Topic Modeling for Gender Bias on Regional News Corpus

Norah Almousa
University of Pittsburgh
nia135@pitt.edu

Modhumonty Das
University of Pittsburgh
mod53@pitt.edu

Gina Ribniscky
University of Pittsburgh
gir20@pitt.edu

1 Introduction

Gender bias is a pervasive societal issue that extends its influence into various aspects of our lives, including the realm of information dissemination. In the context of news articles, gender bias can manifest in subtle yet impactful ways, influencing the language used, perspectives conveyed, and the overall framing of news stories.

Understanding and addressing gender bias is crucial for fostering inclusivity and promoting fair representation within the media landscape. The impact of biased language and narratives can perpetuate stereotypes, reinforce societal norms, and contribute to unequal power dynamics.

In this project, we aim to highlight on gender bias within news articles from Pakistan and Malaysia by employing topic modeling techniques. Our goal is to explore identifying bias in gender representations in different topics in country-specific news. Our analysis of gender bias is structured into two stages. In the first stage, we delve into an examination of the prevalence of each gender within individual news articles across various topics. The second stage involves identifying how much men and women are presented in news articles about different topics.

2 Literature Review

This project is largely inspired by the work done in (Rao and Taboata, 2021). In this article, gender bias in Canadian news media is explored using an LDA topic model. Once split into topics, two metrics are used to identify the representation of men and women in the corpus: number of men/women referenced and number of men/women quoted. Analysis is done on each topic using these bias metrics to show how men and women are represented. Their findings show that women are referenced in lower numbers than men, and are usually included in topics that align with traditional gender roles, like "Lifestyle", "Art and Entertainment", and "Healthcare". Conversely, men take the lead for most associations with "Politics", "Business", and "Sports".

As well as an analysis, these experiments also discuss pre-processing steps in great detail. Steps include tokenization, normalization, lowercasing, stop word removal, lemmatization, and relative pruning. Most interesting of these is the relative pruning, which removes extremely common words that aren't included in the list of stop words, as well as extremely rare words. This pruning method removes the most common words relative to all documents and rarest words relative to all documents, taking care to leave nouns as these are the most important parts of speech for topic modelling, especially when analyzing gender bias.

2.1 Topic Modelling

Topic modeling is an unsupervised learning approach in natural language processing used to group textual documents based on their hidden semantic structure. This field has been under study for many years, resulting in the development of various methods. In recent years, the application of topic modeling has gained significant traction across diverse fields, including health, hospitality, education, social networks (Kretinin and Nguyen, 2022), and finance (Chen et al., 2023). This surge in popularity is driven by its ability to extract valuable insights from large volumes of text data, enabling researchers, businesses, and organizations to better understand trends, patterns, and hidden knowledge within their respective domains. An effective utilization of the topic model was showcased in (Weiss and Muegge, 2019). In this context, the LDA model was employed to perform a comprehensive literature review, encompassing research papers sourced from prominent online databases like Web of Science, Scopus, or Google Scholar. Subsequently, the generated LDA model topics were aggregated into cohesive clusters, thereby de-

riving overarching themes from the original topics. This process proved instrumental in unraveling the interconnections between these topics, facilitating the creation of a concept map. By mapping out the keywords associated with these models, a precise and detailed depiction of the model's underlying topics was achieved.

