INDIAN NEWSPAPERS' DESCRIPTION OF THE TOKYO 2020 OLYMPICS

MEKHALA KUMAR, UNIVERSITY OF MASSACHUSETTS AMHERST

TEXT AS DATA 697D, FALL 2022

INTRODUCTION

Newspapers often reflect the gender biases and gender roles in society.

Rao and Taboada found that English Canadian newspapers quote women more often in the Lifestyle, Entertainment, Arts and Healthcare categories and men more often in the Business, Sports and United States Politics (2021). Even within a field such as sports, the events are described for men while for women, only their achievements are focused upon.

Similarly, Devinney et al. studied Mainstream English news articles, Mainstream Swedish articles and LGBTQ+ web content and found that feminine topics were linked to the private sphere and masculine topics were linked to the public sphere (2020).

Therefore, this project aimed to :
Understand whether there was a difference in the way Indian newspapers reported women's and men's sports during the Tokyo Olympics held in 2021.

DATA

The LexisNexis database was used to collect articles from July 22 to August 9, 2021 (the time when the Olympics were held). The data included articles from Hindustan Times, Times of India, Free Press Journal, The Telegraph, Indian Express, Mint, DNA, India Today, The Hindu and Economic Times (some of these were the online versions of the newspapers).

The key word searched was Olympics and filters including Men's Sports, Women's Sports, Sports Awards and India were used.
The semantic network was created using a corpus which had 1128 articles.

METHODOLOGY

The quanteda package was used for preprocessing. The corpora used were either the entire set of files or a subset depending on the model used. Punctuation and stopwords were removed from the corpora. Additionally, words such as Olympics, India and Tokyo were removed to derive more meaningful results.

Structural Topic Modelling and LDA Topic Modelling were employed using the stm and topicmodels packages respectively. For this, subsets of the dataset were utilised to create corpora. These corpora were made using the metadata which had classification tags such as sports, women's sports and men's sports. The articles were categorised as either men's sports or women's sports.

For structural topic modelling, the corpus had 468 articles. LDA topic modelling was run using separate corpora for men's sports (191 articles) and women's sports (277 articles). I used the search_K() function to determine the number of topics for the LDA and structural topic models.

