user. We observed that the default lemma model provided by Spark NLP did not adequately resolve the range of vocabulary in Canadian English to the level required. For example, the verb *colored*, also spelled as *coloured* in Canadian English, was reduced to 'colore' and not *color* as we might expect.

To address such inconsistencies, we turned to the **spaCy** NLP library⁹, and its publicly available lemma lookup table on GitHub¹⁰. spaCy is an industrial-strength NLP library for Python that is well-maintained and documented to the highest quality. The lemma list maintained by spaCy is quite extensive and contains both British and American spellings, and because Canadian English tends to be based on both British and American English (Dollinger, 2019), this makes it ideal for our use.

⁹ <https://spacy.io/>

¹⁰ <https://github.com/explosion/spacy-lookups-data>

The existing format of spaCy's lookup table as available on GitHub was not suitable for use in our Spark pipeline, so we reformatted it as shown in Table 2. The text on the left of the delimiter (->) is the word lemma, while the text on the right represents the various inflectional forms of the word. This custom lemma lookup table is then used by Spark NLP's lemmatization routine during preprocessing.

Table 2. Snippet of our reformatted lemma lookup table from spaCy

```
color -> color colored coloring colors
colorimeter -> colorimeter colorimeters
colorize -> colorize colorized colorizes colorizing
colossus -> colossus colossi colossuses
colostomy -> colostomy colostomies
colostrum -> colostrum colostrum colostrums
colourant -> colourant colourants
colouration -> colouration colourations
colour -> colour coloured colouring colours
```

4.5 Limitations of our approach

Our topic modelling pipeline was built to scale to huge amounts of data, so some compromises had to be made to maintain robustness, interpretability and performance.

Corpus is decomposed to a collection of unigrams (no bigrams)

During feature transformation, a common approach is to more semantically tokenize the corpus through the use of bigrams. Words that regularly occur together, such as “United” and “States” should, ideally, be tokenized as “united_states” using a master dictionary that defines these terms. We opt against using such bigrams in our workflow mainly because of concerns with memory requirements, owing to the size of data involved. In this study, we consider only unigrams—for example, the terms “United States” and “Islamic State” are tokenized, lowercased and lemmatized to “united”, “islamic” and “state”. While this is not ideal, we reason that this does not have a drastic impact on topic interpretability, and so is acceptable for our purposes.

Difficulty in using custom quantitative metrics

Although external libraries such as Gensim do provide clean, well-tested implementations of popular topic quality metrics such as *Pointwise Mutual Index* (PMI) and ‘UMass’ *topic coherence* ([Mimno et al., 2011](#)), these do not easily fit in with the distributed nature of Spark’s data structures. Spark’s own LDA model does not come with these metrics built-in. In addition, Spark’s interface with the Java Virtual Machine (JVM) makes it non-trivially complicated to implement custom user-defined metrics in Python. As a result, for evaluation of topic quality, we stick to using Spark’s inbuilt perplexity-bound measure, as well as our own subjectivity.

Partitioning the LDA model can have consequences

We perform model training, evaluation (i.e., measuring perplexity) and aggregation/grouping operations on the topic model all in one single sequence. Because Spark is inherently distributed in nature, the

model's ability to reliably make predictions on the topic distribution of a new, unseen document depends on it being partitioned the same way as it was during training. If a trained 'distributed' LDA model in Spark is converted to a 'local' (unpartitioned) model, this could change the model's performance due to shuffling data between its partitions. For this reason, we opt to train new, separate topic models for each set of articles over the time span of interest, as opposed to training a single model over a longer time span and using it to make topic predictions on unseen data.

Perfect reproducibility is not guaranteed due to Spark's distributed nature

A common practice to ensure the reproducibility of random samples in LDA is to fix the random seed across runs. This should, in principle, work well, since it ensures that the Dirichlet samples are drawn from the same random distribution across runs. However, because of the distributed nature of Spark, there is an additional layer of randomness in the methodology—i.e., that of the partitioning system. Due to internal quirks with the way Spark distributes data, the order of samples passed to an executor cannot be guaranteed, nor can the same random number generator function be used on each and every executor. As a result, this stage of the process is non-deterministic. This causes a minor deviation in topic separation across runs, even with the same random seed on the exact same data. However, we show results from our extensive experiments (in Section 5) to verify that this does not drastically affect topic interpretability or the overall reliability of the process.

5 Experiments

To arrive at a robust and reliable workflow for topic modelling, we performed numerous experiments, some of which are described in this section. Maximum perplexity bounds were computed, indicative of the log-likelihood of a model's fit to the given corpus. The same curated stop-word and lemma lists shown in Sections 4.3 and 4.4 were used throughout all experiments described in this section.

As is commonly done in topic modelling studies, the "goodness" of a model is ultimately evaluated using subjective human judgment through a careful inspection of the topic distribution and keywords. The interpretation of what defines a topic is based on word co-occurrence and requires a human-level semantic understanding of the text. For all experiments, topic labels are assigned as per our custom guidelines defined in Section 6.

5.1 Number of topics

Keeping all other hyperparameters constant, we train three topic models varying just the topic number ($k = 10$, $k = 15$ and $k = 25$) on **one month's worth** of news articles from all seven Canadian English-language outlets. The month chosen was July 2019, consisting of approximately 29,000 articles of varying lengths. The resulting keywords are tabulated and human-labelled (using the top 15 words for the topic) to generate topic labels for the word distributions in each case. A summary of this comparison is shown in Table 3.

Table 3. Comparison of topic labels for 10, 15 and 25 topics

| 25 topics | 15 topics | 10 topics |
|-----------|-----------|-----------|
|-----------|-----------|-----------|

Accidents and fire incidents

Arts and entertainment

Aviation incidents

Business and market events

Cannabis and health

Community and indigenous programs

Crime and police investigations

Crime and police investigations

Education programs

Education programs and research

Federal election campaign

Healthcare

Highway safety

Legal issues and court cases

Local food, restaurants and shopping

Local politics

Maritime events and updates

Provincial energy budgets/planning

Provincial energy budgets/planning

Sports (mainstream)

Summer festivals and concerts

US politics

Weather and parks

World politics

World politics

Accidents and aviation incidents

Arts and entertainment

Business and market events

Crime and police investigations

Federal election campaign

Healthcare

Indigenous policy/government

International protests and violence

Legal issues and court cases

Provincial projects/planning

Sports (mainstream)

Sports (summer)

US politics

Weather and parks

World politics

Accidents and aviation incidents

Arts and entertainment

Business and market events

Crime and police investigations

Federal politics and elections

Healthcare

Legal issues and court cases

Provincial projects and planning

Sports (mainstream)

US politics

Red topics are those that are repeated (with similar keyword distributions)

Blue topics are those that are nearly the same across all three cases

We know from the real world that articles published by news outlets fall into broad categories, such as sports, business or politics. Our goal with this experiment is to judge the level of granularity in our model's discovered topics, and whether this is sufficient for our purposes in studying the relationship between topics and the gender distribution of people quoted. As can be seen in Table 3, all three cases show a good degree of topic separation, meaning that the model is capturing realistic themes from real-world news categories. It is important to note, however, that a topic model does not capture semantics of any kind, so the topics themselves may not perfectly correspond with a news outlet's categories from the real world (e.g., international news, business or sports).

The labels in Table 3 indicate that, when 25 topics are used, a certain amount of repetition is present. On the other hand, with 10 topics, some important topics that may emerge in a given month, but not be present across time, are lost. We have found those two trends in multiple experiments with the 25-15-10 topic numbers. Due to issues with repetitive or non-existent topic labels and the difficulty of

labelling a large number of topic keywords by hand, all our experiments going forward use **15** topics ($k = 15$).

5.2 Random seed

Because of the distributed nature of Spark, it is difficult to ensure that the same random number generator gets used across all executors, or that the order of samples being fed to the executor is fixed during model training. To test this, we run some experiments to study the stability of the LDA model over multiple runs for the month of March 2019. This particular month was chosen because it exhibits some interesting events that were of international importance, such as the aftermath of the Boeing 737 Max aviation disaster¹¹, as well as the New Zealand mosque shootings¹². Our goal is to see whether multiple LDA models with different random seeds can consistently capture the thematic structure of such events.

We first vary the random seed in Spark to three different (arbitrarily chosen) values: 1, 99 and 340573. The resulting topics, as interpreted by a human, are shown in Table 4. Note that, although the resulting topic word distributions are not identical across the different models, some domain knowledge of key world events that month, combined with some subjective judgment are sufficient to label the topics. It is clear that certain key events covered in the news that month, including the New Zealand mosque shootings, the Boeing 737 Max aviation disaster and the SNC-Lavalin political scandal, are well-captured in all three models. The non-deterministic nature of the LDA model in Spark does, however, tend to fuse together two different topics (or introduce new topics altogether) into the top 15 topics for a given month (e.g., 'Education & medical research').

Because we limit each model's results to just 15 topics (for ease of labelling), we do not expect that our methodology captures *all* possible topics for a given month. Our primary requirement is that larger, more important themes that compose a given month's timeline be captured as far as possible, which we observe to be true in these results. We tested two other months of data using the same three random seeds and observed similar trends with minimal loss of topic interpretability (with small amounts of word intrusion resulting in the merging of multiple smaller topics in certain cases).

Table 4. Comparison of topic labels obtained using three different random seed values (March 2019)

| Random seed: 1 | Random seed: 99 | Random seed: 340573 |
|---|---|---|
| Arts & entertainment | Arts & entertainment | Arts & entertainment |
| Boeing 737 Max aviation disaster | Boeing 737 Max aviation disaster | Boeing 737 Max aviation disaster |
| Business & market events | Business & market events | Business & market events |
| Crime & police investigations | Crime & police investigations | Education programs & budgets |
| Education & medical research | European politics | Federal politics |
| Community infrastructure | Federal politics | Healthcare & medical research |
| Lifestyle | Healthcare & medical research | Legal & court cases |

¹¹ [CBC News: Canada grounds Boeing 737 Max 8](#)

¹² [CTV News: PM Trudeau condemns fatal shootings at mosques in New Zealand](#)

| | | |
|-------------------------------------|-------------------------------------|-------------------------------------|
| New Zealand mosque shootings | Lifestyle | Lifestyle |
| Provincial politics & programs | New Zealand mosque shootings | New Zealand mosque shootings |
| Severe weather updates | Provincial politics & programs | Provincial politics & programs |
| SNC-Lavalin scandal | Severe weather updates | Severe weather updates |
| Sports | SNC-Lavalin scandal | SNC-Lavalin scandal |
| Transport & highway safety | Sports | Sports |
| US politics | US politics | US politics |
| World politics | World politics | World politics |

Topics marked in orange show slightly different word intrusion and topic separation across cases

The next set of experiments are to study the repeatability of our modelling results. This time, we rerun the same model training step **thrice**, using the same random seed of 1. The resulting human-labelled topics are shown in Table 5.

Table 5. Comparison of topic labels obtained over three runs of a single random seed (March 2019)

| Random seed: 1 | | |
|---|---|---|
| Run 1 | Run 2 | Run 3 |
| Arts & entertainment | Arts & entertainment | Boeing 737 Max aviation disaster |
| Boeing 737 Max aviation disaster | Boeing 737 Max aviation disaster | Business & market events |
| Business & market events | Business & market events | Community infrastructure |
| Crime & police investigations | Consumer products & technology | Crime & police investigations |
| Education & medical research | European politics | Federal politics |
| Community infrastructure | Federal politics | Healthcare & medical research |
| Lifestyle | Healthcare & medical research | Legal & court cases |
| New Zealand mosque shootings | Legal & court cases | Lifestyle |
| Provincial politics & programs | Lifestyle & education | New Zealand mosque shootings |
| Severe weather updates | New Zealand mosque shootings | Provincial politics & programs |
| SNC-Lavalin scandal | Provincial politics & programs | Severe weather & transport safety |
| Sports | SNC-Lavalin scandal | SNC-Lavalin scandal |
| Transport & highway safety | Sports | Sports |
| US politics | Transport & highway safety | US politics |
| World politics | US politics | World politics & violence |

Topics marked in **orange** show slightly different word intrusion and topic separation across cases

As expected, fixing the random seed does not result in perfect reproducibility of the topic labels across multiple runs. While this is not ideal, a closer inspection of the labels indicates that the majority of topics are retained across all cases (including the key transient events for the month, marked in bold). Certain topics exhibit a small amount of overlap, combining keywords from multiple topics (e.g., 'Lifestyle & education', and 'Severe weather & transport safety'). However, this only seems to occur for 'minor' topics that do not feature in that many articles overall (minor topics are those that have weak topic weight intensities across all outlets, shown in more detail in [Section 7.1](#)).

Based on the random seed experiments, we confirm that in Spark, there is an inherent difficulty in producing deterministic topic model results, even with the same random seed on the exact same data. However, considering that our overall goal is to study topic gender breakdown on select topics that feature strongly for any given month, we find that the trade-off between the reproducibility and scalability using our methodology is a reasonable one.