In terms of the underlying algorithms, topic modeling techniques can be broadly classified into three primary categories: algebraic, probabilistic, and neural models(Chen et al., 2023). Models such as latent semantic indexing (LSI) and non-negative matrix factorization (NMF), fall under the algebraic category. Probabilistic models encompass methods like probabilistic latent semantic indexing (pLSI), anchored correlation explanation (CorEx), latent Dirichlet allocation (LDA), and various LDA extensions like hierarchical Dirichlet process (HDP), correlated topic model (CTM), and structural topic model (STM). STM, in particular, caters to social science research, allowing the inclusion of metadata and revealing diverse perspectives on the same topic in texts. Among these conventional models, LDA has been the most widely used over the years due to its popularity(Chen et al., 2023). However, researchers often default to LDA without providing a rationale for their choice. LDA can be seen as an extension of pLSI, incorporating a Dirichlet prior distribution for document-topic and topic-word distributions. One limitation of LDA is its reliance on the bag-of-words (BoW) representation, which overlooks the semantics between words in the text. Conventional topic modeling techniques like LDA require complex corpus pre-processing, careful parameter selection (e.g., the number of topics), proper model evaluation, and interpretation of generated topics based on domain knowledge. In recent years, a new wave of neural models have emerged and gained popularity, particularly since 2016. Examples of neural models include lda2vec, SBM, deepLDA, Top2Vec (Angelov, 2020), and BERTopic (Grootendorst, 2022). These development aligns with the rapid advancement of deep learning technologies. For instance, deep LDA combines LDA with a basic multilayer perceptron (MLP) neural network, while more recent models like BERTopic leverage advanced bidirectional encoders to capture richer contextual information.

2.2 Gender Bias

Gender bias in news media is a persistent issue that impacts the representation of both men and women and their roles in any society or region. Thus, identifying the gender bias in news and understanding the consequences of it is essential for promoting a balanced representation of all genders in the media and society as a whole.

The Canal Sur Noticias' (Spanish television network) (Muñoz Muñoz and Salido Fernández, 2023) Twitter profile highlights the presence of gender bias in public media, even in the digital realm. Findings of the study reveal a significant underrepresentation of women, not only as protagonists but also as authors and external sources. Gender bias is also particularly seen to be evident in the exclusion of women from specific areas of news coverage, such as politics and sports. The research revealed that gender biases and stereotypes, including the use of biased language and assigning traditional gender roles, remain prevalent in digital content. Although the study analyzed only 754 tweets, it identifies the need to address gender bias in public media to promote more equitable presentation of gender in news.

A critical aspect of gender bias in the media, particularly in the context of news headlines and abstracts is addressed in (Dacon and Liu, 2021) which is using speculative and sensational language in news articles to attract readers. It sheds light on the gender disparities in news content, where women are often portrayed as inferior, and their representation in news categories remains significantly underrepresented compared to men. The paper developed a methodology to analyze large scale news content using NLP to discover both implicit and explicit gender biases demonstrating that women are significantly marginalized and suffer from socially-constructed biases in the news.

Neural language models can reflect the bias that is present in the data they are trained on. It is important to address this because this bias can lead to these models generating text that can be offensive or harmful to certain groups of people. These models may often preserve the gender bias or stereotypes while generating large corpora, such as news articles on which they are trained on (Florez, 2020). This can adversely affect the user experience in various applications as well, such as Gmail's smart-compose feature. Thus, the research emphasized on the necessity of eliminating this memorization of gender-related stereotypes and introduced a novel architecture to separate representation learning from memory management, updating memory modules with an equal ratio across gender

types that improves user's experience.

3 Methods

In this section, we discuss the dataset used, describe the pre-processing steps taken, explain the different topic models utilized to generate topics for two countries, and outline our approach to analyze gender bias.

3.1 Data

The news dataset used is obtained from the 'News on the Web' (NOW) corpus (Davies, (2016-), which is comprised of approximately 250,000 news articles spanning 20 countries. These articles are written in English, covering news from various countries. A subset of the news sources from Pakistan and Malaysia is utilized for this work, focusing on the last six months of 2020. This subset includes 28,462 news articles from Malaysia and 30,952 news articles from Pakistan, with a total of 59,414 news articles.

3.2 Pre-processing

The text pre-processing pipeline mirrors that defined by (Rao and Taboata, 2021). It includes several essential steps to enhance the quality of the textual data for subsequent analysis. First, stop word removal is implemented to eliminate commonly used words, such as "and," "the," and "is," as they often do not contribute significantly to the overall meaning of the text. The NLTK library is employed to access a predefined list of English stop words. Subsequently, tokenization and lowercasing are performed. Tokenization involves breaking down the text into individual words or tokens, while lowercasing ensures uniformity by converting all words to lowercase. Following this we apply normalization, which removes punctuation and allows the focus to remain on the core words in the text. Lemmatization is applied to reduce words to their base or root form, aiding in standardizing words and improving the overall analysis of the text.