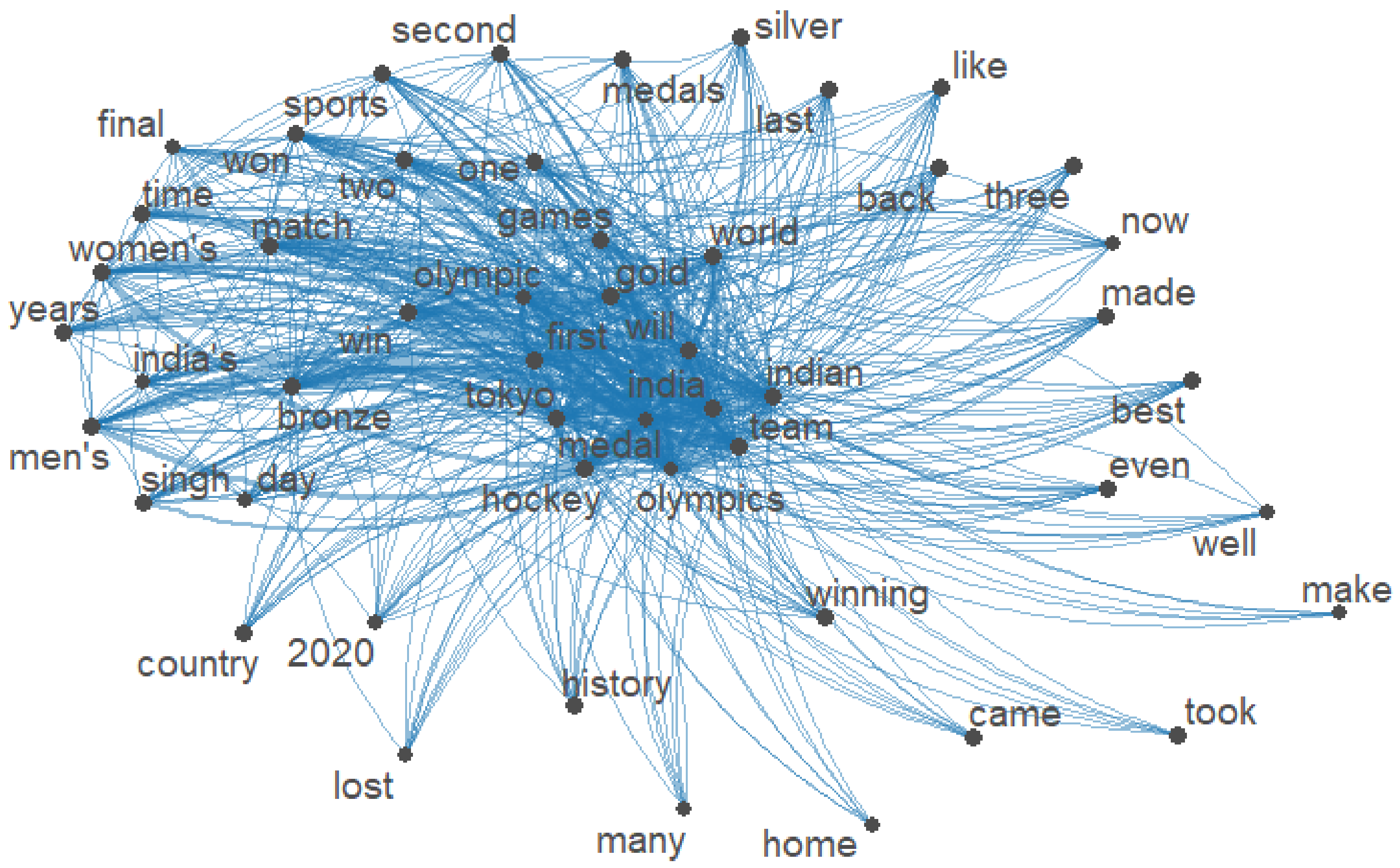
CONCLUSION

To summarise, there was no difference in the language used to describe women's and men's sports. Most of the athletes that were mentioned in the Indian newspapers had won medals or progressed to the final rounds of their respective sports. However, the modelling allowed us to see the popular aspects that were discussed in India during the Olympics, the most popular being the victory of the Indian men's hockey team. Moreover, an interesting find was that the media mainly focused on the sports players who won medals.

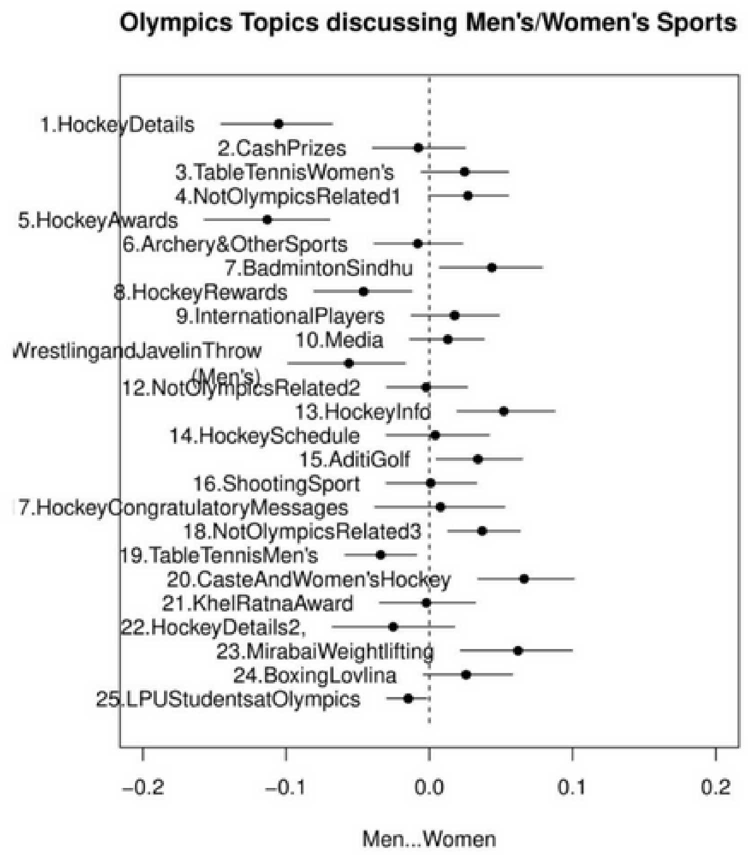
One of the limitations was that the classification was not proper in metadata, there were a few articles where women's sports were labelled as men's sports and vice versa.