5.3 Hyperparameter tuning

This section discusses our observations from the LDA hyperparameter tuning experiments. Each hyperparameter is changed individually, and the quality of the result is measured using a combination of the model's perplexity and human judgment. In Spark, model perplexity is represented by the logarithm of the upper bound on the negative log-likelihood of tokens, divided by the number of tokens in the corpus after relative pruning. A lower perplexity means that the model achieves a better-quality fit to the given data. A "good" topic distribution is defined as one that has *low perplexity* and is *highly interpretable*, with *good topic separation* and *limited word intrusion* (i.e., not too many occurrences of the same words across many topics).

5.3.1 Maximum iterations

The online variational Bayes inference algorithm used in Spark converges quite rapidly (within 100-150 iterations)—this is in contrast to expectation maximization, which we observed can take upwards of 200 iterations to converge. This makes logical sense, since the online algorithm is known to speed up convergence by using stochastic optimization and is less prone to getting stuck in local optima. We did not observe any noticeable differences in terms of perplexity as well as topic interpretability when running the online algorithm for 150 iterations and above. Minor differences in topic keywords were noticed for specific months' results when we used just 100 iterations. As a result, we fix the maximum number of iterations to **150** for all our training runs, regardless of the size of the data involved.

5.3.2 Relative pruning

There are three relative pruning hyperparameters used during data preprocessing: *maximum document frequency*, *minimum document frequency*, and *minimum term frequency*. Controlling these values prunes, i.e., removes those terms that occur both very frequently and very rarely from the vocabulary. As a

result, these hyperparameters must be chosen carefully, so as to not hinder topic interpretability or miss larger themes in the data.

Maximum document frequency

We found that a maximum document frequency of **80%** ($maxDF = 0.8$), in which we prune those tokens that occur in more than 80% of all documents in the corpus, worked well over multiple experiments. This value resulted in minimal loss of information while retaining a good deal of topic interpretability and separation. Too high a value ($maxDF > 0.9$) resulted in a lot of words repeating across topics, increasing word intrusion (and hence perplexity), while reducing topic interpretability.

Minimum document frequency

The minimum document frequency hyperparameter ($minDF$) has a significant impact on tokens that appear very rarely in the data. When setting $minDF = 0.05$, i.e., pruning those tokens that occur in fewer than 5% of all documents in the corpus, we observed to our surprise that there was a significant drop in model perplexity. This was accompanied by a significant loss of topic interpretability, with multiple topics repeating very similar word distributions. Upon closer inspection of the topic keywords, we saw that by removing up to 5% of the least frequent tokens in the data, we were removing a large portion of the useful nouns and adjectives that aided topic interpretability. As a result, the model's perplexity dropped, due to a smaller vocabulary overall and the fact that the model was overfitting to a much-diminished distribution. In the end, we obtained our best results using a minimum document frequency of **2%** ($minDF = 0.02$). This minimizes the loss of crucial low-frequency tokens and improves overall topic interpretability.

This result reinforces the notion that **a model with lower perplexity is not necessarily a better one** ([Maier et al., 2018](#)).

Minimum term frequency

This hyperparameter prunes terms that occur fewer than a specified number of times *within* a given document. Unlike the document frequency hyperparameters (which are specified as percentages relative to the whole corpus), the minimum term frequency in Spark is specified as a positive integer. A minimum term frequency of 1 means that no terms are pruned from a document, whereas setting $minTF = 2$ means that all terms that occur less than twice (i.e., just once) within a document are pruned. Our experiments with $minTF > 1$ once again resulted in the model having a lower perplexity—however, this is primarily because pruning terms that occur once per document causes a huge reduction in the number of available terms (many terms in a document occur just once). Recall that, at this point, we have already removed stop words, so most of the words left in the document have meaningful content. This loss of information is undesirable, and results in niche, overly-specific topics being discovered in the data. Thus, we fall back to keeping the minimum term frequency as **1** so that no terms within a document are pruned, giving the model access to the full set of terms in each document.

5.3.3 Vocabulary size

Setting a finite vocabulary size limits the dimensionality of the random matrices that are solved for during LDA (i.e., θ and β) based on the relative frequencies of tokens in the corpus. In principle, the

LDA algorithm can be run on datasets that are unbounded in size, and as a result, drawing Dirichlet samples from larger and larger distributions of words adds more noise to the inference process. Our survey from existing literature showed that a typical vocabulary size used in LDA ranges from 5,000 to 10,000, regardless of the size of the corpus. We tested the impact of three different vocabulary sizes in training models over one month's worth of data ($vocabSize = 5000 \mid 10000 \mid 20000$).

Our results showed that a larger vocabulary size has close to no impact on model perplexity and a negligible impact on interpretability. The increased sample size from which Dirichlet distributions are drawn resulted in slightly poorer topic separation and slower convergence toward the optimum. Because of its relative lack of impact on model quality, we choose to limit the vocabulary size in our models to **5,000**.

5.4 Time span of data modelled

Because our overall goal is to study the relationship between topics covered in the media and the gender of people quoted, we also look at the effect of time span considered on the topics discovered. We would expect that running a topic model on several hundred thousand articles representing news coverage over a year's time would yield quite different topic labels from one that runs on just a month's worth of data. To study this further, we trained a series of models, over a 1-month, 3-month, 6-month and 12-month period.

Table 6. 12-month topics (Apr 2019 - Mar 2020) and 6-month topics (Oct 2019 - Mar 2020)

| 12 months | 6 months |
|---|---|
| Arts and entertainment | Arts and entertainment |
| Business and market events | Business and market events |
| Community infrastructure | Consumer products, restaurants and services |
| Consumer products, restaurants and services | Crime and police investigations |
| Crime and police investigations | Federal politics and election campaign |
| Federal politics | Government policy |
| Government policy and human rights | Government policy and human rights |
| Healthcare and Covid-19 | Healthcare and Covid-19 |
| Provincial education policy and programs | Healthcare and medical research |
| Provincial projects and planning | Local businesses |
| Public affairs and unions | Provincial education policy and programs |
| Public events | Sports |
| Sports | US politics |
| US politics | Weather and natural disasters |
| World politics | World politics |

Our experiment looks at the topic coverage before and during the COVID-19 pandemic that emerged in early 2020. The 12-month period considers all articles between April 2019 and March 2020, while the 6-month period considers all articles between October 2019 and March 2020. Table 6 shows the human-labelled topics from these two periods. Both periods largely reveal themes that are regularly

covered in the news, such as “Business and stock market”, “Arts and entertainment”, “Sports” and “Federal politics”. It is interesting that the term ‘covid-19’ appears in the topic distributions even for the 12-month span dating back to April 2019—this is primarily because COVID-19 was a global crisis that dominated news coverage in Canada through the early period of 2020, making its terms co-occur very frequently with the “Healthcare” word distribution. This is an undesirable result, as it makes it sound like COVID-19 was present going back to April 2019, which is not the case. Looking deeper at the remaining word distributions and their associated topic labels, we find that no fine-grained labels exist over these large time periods. Smaller and more transient events, expectedly remain absent from the topic labels for this long a time span.

Table 7. 3-month topics (Jan - Mar 2020) and 1-month topics (Mar 2020)

| 3 months | 1 month |
|--|--|
| Arts and entertainment | Arts and entertainment |
| Business and market events | Business and market events |
| Community infrastructure | Covid-19 and healthcare guidelines |
| Coronavirus outbreak | Covid-19 and local communities |
| Covid-19 healthcare initiatives | Covid-19 and provincial updates |
| Covid-19 jobs and education support programs | Covid-19 and travel |
| Crime and police investigations | Covid-19 healthcare and support programs |
| Event cancellations and postponements | Covid-19 tracking and updates |
| Federal politics | Covid-19 business and market impact |
| Healthcare and medical research | Crime and police investigations |
| Indigenous rights and government policy | Event cancellations and postponements |
| Iran aviation disaster and political events | Government policy and support programs |
| Provincial education policy and programs | Sports |
| Severe weather and travel safety | US politics |
| Sports | World politics |

Orange topics are more localized in time, representing more specific events or issues

We then compare the topic distributions from a 3-month period (January 2019 - March 2020) and the 1-month period through March 2020, as shown in Table 7. Unlike the longer time spans, topics from these periods show much more fine-grained labels. As expected, COVID-19 and its related terms dominate the distribution, but even within this larger context, the model is able to disambiguate keywords from more fine-grained themes, such as “COVID-19 and travel” and “COVID-19 business and market impact”. Key world events such as the Iran aviation disaster in January 2020, including the political friction between the US and Iran in January/February 2020 emerged as a topic for the 3-month period.

From our time span topic experiments, we observed that news outlets typically spend a few days or weeks focusing on a particular event or issue, depending on its severity or importance. For transient events such as aviation or natural disasters that have a big impact on local communities, we feel that it is

both insightful and important to model topics over shorter time spans for our source gender analysis. As a result, we settle on a **monthly** topic modelling pipeline for our further analyses and visualizations.

5.5 Final fine-tuned hyperparameters

Our fine-tuned topic modelling workflow in production involves a semi-automated process that trains individual topic models on **1 month's worth** of data at a time. We use the best combination of model hyperparameters (shown in Table 8), and a fixed random seed of **1**, based on our experiments shown in this section. Our data preprocessing methodology incorporates the curated list of stop-words shown earlier in Table 1, as well as a custom lemma lookup table based on the one from the spaCy NLP library.

Table 8. Best hyperparameters found for LDA on monthly news data

| Hyperparameter | Value |
|----------------------------|-------|
| Number of topics | 15 |
| Maximum iterations | 150 |
| Vocabulary size | 5000 |
| Minimum term frequency | 1 |
| Minimum document frequency | 2% |
| Maximum document frequency | 80% |

6 Topic labelling guidelines

Our topic model pipeline is designed as a monthly semi-automated process. On the first day of every month, a new topic model is trained on the previous month's English-language articles from seven outlets. The top 15 topic words for each topic (along with their topic weights) obtained from the LDA model are written to a database, following which they are human-labelled and visualized in greater detail. We maintain a fixed value of **15** topics a month for consistency across months and ease of labelling. Below, we show a few topic labelling guidelines we define as part of our methodology. These guidelines are detailed, and experiments over the last few months have shown that the labels are reproducible over multiple rounds of topic modelling and human-labelling.

6.1 Naming patterns

We adopt a flexible topic naming pattern, in which a given distribution of keywords is interpreted (as best as one can) based on the larger themes that the words cover. Because not all topics can be described in 3-4 words, we occasionally use up to 5-6 words to label a topic more clearly. Some examples are shown in Table 9.

Table 9. Typical naming patterns used in topic labelling

| Keywords | Topic label |
|----------|-------------|
|----------|-------------|

vote, party, election, candidate, voter, campaign, liberal, poll, leader, political, quebec, seat, conservative, support, tory Federal & provincial election campaigns

police, officer, floyd, protest, death, rcmp, charge, george, arrest, incident, black, protester, force, street, investigation George Floyd protests & police investigations

6.2 Specificity

Rather than fixating too much on a single word (or pair of words) to identify a topic, we instead look at entire groups of words to identify larger themes. As an example, consider the keywords '*alberta, oil, price, energy, gas, industry, workers*'. It is clear from the keywords that there is a strong focus on energy as well as the oil and gas sector and its workers, so rather than choosing a vague label such as '*Provincial policy*', we assign it the label '*Energy policy and jobs*'. It is important to remember that there is no hard and fast rule to assigning good labels—this is ultimately down to subjectivity, domain knowledge and human judgment.

In certain months, we observe keywords from different subtopics appearing across multiple topics—for example, different kinds of sports. We avoid assigning the exact same topic label, i.e., '*Sports*' to such cases. To be as specific as possible, we disambiguate the names of the sports by inspecting the keywords and labelling them explicitly, for example, '*Sports (Grey Cup & CFL)*' and '*Sports (Hockey & basketball)*'.

6.3 Topic label reuse

Certain topics with similar keyword distributions appear again and again, regardless of the month of the year. Whenever possible, we reuse past topic labels for word distributions that are quite similar (for the most part), as shown in Table 10. This helps maintain consistency across months and allows for easier comparison of topic trends over time.