The final pre-processing step is relative pruning, where all words present in more than a certain threshold of documents and all words present in less than another threshold of documents are removed. For the experiments, these thresholds were set to 80% for the upper limit and 5% for the lower limit, as recommended by (Rao and Taboata, 2021). Upon implementation, this method did not remove any words over the upper limit, but removed 168,828 words that appeared less than the lower limit.

Together, these pre-processing steps contribute to a refined and standardized representation of the textual data, ready for further analysis in the following natural language processing tasks.

3.3 Topic Modeling

To generate topics for the news articles for both countries, three distinct topic modeling techniques are used: Latent Dirichlet Allocation (LDA) with both Gensim and Sklearn implementations, BERTopic, and Top2Vec. All these topic modeling techniques are configured to extract a specific number of topics, namely 10 topics. Then, we evaluate the topic diversity and coherence scores for each of the models to assess their performance and effectiveness in the given context. Based on the assessment, Sklearn LDA model is chosen for further implementation. For each topic generated by the chosen LDA model, we establish appropriate labels generated by ChatGPT by listing at least 10 representative words from every topic to better define the their characteristics. We additionally calculate topic document matrices for both Malaysia and Pakistan which are used in the subsequent steps of our approach for gender bias analysis.

3.4 Gender Bias Analysis

The gender bias analysis within the corpus of news articles from Pakistan and Malaysia is structured into two stages. The first stage analyzes which gender appears the most in each news article or document across the topics or how often words for men and women are used across each document. The second stage involves identifying how much men and women are talked about in news articles about different topics.

Analysis in the first stage is done by calculating the proportions of male-female gendered words in each document using two methods. The second stage generates male and female gender scores that signifies if men or women are talked about more in news about a particular topic.

3.4.1 Predefined Word List

The first method in the first stage uses a predefined word list of male-female pronouns, such as, "he", "him, and "his" for male and "she ","her", and "hers" for female. By counting how many times each gendered pronouns appear in the documents,

we calculate the ratios or proportions for each gender in each of the news articles or documents.

3.4.2 GloVe Embedding

The second method generates a list of top 10 associated words to a particular seed word for each gender based on GloVe embedding similarity scores. Specifically, we used the Wikipedia 2014 + Gigaword 5 pretrained GloVe word vector, which is uncased and contains 6 billion tokens. When generating our own lists, "man" and "woman" are used as seed words to generate the male and female gendered word lists respectively. The generated list for each gender is shown in Table 1. However, it was noticed that the similarity scores also brought up a few gendered terms from the opposite gender. For example, while using "male" as a seed word, the top associated word found was "woman", which is not ideal for this analysis. The issue is solved by generating the lists which are pruned of opposite-gendered words and gender-neutral words such as "person". The opposite-gendered lists were generated by ChatGPT. The second method now calculates the ratios or proportions for each gender in the news articles the same way as the first method discussed in 3.4.1 based on the generated lists.

3.4.3 Gender Scores

The second stage, then, focuses on calculating gender scores for male and female references across each topic in the two countries. In the previous stages, each document is analyzed to determine the proportion of male and female gendered words and two topic-document matrices are generated for each country as an output of the topic modeling process. The gender proportions and the topic-document matrices are then used to compute a weighted average for each gender in each topic. The computed weighted average for each gender serves a gender score, that indicates gender representations in specific topics. The following formula is used for calculating the gender score. $\sum \left(df_{\text{topics\_c}} \times \text{ratios}_{\text{gc}}\right)$ represents the sum of weighted topics by element-wise multiplication of each topic's proportion in the document from the topic-document matrix for each country with each corresponding gender ratio for that document. The sum of the weighted topics is then divided by the sum of each gender ratio for each country that provides the total weight, resulting into a weighted average. We multiply it by 100 to represent as a percentage.

$$\text{Gender Score}_{\text{c}} = \left(\frac{\sum(\text{df}_{\text{topics\_c}} \times \text{ratios}_{\text{gc}})}{\sum \text{ratios}_{\text{gc}}}\right) \times 100$$

4 Results

In this section, we present the results obtained from the topic modeling and gender bias analysis.