Future research can incorporate more categories beyond sports and a longer time period in order to determine whether a gender bias in Indian newspapers exists.

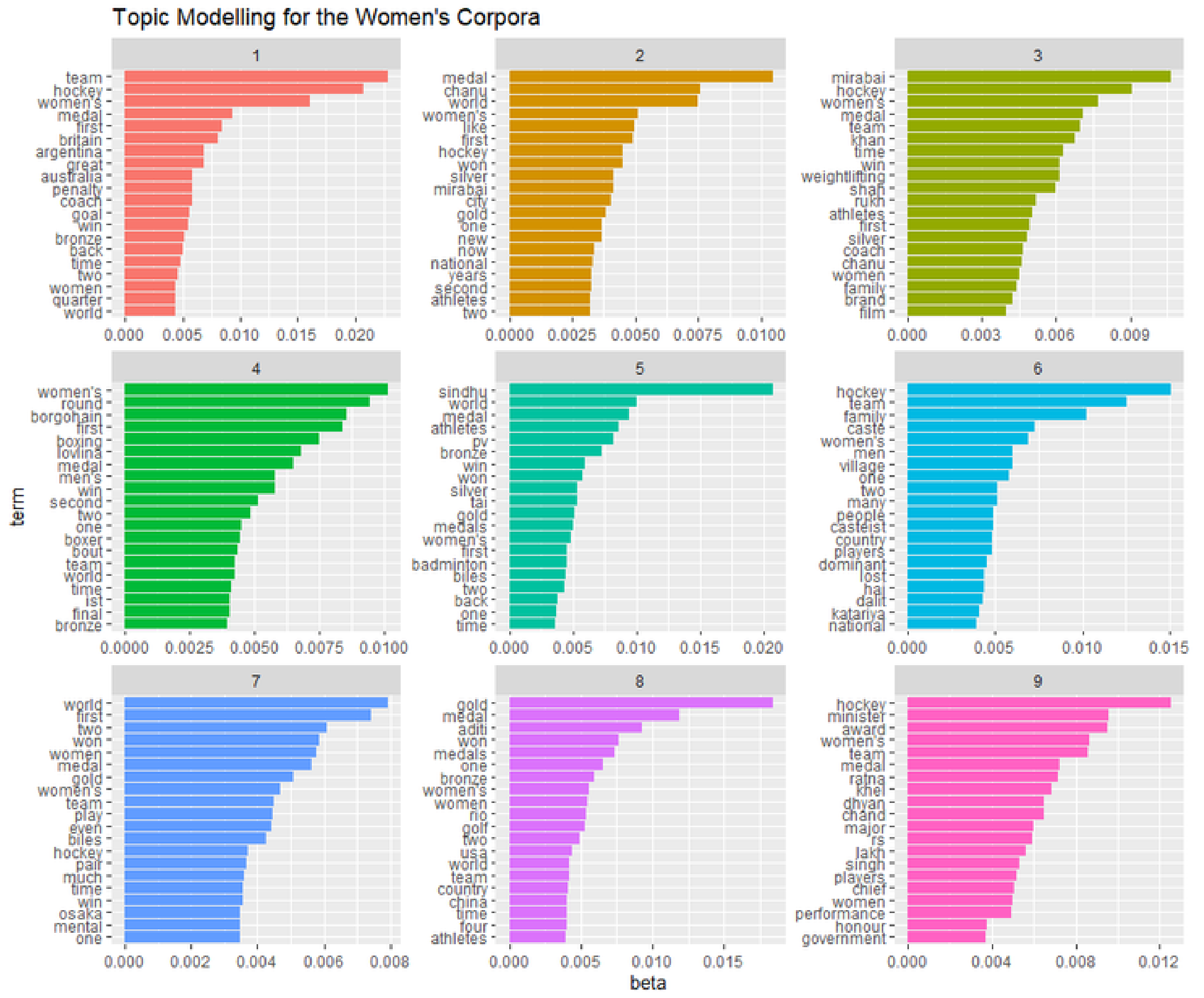
RESULTS

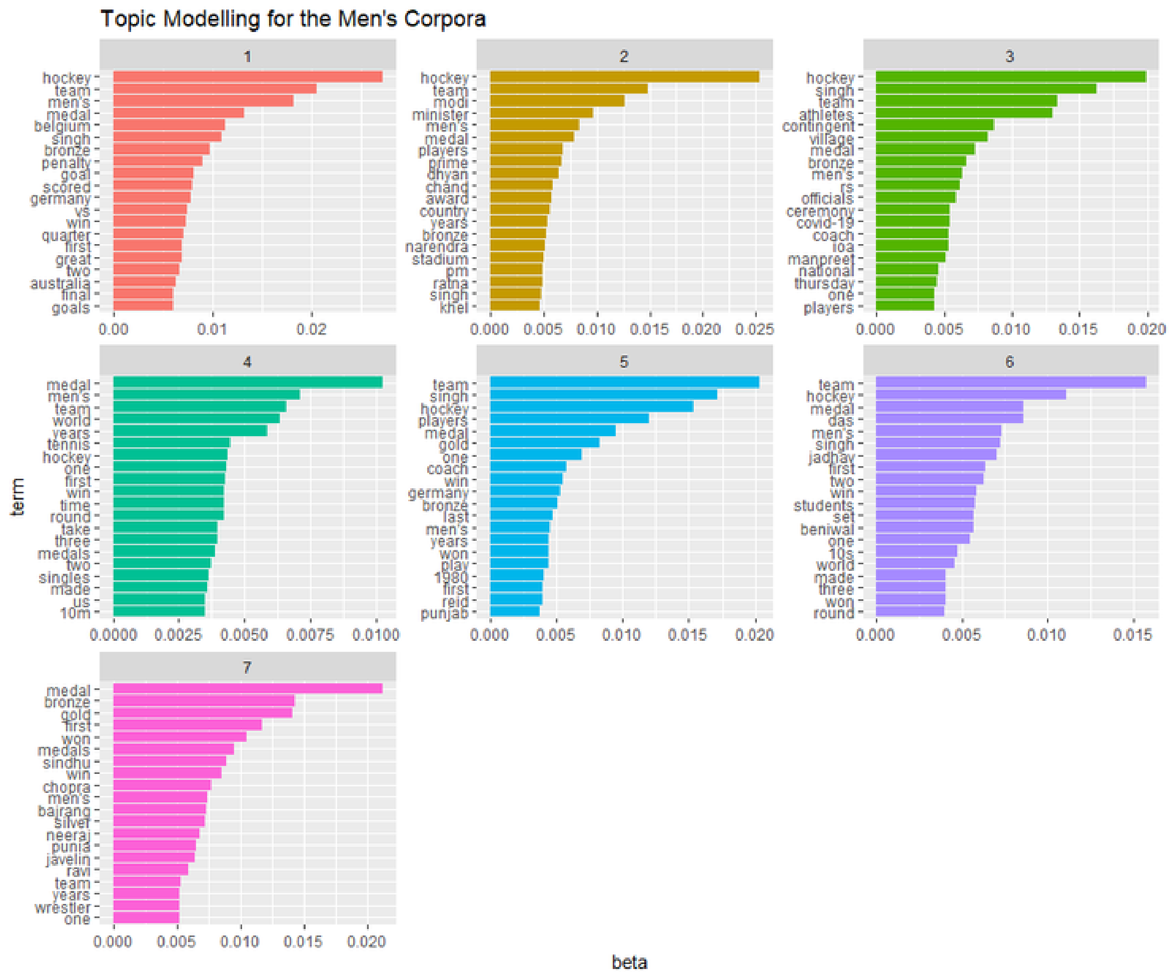


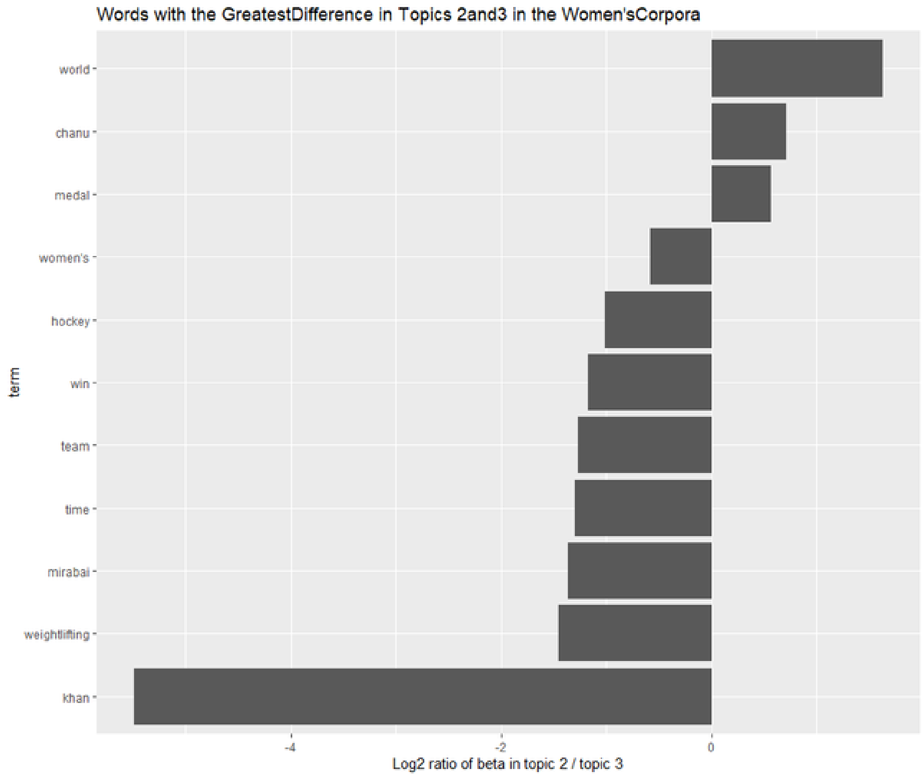
Semantic Network of 1128 articles
One major theme that can be observed is the discussion of the hockey team- the men's team had placed third in over four decades hence marking history and was led by the captain Manpreet Singh. Other significant terms include medals and medal colours pertaining to victories by other Indian athletes.



Estimated effect on Topic Prevalence
The graph illustrates how much more (or less) the topic is mentioned when the article is tagged as Women. It does not give any insightful results because it shows the obvious result that the sports that had women athletes in them were mentioned more when the article was tagged as women and sports that had men athletes were mentioned more when the article was tagged as men.