Table 10. Repeating topic labels we tend to reuse as-is (for similar keyword distributions)

| Topic label | Typical keywords |
|---------------------------------|---|
| Arts & entertainment | <i>book, film, black, part, award, gallery, art, music, movie, toronto</i> |
| Business & market events | <i>company, market, bank, trade, sell, buy, billion, stock, investor</i> |
| Community infrastructure | <i>province, cost, million, project, housing, budget, pay, transit, road</i> |
| Crime & police investigations | <i>police, officer, rcmp, investigation, victim, arrest, kill, die, suspect</i> |
| Lifestyle | <i>child, woman, mother, young, house, daughter, community, experience</i> |
| Federal politics | <i>party, trudeau, liberal, conservative, candidate, vote, ndp, campaign</i> |
| Healthcare & medical research | <i>health, study, research, care, patient, hospital, medical, drug, case</i> |
| Highway & transport safety | <i>driver, car, truck, system, safety, traffic, drive, vehicle, crash, injury</i> |
| Jobs, education & worker unions | <i>worker, community, union, school, student, project, strike, board</i> |

| | |
|------------------------|--|
| Legal & court cases | <i>court, case, judge, lawyer, charge, decision, justice, legal, appeal</i> |
| Sports | <i>game, team, season, player, hit, point, shoot, goal, coach, win</i> |
| US politics | <i>president, unite, trump, state, house, republican, administration</i> |
| Severe weather updates | <i>snow, park, water, fire, winter, road, ice, high, heat, temperature</i> |
| World politics | <i>country, international, minister, national, border, china, unite, state</i> |

7 Topic dashboard

Based on the LDA methodology described in [Section 2](#), we train a **monthly** topic model, on the previous month's data each time, over a period from July 2020 (at the time of writing) all the way back to October 2018. Using this approach, we are able to characterize each document (i.e., an individual news article) for a particular month as belonging to a distribution over topics. Mathematically, this means that a topic model returns a one-dimensional vector of length 15 (we only model a maximum of 15 topics each month) for each document, where each component of the vector represents how strongly or weakly that topic's keywords are associated with that document. Some examples of how these results are represented in our database are shown in Table 11.

Table 11. Example output snippet from topic modelling (Bold weights indicate dominant topics)

| Article ID | Outlet | # Female sources | # Male sources | Topic weight distribution [t1, t2, ... , t15] |
|------------|--------------------|------------------|----------------|---|
| 1 | CBC News | 1 | 0 | [0.996 , 0.002, ... , 0.0001] |
| 2 | Huffington Post | 3 | 1 | [0.0002, 0.992 , ..., 0.0001] |
| 3 | The Globe and Mail | 0 | 2 | [0.0001, 0.0003, ..., 0.675] |
| 4 | CTV News | 1 | 3 | [0.0001, 0.995 , ..., 0.0003] |
| ... | ... | ... | ... | ... |

We know from our analysis of statistics from the Gender Gap Tracker ([Figure 2](#)) that, on average, there are roughly 3-4 times more articles with a majority of male sources than those with a majority of female sources ('majority' here means at least one more source from one gender than the other). As a result, rather than looking at raw counts of articles and which topics dominate in them, we perform *aggregations* (i.e., we compute averages) of the topic distributions over the outlets and/or dominant gender quoted. Because we have at least a few thousand articles each month that quote more female sources than male, we feel that we have a large enough sample size to draw reasonable conclusions from.

All our results, once computed, are visualized on an [interactive dashboard](#)¹³ for easy exploration. The dashboard serves a dual purpose, of both allowing us to explore topics and gender breakdown, as shown in [Section 8](#), and providing visualizations for journalists, news organizations and the public.

¹³ <https://gendergaptracker.research.sfu.ca/apps/topicmodel>

7.1 Topic intensity

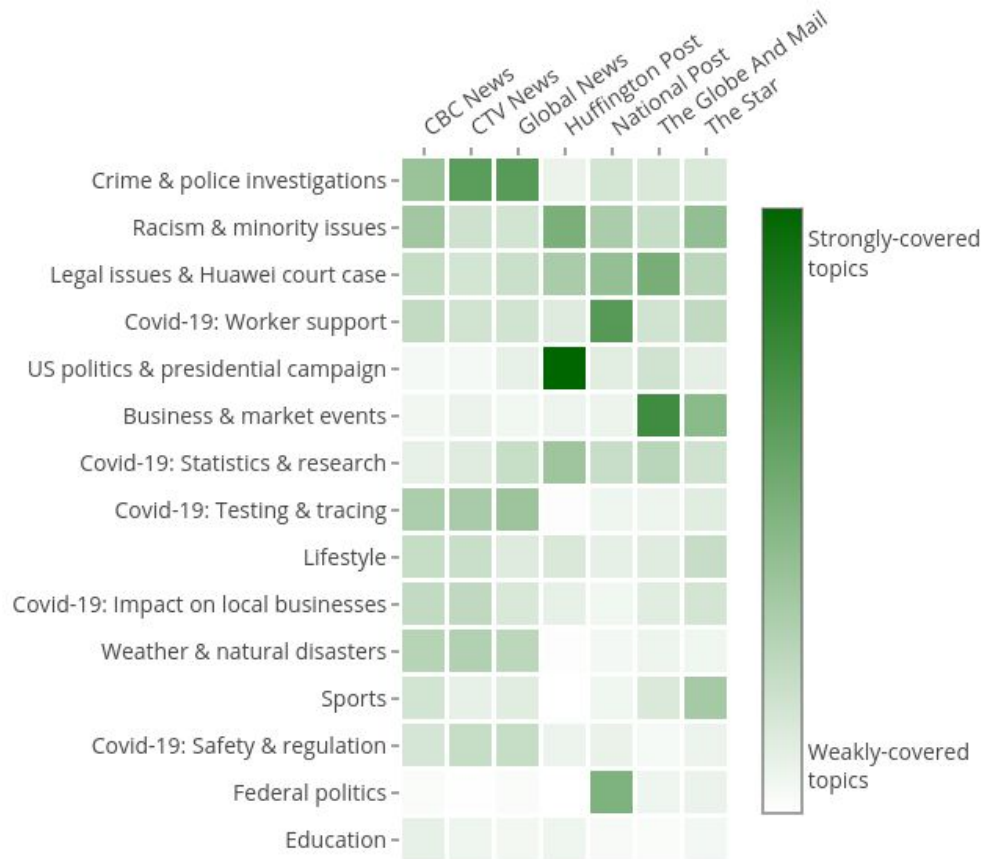


Figure 5. Heat map of mean topic intensity per outlet for news articles in July 2020

As a first step, we are interested in answering the question: *Which topics were covered more extensively by each outlet in a particular month?* To do this, we first group our results by outlet, and compute the element-wise mean topic weight for all articles from that outlet. This returns a [15 x 7] matrix, representing the mean topic weights over 15 topics for articles from all seven English news outlets in that month. This matrix is easily visualized as a heat map as shown in Figure 5.

The heat map of mean topic intensity is ordered by the sum of means for all outlets, with the most strongly covered topic for that month (on average, across all outlets) appearing on top. From Figure 5, for the month of July 2020, it is clear that a large proportion of articles contain keywords pertaining to crime, police investigations, and racism and minority issues. Relatively fewer articles contain keywords pertaining to the lifestyle, sports, or education topics.

7.2 Topic gender prominence

In order to study the difference in representation between male and female sources for each topic discovered, we perform one additional step prior to aggregation. The corpus of articles shown in Table 11 is separated into two smaller corpora—those with majority male sources, and those with majority

female sources¹⁴. The majority condition is easily calculated by comparing the columns containing the source counts for either gender from Table 11. We refer to these corpora as the **female** and **male** corpora from this point on. Next, we once again group our results by outlet and aggregate the topic weights, but this time, we do so for *each corpus* (with male/female majority sources) separately. This results in two [15 x 7] matrices (one for either corpus), each representing the mean topic weights over 15 topics for articles from the seven outlets.

Per outlet gender prominence

Here, we introduce the term ‘gender prominence’ to help disambiguate how different topics are related to the number of female/male sources quoted. For the purposes of this study, we define gender prominence as the difference in mean topic weights between the female and male corpora 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. Similarly, a topic can be said to exhibit female prominence if the mean topic weights from the female corpus are greater than those for the male corpus. Mathematically, this is calculated as the element-wise difference between the two [15 x 7] topic weight aggregation matrices. A positive difference indicates that the topic exhibits female prominence, whereas a negative difference indicates male prominence. Figure 6 showcases this result as a heat map for the topics discovered in July 2020.

The heat map of topic gender prominence uses a divergent colour scale. Topics that are at the extreme ends of the heat map (‘Lifestyle’ at the top and ‘Sports’ at the bottom) exhibit the strongest disparity per-topic in the mean topic weights of the female/male corpora. For July 2020, the ‘Lifestyle’ topic was much stronger (i.e., had a higher mean topic weight) in the female corpus, leading to a greater positive difference (red) between the topic weight matrices. Conversely, the ‘Sports’ topic was much stronger in the male corpus, leading to a greater negative difference (blue) between the topic weight matrices.

¹⁴ We define the ‘majority’ condition here as any case where the number of sources from one gender is **one or more** greater than the number of male sources from the other gender. For example, an article with 3 female sources and 2 male sources is categorized as “female-majority”, and is assigned to the female corpus.

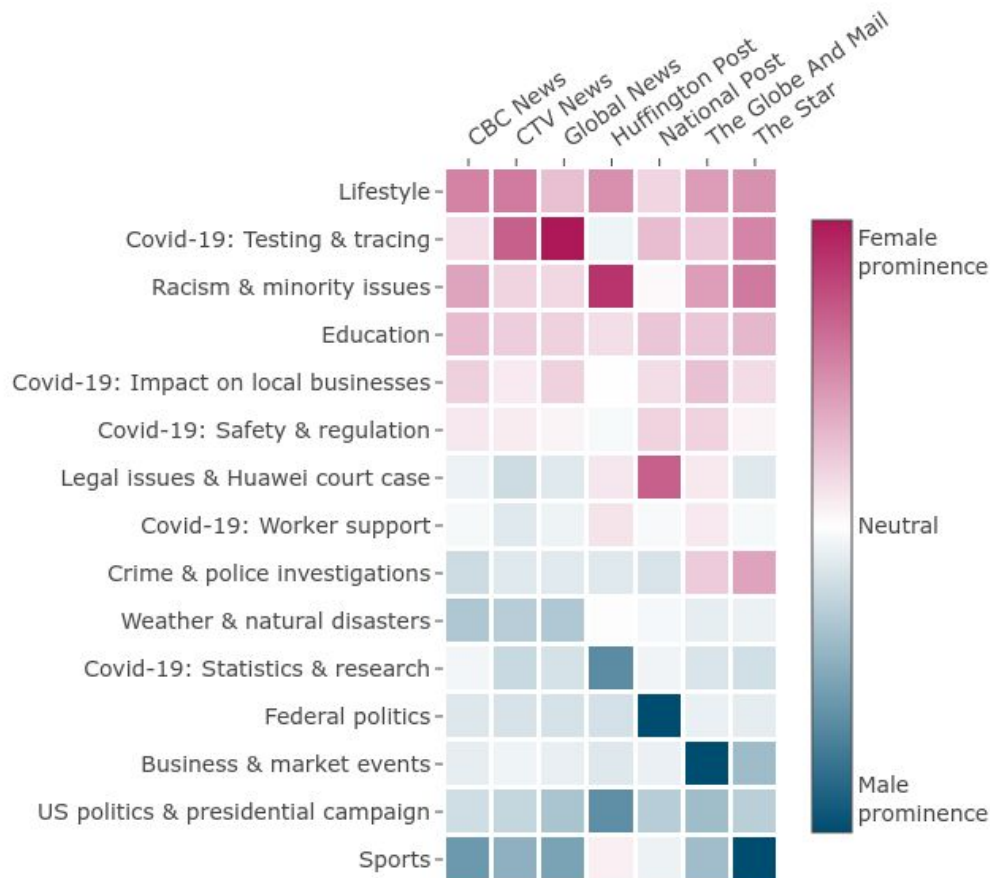


Figure 6. Heat map of mean topic gender prominence per outlet for news articles in July 2020

Note that in the gender prominence heat map, a zero value (white) indicates neutrality. A topic can be 'gender-neutral' in one of two ways. First, there might exist true parity in topic intensity between the two corpora (with female/male-majority sources), where both corpora exhibit the same mean topic weight, leading to their difference being zero. Alternatively, the topic might just have been non-existent for that particular month, resulting in both the male and female corpora showing a zero topic weight for that topic.

Overall gender prominence

In addition to the heat maps, we also provide bar charts of mean topic weights across *all* outlets for the female and male corpora, as shown in Figure 7. The data for these plots is obtained by simply aggregating topic weights over all articles in either corpus (i.e., with female or male-majority sources).

The overall gender prominence plots shown are ranked in decreasing order of mean topic weights. Higher values indicate that the topics were particularly strongly covered for the corpus in question. For July 2020, we can see that topics such as 'COVID-19: Testing & tracing' and 'Lifestyle' show the highest gender prominence in the corpus with majority female sources, whereas topics such as 'Legal issues & Huawei court case', 'Business & market events' and 'Sports' show higher gender prominence in the male corpus.