4.1 Topic Modeling

Implementing four different topic modeling techniques with distinct approaches yielded 10 topics from each model. Some of these generated topics were somewhat similar in all the models. Table 2 displays some of the outcomes for each model along with the associated words representing each topic.

We evaluate the performance of topic models by using two key metrics: topic diversity and coherence score. Coherence focuses on the quality and interpretability of individual topics, while topic diversity focuses on the variety and distinct nature of the entire set of topics. Table 3 shows the evaluation results for the four topic models. A more detailed analysis and discussion of these topics can be found in Section 5.

4.2 Gender Bias Analysis

We present three visual comparisons of our gender bias results. Figure 1 shows bias scores across all genders for all regions. It contains visualizations for scores calculated with both the predefined lists of words as well as GloVe-generated lists of words.

In Figure 2, we show each topic per region and the percentage of male or female references that make up each topic. Each topic is stretched to 100% so we can see the fine-grained details within each topic, as Figure 1 makes it difficult to see the differences in lower-scoring topics such as "Media" and "Property".

Finally, we include Figure 3, which compares same-gender bias scores between regions (e.g. male scores in Pakistan vs male scores in Malaysia). This chart is useful to see how bias changes as we move between regions.

A more detailed analysis and discussion of these charts can be found in Section 5.

5 Discussion

In this section, we discuss the results obtained from the topic modeling and gender bias analysis.

Gender	Generated Word List
Female	woman, girl, mother, child, herself, victim, wife, she, teenager, couple
Male	man, boy, one, turned, another, whose, once, life, thought, victim

Table 1: Generated word lists by GloVe embedding for each gender

Model : Topic	**Topic Words**
Sklearn LDA:1	Said, Minister, Government, Project, Meeting, Development, Education, Country, Also, Student
Sklearn LDA:2	News, Day, Pakistan, One, Best, People, Entertainment, Time, Work, Year.
Sklearn LDA:3	Said, Opposition, Party, Minister, Government, Election, Political, Leader, People, Prime, Country, President.
Gensim LDA : 1	Run, One, Back, Time, Well, People, Year, Also.
Gensim LDA : 2	New, China, Technology, Digital, Mobile, Also, User, Company, Online, Feature.
Gensim LDA : 3	Government, Said, Also, Project, Development, Would, Area, Province, Meeting, Sector.
BERTopic : 1	Said, Also, Pakistan, Government, Property, New, Case, Year, Minister, People
BERTopic : 2	Game, Account, Social, Also, Get, User, Time, Facebook, First, Video
BERTopic : 3	Car, Model, New, Also, Note, Vehicle, Like, Get, Name, Come
Top2Vec : 1	News, Communication, Story, New, Fresh, Announced, Published, Press, Tuesday, Daily
Top2Vec : 2	Police, Officer, Political, Policy, Official, Deputy, Authority, Reported, Government, News
Top2Vec : 3	China, Chinese, Economy, Economic, Government, President, News, Market, Billion, Global

Table 2: Visualization of words generated for each topic model.

Model	Diversity	Coherence
Sklearn LDA	0.787	0.559
Gensim LDA	0.85	0.479
BERTopic	0.716	0.464
Top2Vec	0.482	0.419

Table 3: Evaluation results for our four topic models.

5.1 Topic Modeling

The topics generated by the different topic models can be seen to provide meaningful and insightful associated words for each topic. In addition, we can also see in Table 2, some generated topics are very similar across different models. For example, topics related to government, health and technology are present in most of these models, with specific words consistently appearing across each of these topics.

Further, we observe that Sklearn LDA and Top2Vec models provide valuable insights by generating meaningful words for each topic. However, while default settings in BERTopic generate meaningful words for each of the numerous topics (around 75), limiting the number of topics, especially when set to 10 like other models, does not consistently provide the same meaningful results. We also notice that Sklearn LDA and Gensim LDA models generates common words for topics such as Politics.

The model that we choose to work with for gender bias analysis is based on the evaluation of the diversity score and coherence score. We find that the Top2Vec model provides a lower score for both of the metrics, but it generates meaningful topics with relevant words for each topic. BERTopic consistently ranked lower compared to the LDA models. In case of Sklearn LDA and Gensim LDA, the evaluation metic scores are closely aligned, with almost no significant difference between them. However, we choose to proceed with Sklearn LDA because it has the highest coherence score and the topic diversity score is also desirable. The 10 topics generated by Sklearn LDA are assigned labels generated by ChatGPT to provide more context. The labelled topics include "Policy", "Lifestyle", "Politics", "Pandemic", "Property", "Media", "Regional", "Legal", "Tech", and "Economy". Table 4 displays some of the topic labels generated by ChatGPT and associated words in each of the topics.

5.2 Gender Bias Analysis

Our results revealed insights similar to what we expected. As shown across all bias results (Figures 1, 2, and 3), the topic "lifestyle" was always skewed to primarily focus on women and "politics" and "economy" were skewed with a focus on men.

Because our gender bias metric uses percentages of gendered words that skew towards either gender, our results also show us which topics contain the most gendered words. For example, our results for topics such as "Property" and "Pandemic" have small gender bias scores, primarily because few gendered words are used in those articles compared to other more human-centric topics like "Lifestyle" and "Politics". Because of this, it is important to consider the differences within single topics when analyzing the severity of biases instead of comparing topics to one another.

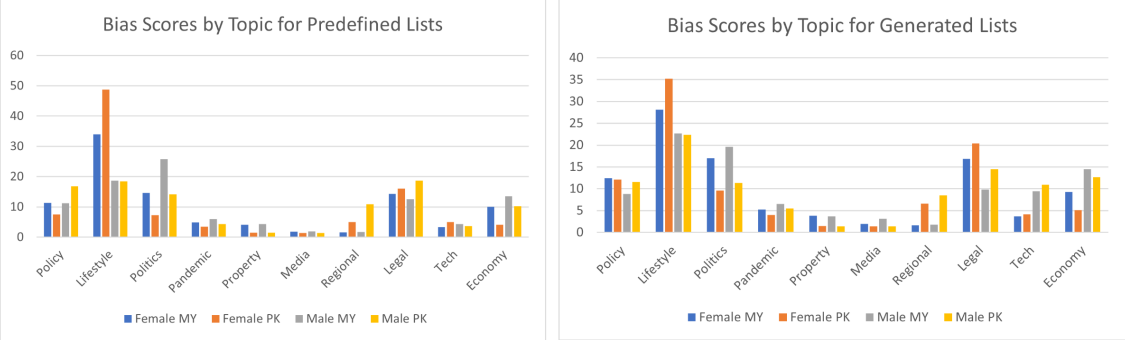

Figure 1: A comparison of all gender bias scores across both regions. Predefined lists of words, described in section 3.4.1, are used on the left and generated lists of words, described in section 3.4.2, are used on the right side.

Labeling : Topic	Topics Words
Policy: 1	Government, Project, Development, Education
Lifestyle: 2	Entertainment, People, World, Life, Medium
Politics: 3	Opposition, Party, Government, Political, Leader
Pandemic: 4	Coronavirus, Hospital, Health, Death, Infection
Property: 5	Property, Land, Sale, Price, Comprehensive

Table 4: Labeling topics for an Sklearn LDA.

When comparing gender bias scores in individual regions, we see many similarities between the PK and MY datasets. Specifically in Figure 2, we see that both regions share the same female-bias in the "Lifestyle" topic. However, we also see that some biases are amplified in different regions, such as "Policy", "Regional", and "Economy" being more biased towards men in the PK dataset than they are in the MY dataset. Interestingly, we also see that the topic "Technology" is skewed slightly more towards women in Malaysia but slightly more toward men in Pakistan.