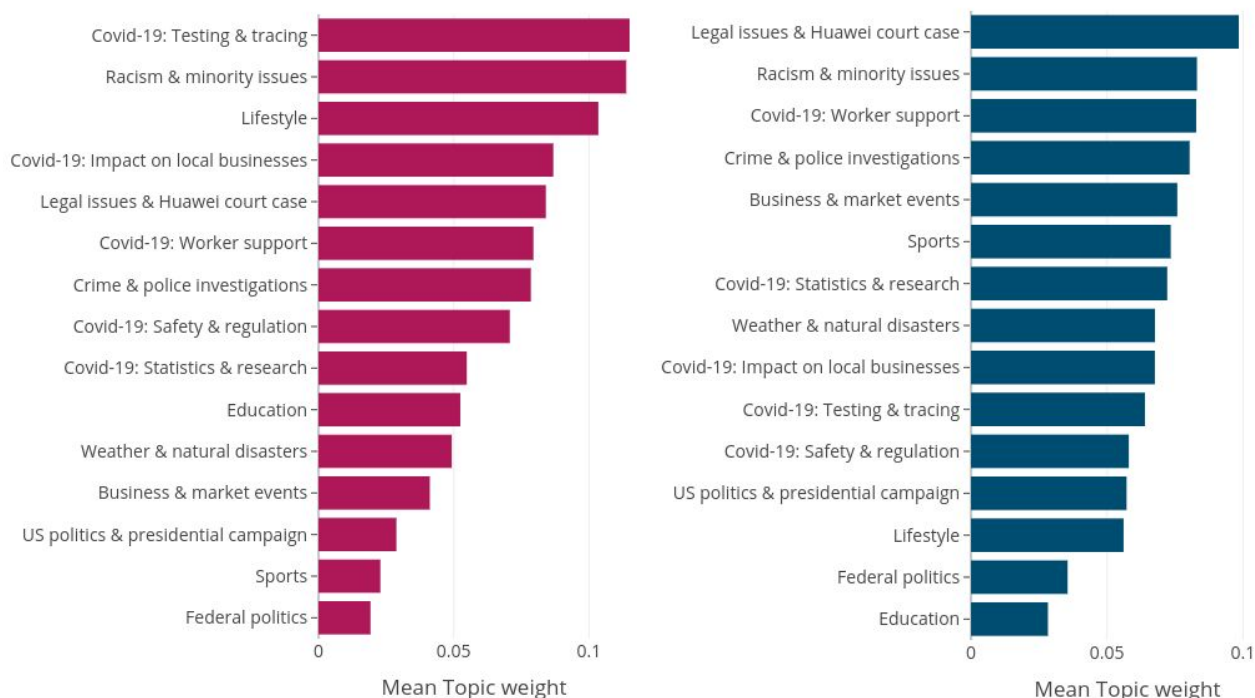


Figure 7. Bar plots of topic gender prominence for news articles over all outlets in July 2020

Interestingly, the variance between the top and middle topic weights is much greater for the female corpus than for the male corpus. We think this is primarily due to the fact that many, many more articles exist that quote more men than women¹⁵, so it is natural that men's voices are more equitably distributed across the topics. We observe a similar trend for almost all the months for which we have data, meaning that there could exist a correlation between the gender distribution of sources in an article and the likely content it covers.

8 Analysis and observations

In this section, we make some high-level observations on the average topic trends over time, and perform a deeper analysis on certain representative topics using techniques from corpus linguistics. We show that, by exploring the distribution of topics, proportions of those cited, and certain keywords in the data, we can gain important insights about gender representation in Canadian mainstream media.

8.1 Monthly gender prominence for nine recurring topics

First, we look at nine topics that feature quite regularly in most months between October 2018 and July 2020. The topics cover a broad range of domains typically seen in the news, and are shown on the vertical axis of Figure 8. From [our dashboard](#), we obtain per-month, per-outlet mean topic gender prominence measures for each of these topics over the entire time period we have results for. The values range from negative (red), indicating female prominence for that topic, to positive (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

¹⁵ In July 2020, there were 12,723 articles in the male corpus, and just 4,303 articles in the female corpus.

as a time-series heat map, as shown in Figure 8. The white (neutral) squares in the heat map indicate that a given topic did not appear at all in that particular month (or, in some cases, where perfect parity exists).

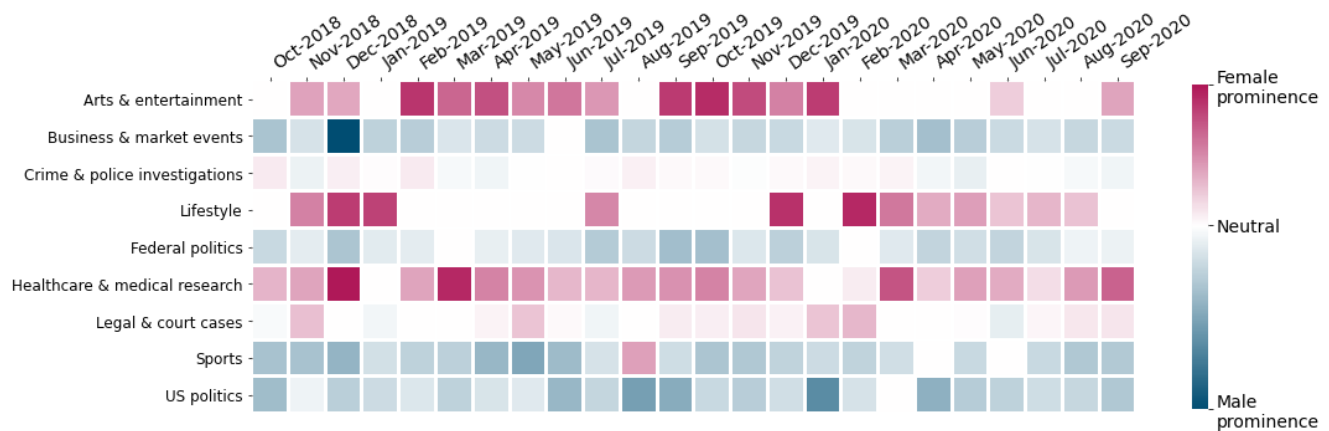


Figure 8. Monthly topic gender prominence for 9 recurring topics (avg. over all outlets)

We can see from Figure 8 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 & entertainment', which primarily discusses news about public events, art exhibitions, films, television, and celebrities, tends to show a high topic intensity in the female corpus. This same topic does not, on average, show as high an intensity in the male corpus (relative to all other topics in the male corpus). Because of this disparity, the 'Arts & entertainment' topic regularly shows a strong female prominence over all months for which it occurs. 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 art events were cancelled.

Other topics that show strong female prominence over time include 'Lifestyle' (which contains keywords related to families, home/holiday activities and individual/personal experiences), 'Healthcare & medical research' and 'Legal & court cases'. The primary reason for the 'Lifestyle' topic being female-prominent is that it regularly features mothers involved in childcare, women in the family reliving their past experiences, or women experts offering personal care advice. The 'Legal & court cases' topic tends to quote a good number of women sources with a relatively high topic intensity mainly due to cases that pertain to sexual assault or women's rights, featuring both victims and experts who are women. The 'Healthcare & medical research' topic is particularly interesting, because, at least in Canada, a sizable proportion of healthcare experts and provincial medical 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 a number of prominent expert women being quoted), thus obtaining a higher female prominence in most months.

On the other hand, the 'Business & market events', 'Sports', 'Federal politics' and 'US politics' topics are almost entirely male-prominent throughout the time period studied. The 'Business' topic typically includes keywords pertaining to investments, analysts' predictions on the stock market, and in-depth discussions on company futures, interest rates and profits. Historically, these areas tend to be

dominated by men, partly explaining why the topic intensity for 'Business' is always much lower for the female corpus. Similarly, the 'Sports', 'Federal politics' and 'US politics' topics feature discourse that generously quotes male sportspersons, coaches and politicians, who are not only more in number, but are also quoted much more often than women in the sector (as can be seen in [Figure 2](#)).

In general, the heat map shown in Figure 8 highlights aspects of societal bias that we know exist today. It seems to uphold our topic modelling methodology, as it confirms the intuitions that we had when we started this project. We explore these trends for certain representative months in more detail in [Section 8.3](#).

8.2 Monthly gender prominence per outlet

Next, we would like to see whether a gender is prominent in specific topics based on the outlet that produced the content. We first tabulate mean gender prominence results by outlet, as shown in Figure 9. For consistency, we consider just the nine recurring topics shown in Figure 8 for averaging.

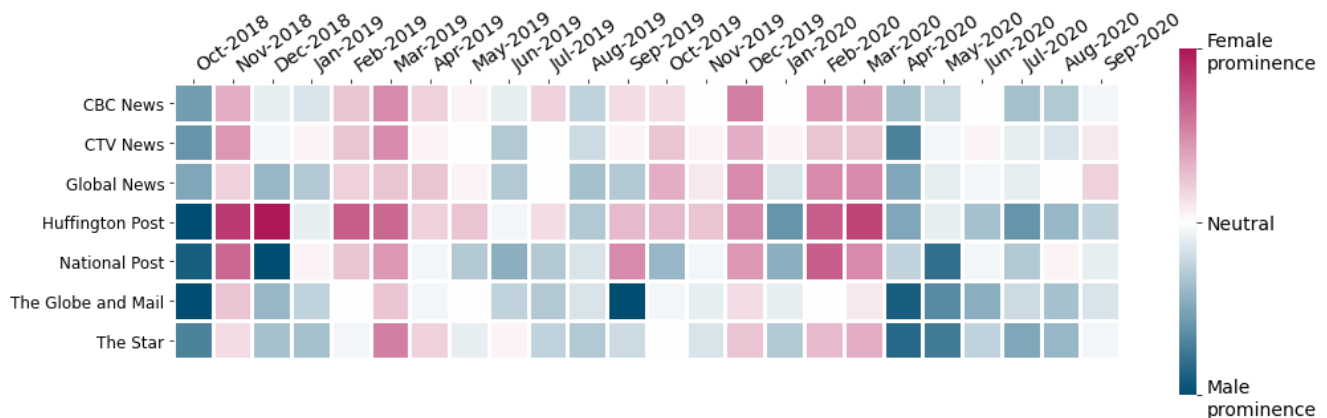


Figure 9. Monthly topic gender prominence per outlet (avg. over 9 topics)

The heatmap from Figure 9 seems to highlight a similar trend as observed in our main dashboard that tracks the proportion of female sources by outlet (across all topics). *Huffington Post*, does indeed, have a higher proportion of female sources (on average) than *The Globe and Mail* over 2 years of data, as can be seen in Figure 10.

The *Huffington Post* seems to have the most months with female-prominence (averaging over these nine topics), spanning 13 of the 24 months studied (Figures 9 and 11). *The Globe and Mail* (Figure 12) shows the least female prominence amongst its topics (just 4 out of the 24 months). Specific months (November 2018, February 2019/2020 and March 2019/2020) show recurring patterns where more female prominent topics are visible across all outlets. During the early months of the COVID-19 outbreak in Canada (February/March 2020), a significant proportion of all topics discovered featured a high level of female prominence, mainly due to the fact that many provincial medical officers and health experts (who were quoted quite heavily) happen to be women. From April onward, however, we see a shift back to male prominence on average, because of an uptick in political coverage in Canada and the U.S.