Another interesting result that occurs throughout the dataset is the effect of the GloVe-generated lists of gendered words. In some topics, such as "Lifestyle" and "Politics", we see that the generated word lists reduce the severity of the biases present in the predefined word lists. But in other topics, such as "Technology" and "Legal", the generated word lists amplify the bias. In the case of the topic "Legal" in the PK dataset, the generated lists actually flips the bias in the opposite direction, skewing towards more women than men. These differences may be due to the inclusion of many non-gendered words in the generated word lists, such as "whose", "once" and "life" being associated more with men.

When we compare both regions to one another, such as in Figure 3, we see many similarities in where bias are present. Both regions give an almost identical score to men in "Lifestyle", and also fairly similar scores to men in "Media" and "Tech". However, we see very large differences between men in "Regional" and "Politics", suggesting that there is a larger bias towards men in these topics in Pakistan.

6 Ethical Considerations

As with all studies on bias, it is important to remember that all regions analyzed have rich and diverse cultures that may not be perfectly captured in a single bias metric. While our results revealed certain potentials for bias, these biases are difficult to quantify and may not properly reflect the reality. It is also important to note that all studied biases are localized to news outlets and again may not reflect the day-to-day biases different groups encounter on a personal scale.

7 Limitations

One limitation of our project is the selection of countries from the corpus. We opted with two countries, Malaysia and Pakistan, as the countries are considered to share common aspects such as the religion, economy, and political structure [1]. This

[1]Comparison metrics can be found on the sidebar at https://en.wikipedia.org/wiki/Pakistan and https://en.wikipedia.org/wiki/Malaysia

Figure 2: A comparison of gender bias scores per topic per region. Predefined lists of words, described in section 3.4.1, are used on the left and generated lists of words, described in section 3.4.2, are used on the right side.

similarity influences our results, with nearly every topic showing a similar pattern for gender bias. It might be beneficial to choose different countries for a more diverse and representative analysis. Additionally, our dataset only covers news from the last 6 months of 2020. It would be more effective to expand it to the entire year to achieve better results and analyze the gender bias for the entire year of 2020 or more recent years. However, this process will be time consuming to run and execute the code considering our current dataset contains around 59,000 news articles.

8 Conclusion

Our findings were consistent with our expectations and showcased gender biases present in Malaysian and Pakistani news sources. These biases were most present in human-centric topics like "Lifestyle" and "Politics", and had similarities and differences across regions. Our research showed us that the LDA topic model has earned its spot as the most popular topic model for most applications and that other neural models, while interesting, have their own unique use cases. Future work includes extending this analysis to cover additional regions and experimenting with new word sets and gender bias metrics.

Code for this project can be found at https://colab.research.google.com/drive/1DQWquJLpyJMnP1Ir4baCuYsAWAu-RiMg

9 Roles and responsibilities assigned to each member of the group

We collaborate as a team and hold regular meetings to discuss our progress. We support each other in areas where we face challenges. In this section you observe that certain tasks are shared among team members. Many smaller tasks, such as debugging and proofreading, were shared across all members as well.

Norah:

- Coding Task: Working on datasets, implementing Sklearn LDA, Gensim LDA, and BERTopic, calculating topic diversity and coherence. Collaborating with Modhumonty on calculating the weighted average for gender scores.
- Writing Task: Writing introduction, literature review on topic modeling, datasets, preprocessing, methodology, result, discussion for the topic modeling section and limitations.

Modhumonty:

- Coding Task: Working on datasets, some parts of initial pre-processing, implementing

Figure 3: A comparison of same-gender bias scores between regions. Predefined lists of words, described in section 3.4.1, are used on the left and generated lists of words, described in section 3.4.2, are used on the right side.

Top2Vec, calculating topic diversity and coherence, weighted average calculations for gender scores in each region for for gender bias evaluation

- Writing Task: Literature review on gender bias, datasets, methodology for gender bias analysis, result and discussion on topic modeling.

Gina :

- Coding Task: pre-processing and relative pruning, working on datasets, calculating topic diversity and coherence, glove embeddings, gender bias calculations per document with predefined/generated lists
- Writing Task: (Rao and Taboata, 2021) Literature Review, pre-processing, gender bias results and charts, discussion of gender bias results, ethical considerations, and conclusion

References

Dimo Angelov. 2020. Top2vec: Distributed representations of topics.