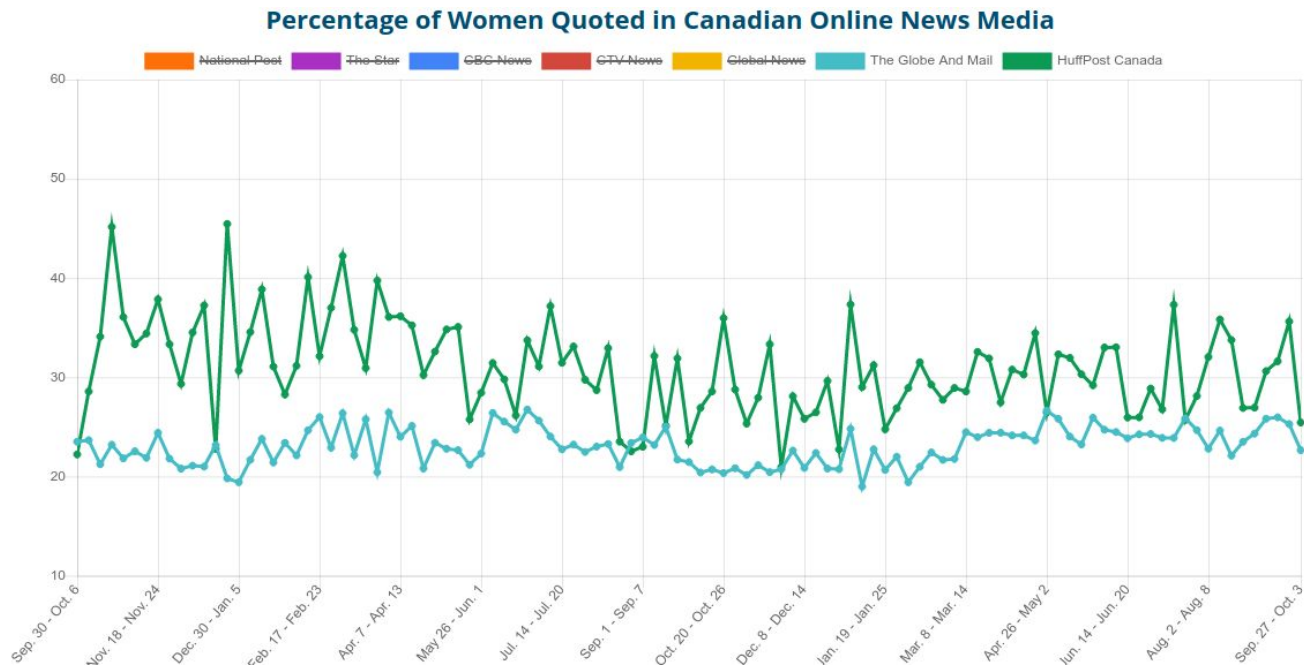


Figure 10. Snapshot of the daily proportion of female sources for two specific outlets (<https://gendergaptracker.informedopinions.org/>)

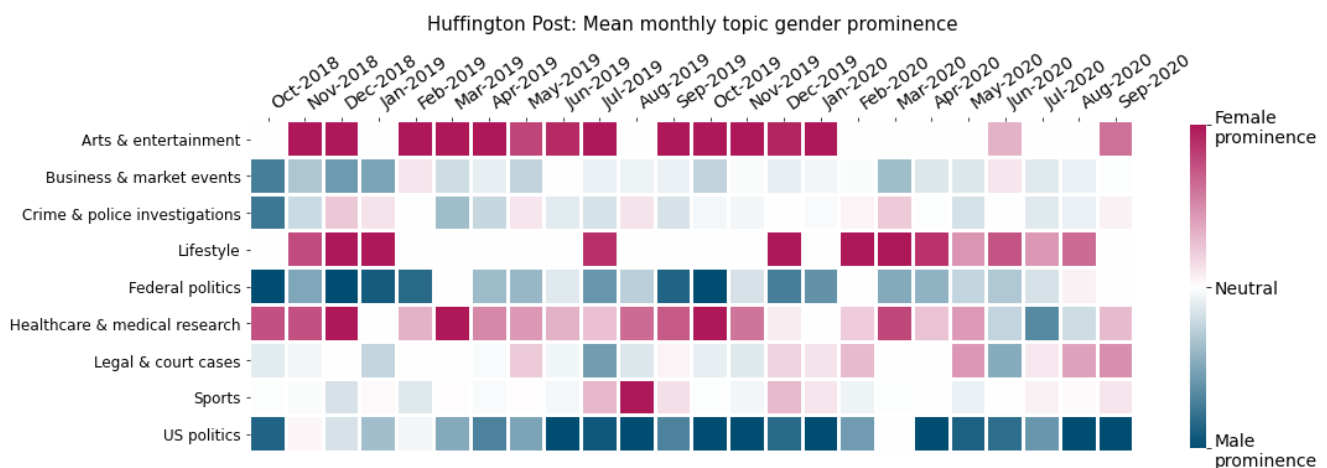


Figure 11. Monthly gender prominence per topic, *Huffington Post*

The Globe and Mail, historically known to cater to the “Canadian political and intellectual elite, such as managers and professionals” (Winter, 2011), focuses heavily on the ‘Business & market events’ topic throughout the year. This is also the topic for which *The Globe* exhibits the strongest male prominence, as can be seen in Figure 12. This is not to say, however, that *The Globe* fails to cover any topics whatsoever that feature female voices. Topics such as ‘Crime & police investigations’, ‘Legal & court cases’, as well as ‘Lifestyle’ and ‘Arts & entertainment’ regularly feature as female-prominent in *The Globe*’s content. It does seem, though, that covering the ‘Business’ topic predisposes authors to quote men much more heavily than women, leading to a strong male prominence for this topic throughout.

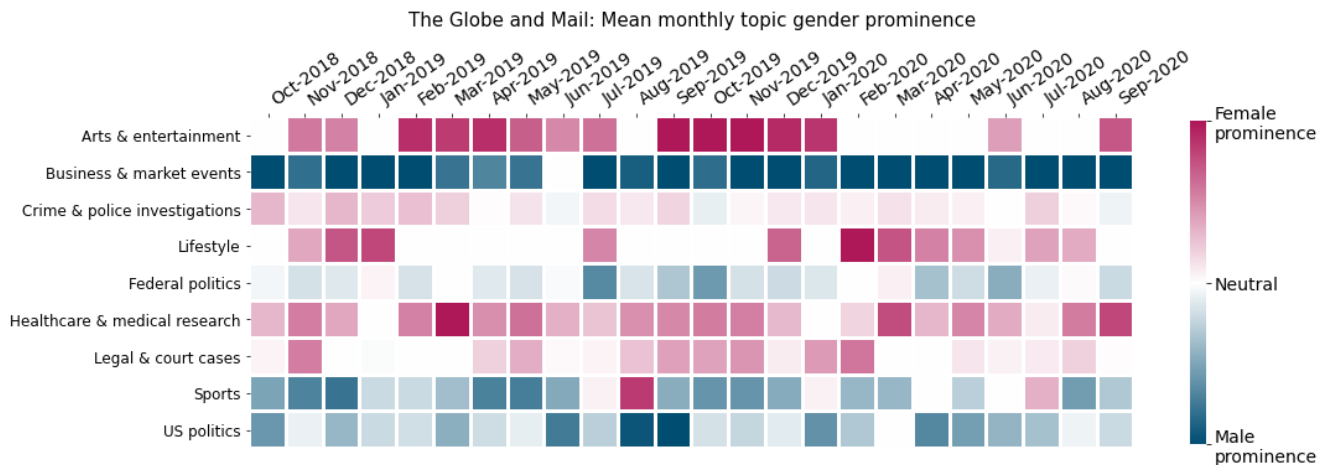


Figure 12. Monthly gender prominence per topic, *The Globe and Mail*

8.3 Corpus analysis

The heat maps shown in Sections 8.1 and 8.2 provide a high-level overview of which topics tend to show male or female prominence over time. However, to gain an understanding of *why* certain topics in certain months show particular gender distributions in their top quoted sources, a deeper linguistic analysis is required. Because we already divide our news article content into two separate corpora based on which gender is most quoted, our scenario is well-suited to *corpus studies*, i.e., a set of techniques that are known to “help deconstruct hidden meanings and the asymmetrical ways people are represented in the press” ([Caldas-Coulthard and Moon, 2010](#)).

In this section, we utilize two techniques from corpus linguistics (keyness and dependency bigrams) in conjunction with our topic modelling results to study asymmetry (if any) in the way the same topic is covered between the male and female corpora introduced in Section 7.2. To begin, we identify a topic of interest that exhibits strongly male or female gender prominence for a particular month. We then query that month’s data from our database and sort in descending order of topic weights for that topic (recall that we store the topic distribution vector for every article). Sorting the articles in this order puts all articles that are strongly related to that topic’s keywords on top. The full corpus, in sorted order for a particular topic, is then split into two corpora, each with male-majority and female-majority sources (just as we did earlier in Section 7). An illustration of this workflow is shown in Figure 13.

Once we have the two corpora, we then extract the full body article text (using the article IDs) for the top 200 articles in either corpus from our database. We chose 200 articles for empirical reasons—we observed that in most cases, the maximum topic weights for each article rapidly dropped to less than 0.5 after a few hundred samples (sorted in descending order of weights), so it didn't make sense to go too far down in the list of articles strongly associated with a particular topic.

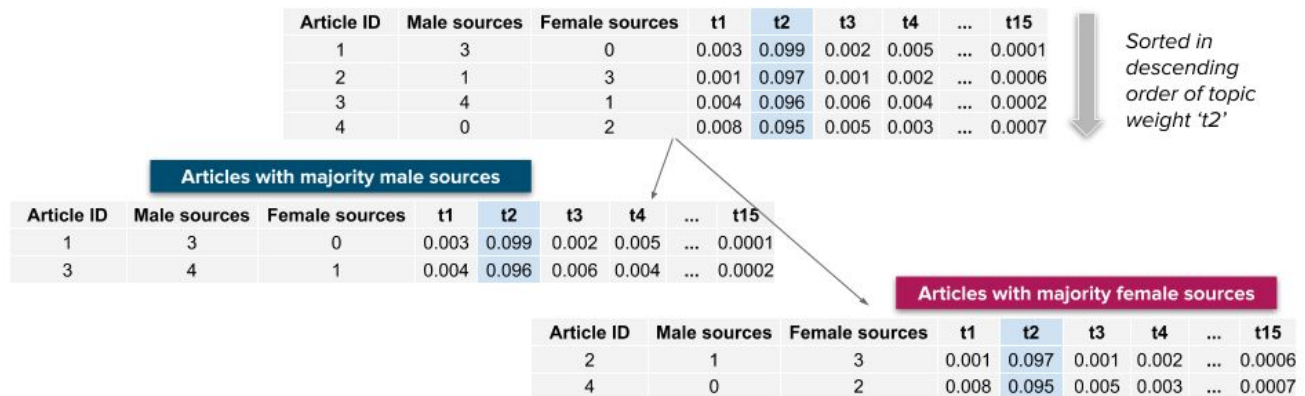


Figure 13. Topic-wise sorting and article extraction for corpus analysis

Following these steps, we use the 'corpus-toolkit' Python library (built on top of spaCy)¹⁶ to perform keyness analysis and extract dependency bigrams.

Keyness

Keyness is a concept from corpus linguistics that aims to identify large differences between the frequency of word-forms in two corpora (Gabrielatos, 2018). Keyness analysis introduces the term 'key word', which is a word that occurs in a text more often than we would expect to occur by chance alone. Historically, the *Chi-squared* metric was used to calculate the statistical significance, i.e., the extent to which we can trust that there exists an observed frequency difference of a key word's occurrence in either corpus, termed the 'keyness score'. However, recent work by Hardie (2014) has marked a shift in corpus linguistics towards the use of *effect-size* metrics for keyness. The 'log-ratio' metric is an example of an effect-size metric, as it can be directly used to interpret the magnitude of the frequency difference for a given set of key words between two corpora. As a result, in this section, we make use of the log-ratio keyness metric (as implemented in the *corpus-toolkit* library) for all our case studies.

Dependency bigrams

A relatively more advanced form of corpus analysis, dependency bigrams help identify verb-direct object combinations and their frequencies within a corpus (Mertz et al. 2014; Sidorov et al. 2012). To extract these bigrams, a dependency parser (such as the one provided by spaCy) identifies words that are syntactically connected by a head-dependent relationship¹⁷. For example, in the clause "The player **kicked** the **ball**", the main verb **kicked** is connected to the noun **ball** via a direct object relationship, wherein **kicked** is the head and **ball** is the dependent. The dependency bigram returned using this process is 'ball_kick'.

Studying dependency bigrams helps understand linguistic aspects of the male and female corpora, unlike keyness (which mainly highlights textual differences between the corpora). From our experiments with our data, we obtained much more useful results with dependency bigrams rather than the more standard corpus analysis methods such as collocation or n-gram frequencies. This is because the

¹⁶ https://kristopherkyle.github.io/corpus_toolkit/

¹⁷ https://github.com/kristopherkyle/corpus_toolkit#dependency-bigrams

dependency bigrams capture syntactic and semantic relationships between words, rather than just bags of words from a text (unigrams) or linear relationships ('classic' bigrams).

8.3.1 Sports

The first topic we study in greater detail is 'Sports', during the months of June, July and August 2019. August 2019 was an outlier because it was the only month for which the 'Sports' topic showed an overall female prominence. August 2019 was also particularly interesting because 'Sports' was the strongest topic in *both* the female and male corpora. To understand these trends a little better, we performed keyness and dependency bigram analyses for the top articles from each of these three months. The keyness results show *textual* differences between the female and corpora, whereas the dependency bigrams show *linguistic* differences based on subject-direct object relationships.

Figure 14 shows the results for June 2019—for this month, we obtained two topics for sports (one for baseball/football, and another for basketball due to the Toronto Raptors becoming NBA champions). As a result, we consider the top 200 articles from *both* sports topics for this month's analysis, giving us a total of 400 articles in either corpus for this month.

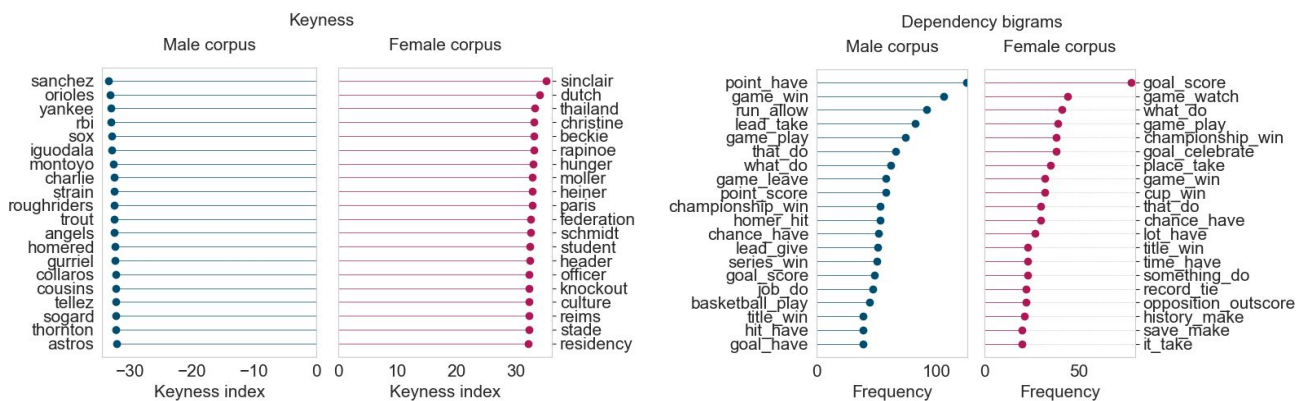


Figure 14. Corpus analysis results for top 400 articles from 'Sports' (June 2019)

From the keyness plot, we can see that the female corpus contains numerous terms pertaining to the FIFA women's world cup held in France that month. The names of Canadian and American soccer players and those of their female coaches appear, as do other terms related to women's soccer federations. The male corpus, on the other hand, shows terms largely related to baseball (*Yankees, Red Sox, Orioles*) and teams from the Canadian Football League (*Lions, Roughriders, Eskimos*). The dependency bigram plots show differences in language used across the female and male corpora based on frequency counts of the most common bigrams¹⁸. The male corpus shows much higher occurrences of general verbs describing events in baseball/football games, such as 'hit homer', 'score goal', 'allow run' and 'give lead'. The female corpus, on the other hand, seems much more focused on winning, celebrating and breaking records (although 'score goal' is the highest ranking bigram in terms of raw frequency).

¹⁸ The dependency bigrams in the plots should be read right-to-left, that is, "game_win" is part of a sentence involving "win a/the game", and "lead_take" refers to taking the lead in a game.

One interesting observation from the sports topics for all three months is the relative lack of basketball terms, including player and coach names in either set of top keywords for the month of June 2019. This is despite the fact that the Toronto Raptors became NBA champions that month, resulting in heavy basketball news coverage during this period. It does seem that the coverage involving the Raptors' success was more equitably distributed across both the male and female corpora, at least when compared to sports like baseball or women's soccer, explaining why these keywords do not appear.

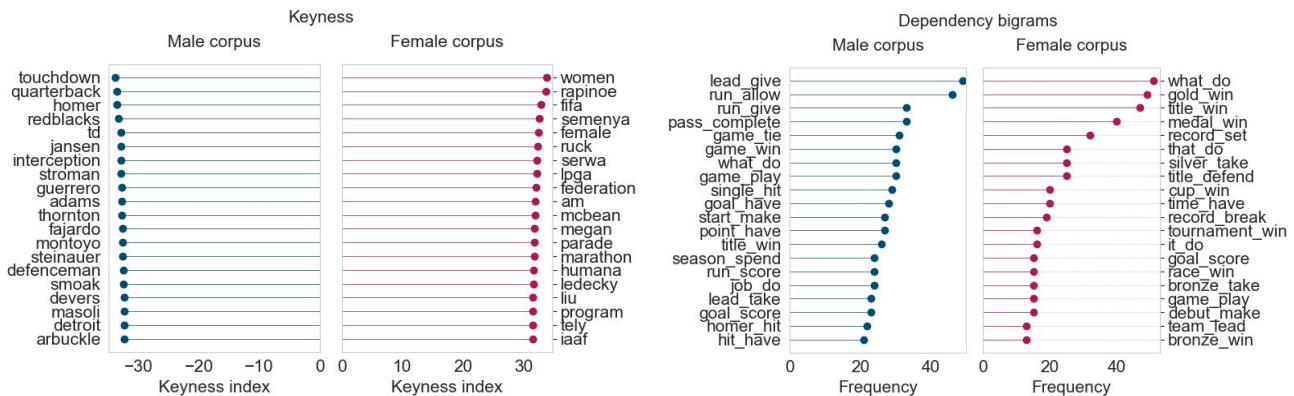


Figure 15. Corpus analysis results for top 200 articles from 'Sports' (July 2019)

July 2019 shows a similar trend, as can be seen in Figure 15. The female corpus continues to show terms related to the FIFA women's world cup, with some added key words and person names related to summer athletics and golf. The male corpus continues to highlight terms and player/coach names from baseball (*Blue Jays*) and the CFL (*Redblacks*). The dependency bigram plots also show a similar trend, with the female corpus using more vocabulary related to winning cups/medals and setting/breaking records. On the other hand, the male corpus much more frequently highlights terms that describe finer aspects of games, such as giving the opponents the lead, or allowing them a run.

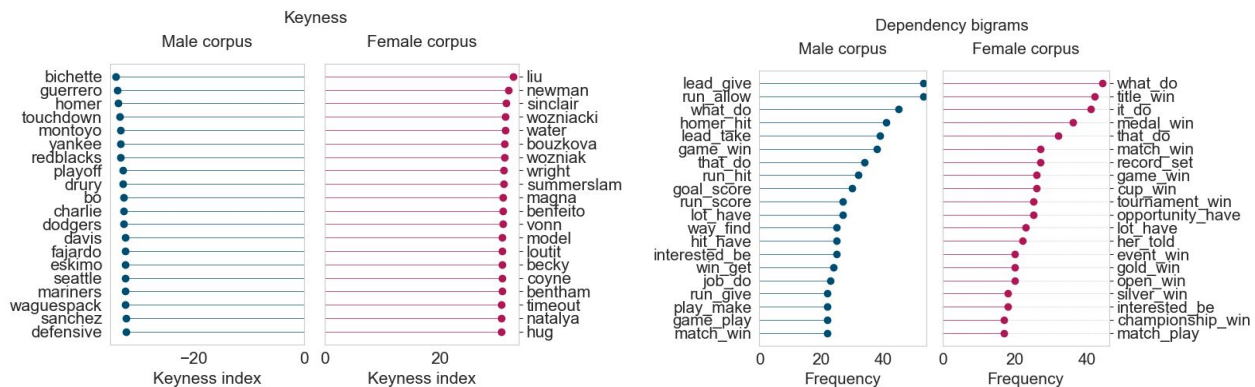


Figure 16. Corpus analysis results for top 200 articles from 'Sports' (August 2019)

Finally, in August 2019, the keywords show that the female corpus largely switches focus to the Wimbledon tennis tournament and summer athletics. The male corpus continues to focus heavily on baseball and the CFL season and their personnel. Just as in the prior two months, the female corpus once again shows a high frequency of bigrams related to winning matches, titles or cups, and setting records.

From our analysis of the top 200 'Sports' articles across three summer months in 2019, we observed that articles that feature women's voices tend to focus more heavily on achievements and records ('win cup' and 'break record') as opposed to articles that feature men's sports, which focus a lot more on the finer details of each game. Perhaps authors that write about women's sport tend to highlight women's achievements and success stories because of the limited time/space they have compared to authors that cover men's sports. After all, the baseball and football seasons in Canada are each several months long, whereas events like the FIFA world cup and the Wimbledon tournament occur over the span of just one month. It is important to remember that the male corpus also contains roughly 3 times as many articles as the female corpus, so both these factors could also play a role in the kind of language used by authors. It seems that men just need to play to feature in the news; women need to win.

However, as August 2019 shows, sports doesn't always have to be a male-dominated arena in the news. In sports that have well-known female role models (Bianca Andreescu for tennis, Christine Sinclair and Megan Rapinoe for soccer and Michelle Liu for golf), we see considerable amounts of coverage in the female corpus during certain times of the year. However, sustaining a high level of coverage seems to require more women role-models across a wider range of sports, as well as longer seasons over which women's sports are regularly featured. Since the COVID-19 pandemic in early 2020, our topic model gender prominence results show that women's sport seems to have once again taken the back seat in the news.

8.3.2 Business & market events

We consider two months, October and December 2018, for which this topic exhibits a particularly strong male prominence. Just as before, the top 200 articles (with strongest topic intensities) were chosen for either corpus for comparison.

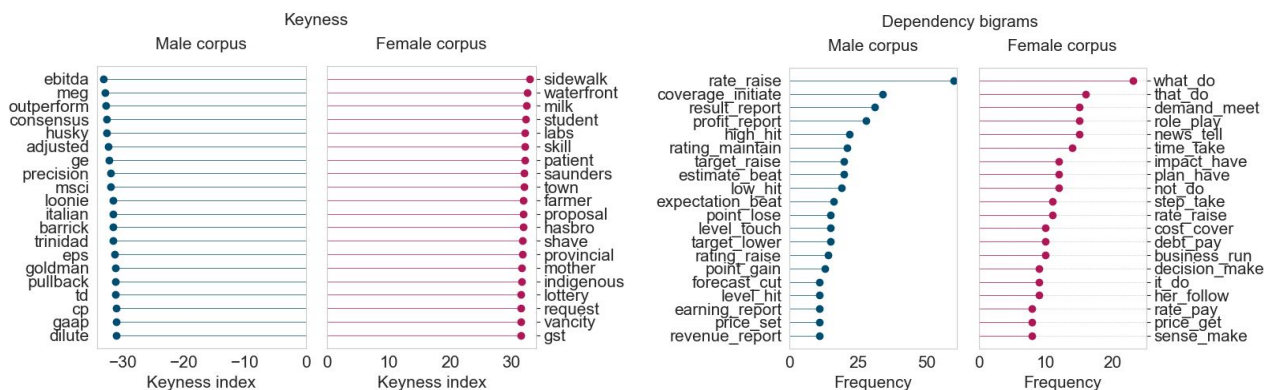


Figure 17. Corpus analysis results for top 200 articles from 'Business' (October 2018)

Figure 17 shows the results for October 2018. The keyness plot highlights terms related to small-time transactions and shopping ('milk', 'shave') and small/local businesses ('town', 'mother' and 'baby') for the female corpus. This can be further clarified by inspecting the dependency bigrams, where we see high frequencies of terms like 'have impact', 'cover cost', 'pay debt' and 'run business'. The male corpus, on the other hand, shows terms related to larger financial instruments, stocks, trading and company performance metrics. The male corpus also frequently features verbs pertaining to high-level business

policy and decision-making, as can be seen from the 'raise rate', 'maintain rating' and 'report profit' bigrams.

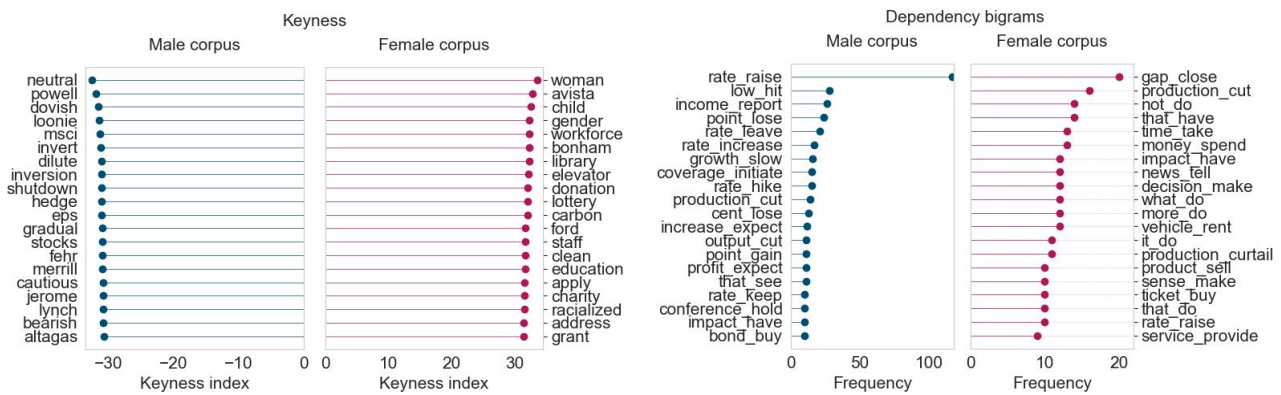


Figure 18. Corpus analysis results for top 200 articles from 'Business' (December 2018)

December 2018, the month for which the 'Business' topic exhibits the strongest male prominence in all our results, showcases a similar disparity in the kind of language used across both corpora. Keyness analysis shows that the female corpus uses keywords pertaining to women and minorities in the workplace, along the lines of 'gender', 'workforce', 'child' and 'racialized'. The male corpus, as always, contains specific terms related to financial and business operations ('stock', 'hedge', 'dilute', 'bearish'). From the dependency bigrams, we can see that the female corpus for this month highlights terms such as 'cut production' and 'make decision'—this was primarily due to the then-Alberta Premier Rachel Notley's numerous comments on the oil industry's production cuts. However, other terms such as 'close gap', 'spend money' and 'sell product' clearly point to a certain level of gender disparity between the two corpora.