Weisi Chen, Fethi Rabhi, Wenqi Liao, and Islam Al-Qudah. 2023. Leveraging state-of-the-art topic modeling for news impact analysis on financial markets: A comparative study. *Electronics*, 12(12):2605.

Jamell Dacon and Haochen Liu. 2021. Does gender matter in the news? detecting and examining gender bias in news articles. In *Companion Proceedings of the Web Conference 2021*, WWW '21, page 385–392, New York, NY, USA. Association for Computing Machinery.

Mark Davies. (2016-). Corpus of news on the web (now). Available online at https://www.english-corpora.org/now/.

Omar U. Florez. 2020. On the unintended social bias of training language generation models with news articles.

Maarten Grootendorst. 2022. Bertopic: Neural topic modeling with a class-based tf-idf procedure.

Mykyta Kretinin and Giang Nguyen. 2022. Topic modeling on news articles using latent dirichlet allocation. In *2022 IEEE 26th International Conference on Intelligent Engineering Systems (INES)*, pages 000249–000254.

Ana María Muñoz Muñoz and Juana Salido Fernández. 2023. Sesgos de género en las redes sociales de los medios públicos autonómicos: el caso del twitter de @csurnoticias. *Revista Latina de Comunicación Social*, (82):1–16.

Prashanth Rao and Maite Taboata. 2021. Gender bias in the news: A scalable topic modelling and visualization framework. *Frontiers*. Https://doi.org/10.3389/frai.2021.664737.

Michael Weiss and Steven Muegge. 2019. Conceptualizing a new domain using topic modeling and concept mapping: A case study of managed security services for small businesses. *Technology Innovation Management Review*, 9(8).

A Additional Results

Raw values for the results of gender bias analysis are included for completion in Tables 5 and 6. Topic words generated by Sklearn topic modeling technique for the 10 generated topics are included in Table 7.

	Female		Male	
Topic	MY	PK	MY	PK
Policy	11.34	7.54	11.24	16.85
Lifestyle	33.98	48.8	18.67	18.46
Politics	14.64	7.29	25.79	14.18
Pandemic	4.84	3.49	5.94	4.28
Property	4.15	1.43	4.3	1.43
Media	1.84	1.41	1.97	1.41
Regional	1.54	4.98	1.69	10.87
Legal	14.28	16.06	12.51	18.68
Tech	3.37	4.94	4.37	3.64
Economy	10.02	4.06	13.52	10.2

Table 5: Raw gender scores for all regions using predefined word lists.

	Female		Male	
Topic	MY	PK	MY	PK
Policy	12.4	12.13	8.79	11.55
Lifestyle	28.14	35.21	22.71	22.34
Politics	17.03	9.57	19.58	11.33
Pandemic	5.27	3.99	6.53	5.49
Property	3.85	1.43	3.69	1.41
Media	1.93	1.4	3.15	1.38
Regional	1.6	6.61	1.8	8.45
Legal	16.81	20.42	9.79	14.49
Tech	3.67	4.13	9.46	10.92
Economy	9.3	5.11	14.51	12.64

Table 6: Raw gender scores for all regions using generated word lists.

Labelled Topics	Topic Words
Policy	said, minister, government, project, meeting, development, education, country, also, student
Lifestyle	news, day, pakistan, one, best, people, entertainment, time, work, year
Politics	said, opposition, party, minister, government, election, political, leader, people, prime
Pandemic	coronavirus, case, virus, hospital, health, death, hour, test, 24, reported
Property	property, breaking, land, malaysia, collection, launch, sale, commercial, price, type
Media	article, access, july, pic, 2018, full, free, effective, 2020, available
Regional	pakistan, karachi, sindh, indian, india, imran, islamabad, punjab, khan, pakistani
Legal	court, police, said, case, pakistan, justice, punjab, government, officer, report
Tech	mobile, device, user, phone, pakistan, feature, new, application, company, system
Economy	pakistan, billion, million, price, company, per, bank, year, market, rate

Table 7: Sklearn LDA Topic Model: Topic Words for 10 Generated Topics