We find that the 'Business and market events' topic regularly showcases the largest disparity between the male and female corpora, in terms of language used, for all months of data studied. It seems that women's voices are rarely featured in articles that cover business, especially big business, finance and the stock market. The kind of language used (as seen in the dependency bigrams) clearly shows that the focus of articles that quote women tends to be primarily on small/local businesses and the challenges of working at or running a business. It is also important to note that a large portion of articles in this topic were published by *The Globe and Mail*, which we know from our database statistics (Figure 10), quotes the smallest proportion of women (on average).

8.3.3 Lifestyle

The 'Lifestyle' topic, which covers content related to families, personal experiences, holidays and shopping, showcases strong female prominence for all months in which it appears. Figure 19 and Figure 20 show the results from two months (December 2019 and February 2020), for which we observed particularly strong female prominence in this topic.

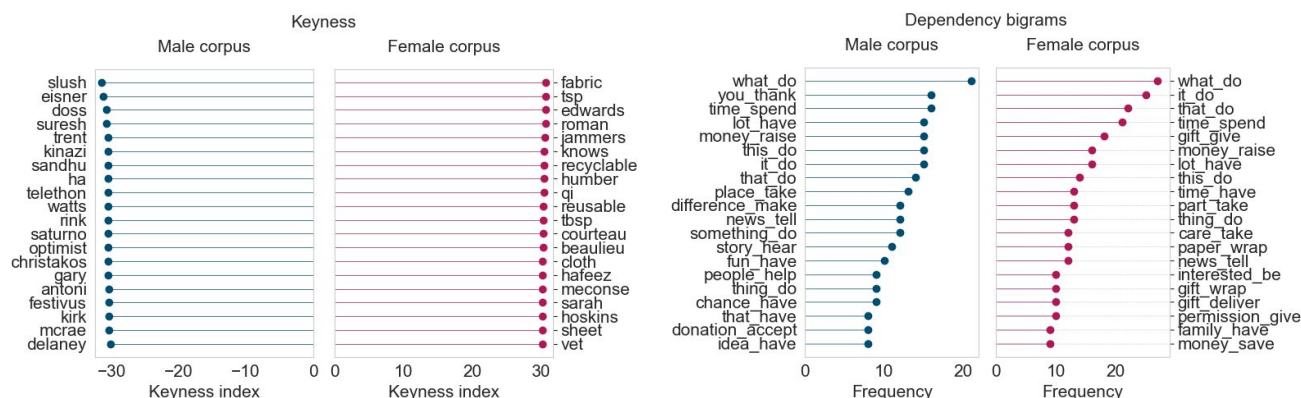


Figure 19. Corpus analysis results for top 200 articles from 'Lifestyle' (December 2019)

Keywords from the female corpus for December 2019 highlight a range of lifestyle topics, from food recipes ('tbsp', 'tsp') to clothing and fabrics. Interestingly, not as many women's names are seen for this month's keywords, as we do for other topics in other months. The male corpus keywords include several names of people, including Suresh Doss, a Toronto-based food writer. The dependency bigrams for both the male and female corpora focus mainly on the festive season, as well as acts of gift-giving and donating to charities.

In February 2020, an interesting new term appeared in the keywords for the female corpus—'parentese'. This is a word related to 'motherese' (a speaking style used by mothers of very young children). Although the term 'parentese' was introduced as a gender-neutral term, it still tends to be used in articles in which expert women that studied the phenomenon are quoted, making it particularly strong in the female corpus. Once again, keywords related to women's clothing and fashion appear in the 'Lifestyle' topic. In contrast, the male corpus contains keywords related to music, musicians and a number of prominent men's names.

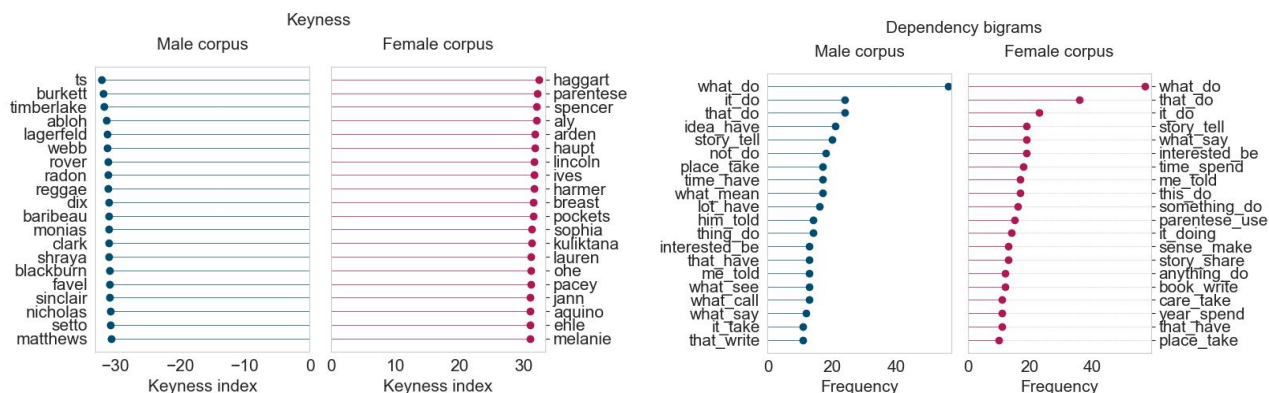


Figure 20. Corpus analysis results for top 200 articles from 'Lifestyle' (February 2020)

We found through our exploration that the 'Lifestyle' topic, unlike other topics, tends to name a lot of regular people who share their life experiences with the media—this can lead to the drowning out of other keywords such as common nouns or adjectives in the keyness results. As a result, keyness is slightly less useful in interpreting textual disparities across the corpora for this topic. On occasion,

interesting terms (such as 'parentese') emerge, mainly because multiple outlets tend to write about the same topic, typically quoting the same people. This points to a trend where authors from different outlets seem to be influenced by other authors writing about a hot/trending term. Some interesting questions also arise through this analysis: Do specific keywords emerge in the female corpus primarily because they are talked about by expert women? And if so, do expert women tend to talk more often about topics that are relevant to women?

8.3.4 Crimes & sexual assault

In exploring our results, we observed that topics pertaining to women's rights and sexual assault emerged around February and March, in both 2019 and 2020. We explore these topics for these periods using corpus studies, as shown below.

Figure 21 shows the results for February 2020. A number of terms pertaining to women's and trans rights appear in the female corpus ('hijab', 'trans', 'hormone'). The issue of equality within women's sports was also a major one during this period, as can be seen in the 'fifa' and 'cup' keywords. The male corpus keywords primarily feature names, presumably of men involved in crimes and their victims. Some key dependency bigrams from the female corpus include 'protect athlete', 'involve victim' and 'take action', indicating that coverage was focused on the action taken towards women's justice. The male corpus highlights bigrams related to the crimes themselves, convictions and legal matters.

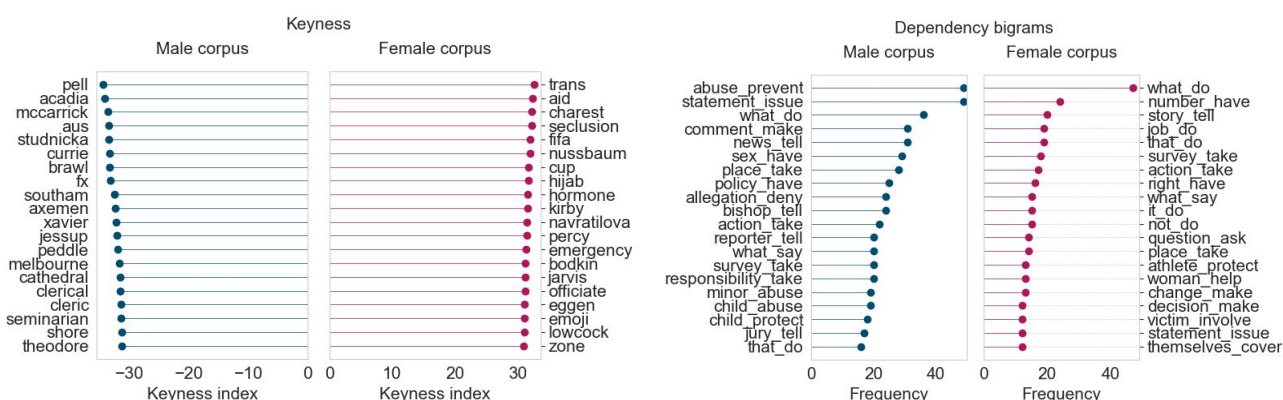


Figure 21. Corpus analysis results for top 200 articles from 'Crimes & sexual assault' (February 2019)

In February 2020, as seen in Figure 22, a number of keywords in the female corpus relate to the Harvey Weinstein sexual crime convictions, as well as other crimes against women. A number of victims' names appear in the keywords ('mcgowan', 'rosenbaum', 'cunningham'). The male corpus keywords once again consist of names of perpetrators or victims of other crimes. In general, the language used (as seen in the bigrams) tends to focus on the victim's accusations and details of the crimes, in the female corpus, and on the convictions, defense and accounts of the victims in the male corpus.

It is interesting that topics around women's rights (encompassing themes from sexual assault, consent, #MeToo, and gender equality) tend to emerge at the same time each year (February/March)—which is also the time when International Women's Day¹⁹ is celebrated. Analyzing the language used for these

¹⁹ <https://www.unwomen.org/en/news/in-focus/international-womens-day>

months, we find it heartening that a number of strong action verbs ('take stand', 'tell story' and 'have right') are used quite frequently in the female corpora, highlighting the important role the media plays in spreading awareness about women's issues and rights in Canada.

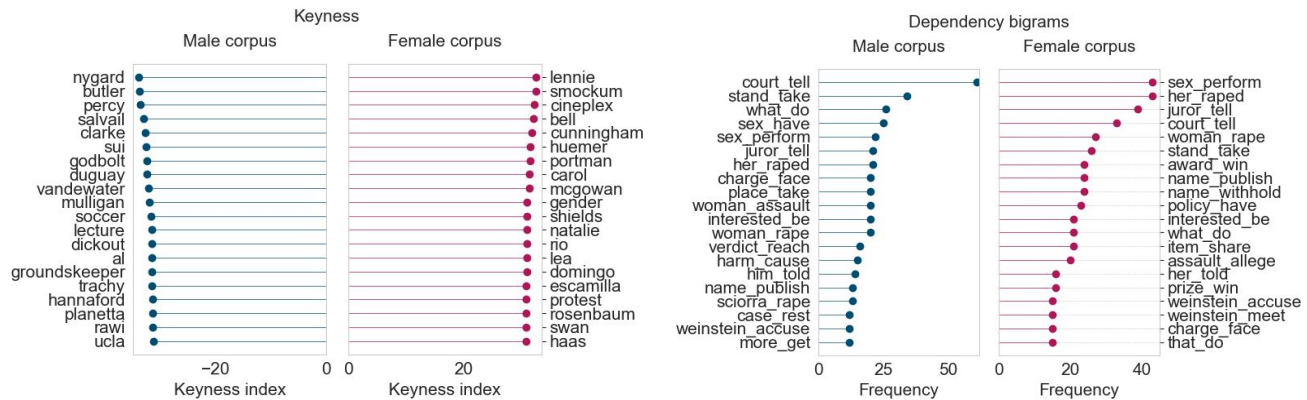


Figure 22. Corpus analysis results for top 200 articles from 'Legal & sexual crime cases' (February 2020)

8.3.5 Healthcare & medical research

In the Gender Gap Tracker dashboard²⁰, we observed during the period February-May 2020 (at the peak of the COVID-19 global pandemic), that the proportion of women quoted across all outlets increased by 3-4%, compared to pre-COVID levels. To study this in greater detail, we first showcase the keyness and bigram analysis results for March 2019 (a year before the COVID-19 pandemic), followed by a similar analysis for March 2020, when COVID-19 began spreading in North America.

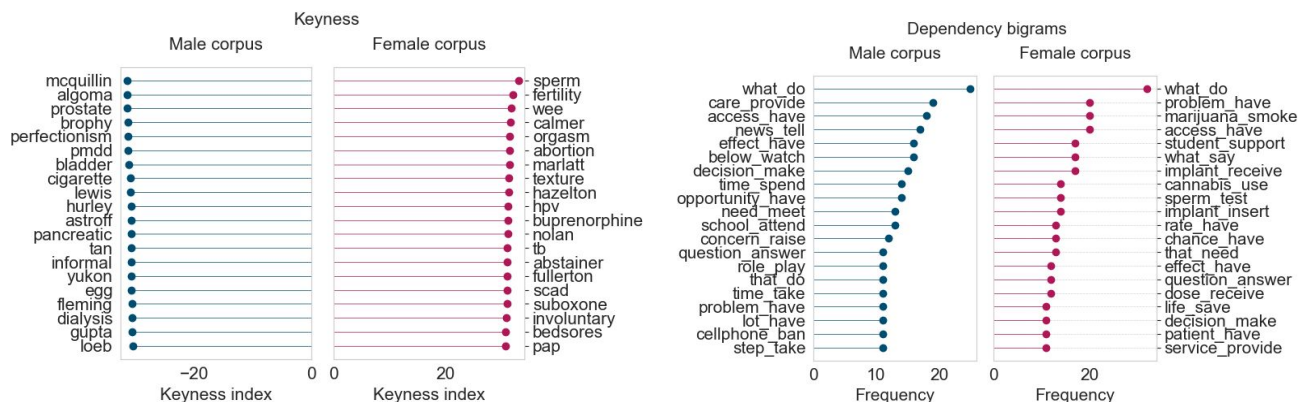


Figure 23. Results for top 200 articles from 'Healthcare & medical research' (March 2019)

Figure 23 shows the results for March 2019. The female corpus shows a number of terms related to women's fertility and their sexual/overall health ('sperm', 'hvp', 'abortion'), while the male corpus shows a similar focus toward terms pertaining to men's health ('prostate', 'bladder', 'pancreatic'). Interestingly, the bigrams from the female corpus seem to showcase an additional emphasis on the effects of women's personal choices on fertility and pregnancy ('smoke marijuana', 'use cannabis' and 'insert implant'). The

²⁰ <https://gendergaptracker.informedopinions.org/>

male corpus bigrams, in this case, do not seem to have a particular focus other than the availability and effects of various treatments in the healthcare domain.

Figure 24 highlights the stark shift in healthcare coverage once the COVID-19 pandemic began spreading in Canada in March 2020. Because of the seriousness of the pandemic, there were multiple COVID-related topics discovered in March—we focus our attention on **four** specific topics that were clearly female dominant throughout this period ('COVID-19: Testing & tracing', 'COVID-19: Cases, deaths & spread', 'COVID-19: Community impact & closures' and 'COVID-19: Provincial updates & case counts'). The top 100 articles' full body text for each topic (ordered by topic intensity) were collected from our database and separated into male and female corpora, just as before. This gave us **400** articles in either corpus to analyze for their language used during this month.

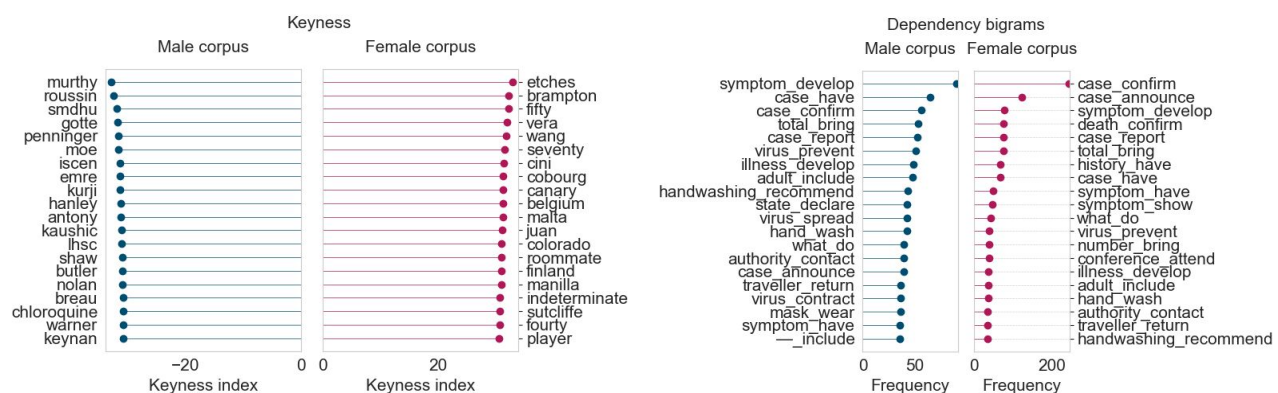


Figure 24. Results for top 400 articles from four 'Healthcare & COVID-19' topics (March 2020)

A number of the keywords in the female corpus are direct names of women who are medical experts (Dr. Vera Etches and Dr. Penny Sutcliffe), or are related to the spread of COVID-19 in various regions of Canada and the world. Similarly, the male corpus also names prominent medical researchers and scientists (Dr. Srinivas Murthy and Dr. Brent Roussin). Unlike the months prior to the pandemic, the dependency bigrams showcase a very similar distribution across both male and female corpora, with terms pertaining to symptoms, case counts and public health guidelines. Importantly, the frequency count of the bigram 'confirm case' is significantly higher (> 200) in the female corpus than in the male corpus, which makes sense, considering that for the most part, the provincial health officers confirming the numbers happened to be women.

In general, we observe that Canadian media outlets tend to provide a good degree of coverage to women's healthcare and health issues throughout the year. This is helped by the fact that a relatively large proportion of healthcare experts and medical officers in Canada are women, which explains the high female prominence of the topic in most months of the year. The most notable of these is Bonnie Henry, the Chief Provincial Health Officer of British Columbia, who is the top quoted woman by far throughout the spring/summer of 2020, due to the seriousness of the COVID-19 pandemic). This clearly shows that having women in prominent positions increases their chances of being quoted by the media. This is, naturally, beyond the control of news organizations and journalists.

9 Conclusions

Our goal through this topic modelling study was to probe deeper into the numbers from the Gender Gap Tracker's data, and to understand the relationship between topics in the news and the gender of those quoted. We deployed our topic modelling workflow as a monthly automated (unsupervised) process, allowing a human to come in and label the topics after the fact, using guidelines we developed specifically to handle our news corpus. Although we do trade off a certain degree of reproducibility for scalability (in order to handle huge datasets), we observed that our methodology still produces topic word distributions that are human interpretable, and repeatable (for most major topics discovered). We tested this approach over multiple models from nearly two years' worth of data to produce visualizations that reflect events and trends from the real world.

We define a term, 'gender prominence' that measures the difference in topic weight intensities between two corpora (one that contains articles quoting a majority of women, and another containing articles quoting a majority of men) to better understand which topics feature women's voices more prominently, on average. Our findings indicate that, although women are always quoted much less frequently than men (in overall numbers) regardless of topic, certain topics, such as arts, lifestyle and healthcare, tend to feature women more prominently than they do men, perhaps confining women to a few areas of expertise in the public domain.

We also perform corpus-based language analyses based on the dominant topic, and investigate whether articles that quote more men than women tend to more frequently use different action verbs in specific contexts. We find that, on average, business articles show the largest disparity in the kind of language used when more men are quoted than women—articles that quote men heavily tend to highlight aspects of big-business and the stock market, whereas articles that quote more women than men tend to focus more on small/local businesses and shopping transactions. On the other hand, women appear to be relatively better represented in healthcare topics (at least in Canada). From our language analysis, healthcare topics regularly feature content relevant to women, and, on average, quote a sizable number of expert women—this effect is even more pronounced since COVID-19.

Gender equality is one of the United Nations' 17 Sustainable Development Goals ([United Nations, 2019](#)). We are sadly far from achieving gender equality in many areas of our societies. Gender representation in the media is, however, within our reach, if enough effort is devoted to this goal and if we incorporate accountability into the effort. Our hope with the Gender Gap Tracker's topic dashboard is that it be used as an accountability tool to encourage and facilitate gender parity in sources. Providing organizations with the means to narrow down on which topics exhibit the strongest disparity in gender focus can, in our view, have a tangible impact on the gender balance of sources by topic.

Due to the size and richness of data in the Gender Gap Tracker, we believe that a deeper analysis of gendered language by topic can yield further interesting insights. For instance, we are interested in studying whether quotes by men and women are presented differently in terms of endorsement or distance (*stated* vs. *claimed*) or in the way that the speaker is presented, as an expert or merely a source. While we plan to pursue such additional areas of inquiry, we also invite researchers to join in this effort. The data collected for this project can be made available, upon request, for non-commercial research purposes.

10 References

- Asmussen, C.B., Møller, C. (2019) 'Smart literature review: a practical topic modelling approach to exploratory literature review', *J Big Data* 6, 93 (2019). <https://doi.org/10.1186/s40537-019-0255-7>
- Balakrishnan, V. and Lloyd-Yemoh, E. (2014). 'Stemming and lemmatization: A comparison of retrieval performances', *Lecture Notes on Software Engineering*, 2(3):262 – 267.
- Blei, D., Ng, A. and Jordan, M. (2003) 'Latent Dirichlet Allocation', *Journal of machine learning research: JMLR*, 3, pp. 993–1022.
- Blei, D. M. (2012) 'Probabilistic topic models', *Communications of the ACM*, pp. 77–84. doi: 10.1145/2133806.2133826.
- Butt, M. (2003) 'The light verb jungle: still hacking away', *Complex Predicates*, pp. 48–78. doi: 10.1017/cbo9780511712234.004.
- Caldas-Coulthard, C.R. and Moon, R. (2010). "'Curvy, hunky, kinky": using corpora as tools for critical analysis'. *Discourse and Society* 21 (2), pp. 99–133.
- Dollinger, S. (2019). *Creating Canadian English: The Professor, the Mountaineer, and a National Variety of English*. Cambridge: Cambridge University Press.
- Gabrielatos, C. (2018). Keyness analysis: Nature, metrics and techniques.
- Hardie, A. (2014) 'Log ratio – an informal introduction'. Post on the website of the ESRC Centre for Corpus Approaches to Social Science CASS. Retrieved from <http://cass.lancs.ac.uk/?p=1133>.
- Hoffman, M., Blei, D. and Bach, F. (2010) 'Online Learning for Latent Dirichlet Allocation', *Advances in neural information processing systems*, 1, pp. 856–864.
- Jespersen, Otto (1965) *A Modern English Grammar on Historical Principles*. George Allen and Unwin Ltd., p. 117.
- MacKay, D. J. C., & Peto, L. C. B. (1995) 'A hierarchical Dirichlet language model', *Natural Language Engineering*, 1, pp. 289–307.
- Maier, D. et al. (2018) 'Applying LDA Topic Modeling in Communication Research: Toward a Valid and Reliable Methodology', *Communication Methods and Measures*, pp. 93–118. doi: 10.1080/19312458.2018.1430754.
- Manning, C. D., Raghavan, P., & Schütze, H. (2009) 'An introduction to information retrieval', Cambridge, UK: Cambridge University Press.
- Manning, C. D., & Schütze, H. (2003) 'Foundations of statistical natural language processing' (6. print with corr.). Cambridge, MA: MIT Press.
- Martin, F. and Johnson, M. (2015) 'More Efficient Topic Modelling Through a Noun Only Approach', *ACL Anthology*, Proceedings of the Australasian Language Technology Association Workshop 2015. Available at: <https://www.aclweb.org/anthology/U15-1013>.
- Mertz, M., Korfiatis, N., & Zicari, R. V. (2014). Using Dependency Bigrams and Discourse Connectives for Predicting the Helpfulness of Online Reviews. In M. Hepp & Y. Hoffner (Eds.), *Proceedings of the 15th*

International Conference on E-Commerce and Web Technologies (pp. 146-152). Cham: Springer.

Mimno, D., Wallach, H., Talley, E., Leenders, M. and McCallum, A. (2011) 'Optimizing semantic coherence in topic models', in *Proceedings of the 2011 Conference on Empirical Methods in Natural Language Processing. Association of Computational Linguistics.*, pp. 262–272.

Puschmann, C. and Scheffler, T. (2016) 'Topic Modeling for Media and Communication Research: A Short Primer', *SSRN Electronic Journal*. doi: 10.2139/ssrn.2836478.

Sidorov, G., Velasquez, F., Stamatatos, E., Gelbukh, A., & Chanona-Hernández, L. (2013). Syntactic Dependency-Based N-grams as Classification Features. In I. Batyrshin & M. G. Mendoza (Eds.), (pp. 1-11). Berlin: Springer.

United Nations (2019) 'The Sustainable Development Goals Report'. Report, United Nations. URL <https://unstats.un.org/sdgs/report/2019/>

Wallach, H. (2006) 'Topic modeling: Beyond bag-of-words', In *Proceedings of the 23rd International Conference on Machine Learning, January 2006*. doi: 10.1145/1143844.1143967

Wang, X., McCallum, A. and Wei, X. (2007) 'Topical n-grams: Phrase and topic discovery, with an application to information retrieval'. In *Proceedings of the Seventh IEEE International Conference on Data Mining (ICDM 2007)*, Omaha, USA, pp. 697–702.

Winter, E. (2011) 'Us, Them and Others: Pluralism and National Identities in Diverse Societies', *University of Toronto Press*, p. 96.

Zhao, W., Chen, J. J., Perkins, R., Liu, Z., Ge, W., Ding, Y., and Zou, W. (2015). 'A heuristic approach to determine an appropriate number of topics in topic modeling,' *BMC Bioinformatics*, 16(13), S8.