Topic Modelling for the Women's Corpora
Most of the topics in the women's sports corpora are about the Indian women athletes who won medals at the Olympics or were in the final rounds. Other than this, there was an incident where casteist remarks about Indian women hockey players were made after the women's team had lost a semifinal which is reflected in topic 6. Finally, when it came to international athletes and events, the only topic found was about Simon Biles and her decision to leave the Olympics early due to mental health reasons.

Topic Modelling for the Men's Corpora
Most of the topics are regarding the men's hockey team's victory, including the details of the match and people's reaction to the same. Other people discussed in the corpora as well are medallists. Moreover, even though this was the men's corpora, the female Badminton player PV Sindhu was among the top terms in topic 7. This shows that the tags present in the metadata were not completely accurate.

Words with Greatest difference in Topics 2 and 3 of the women's corpora
Topics 2 and 3 both have words related to Mirabai Chanu's success in weightlifting and about hockey so the words with the greatest difference in the 2 topics was looked into. The words that are more common in topic 2 include world, chanu and medal whereas the words in topic 3 include hockey, win, team, time, mirabai, weightlifting and khan. This is indicates that topic 2 has information specific to weightlifting and topic 3 is a mixture of the two sports.

REFERENCES

Devinney,H., Björklund,J. & Björklund,H.(2020). Semi-Supervised Topic Modeling for Gender Bias Discovery in English and Swedish. Proceedings of the Second Workshop on Gender Bias in Natural Language Processing, 79–92.
https://aclanthology.org/2020.gebnlp-1.8

Nexis Data Lab (2022). Olympics. Retrieved October 24, 2022,https://advance.lexis.com

Rao P and Taboada M (2021) Gender Bias in the News: A Scalable Topic Modelling and Visualization Framework. Front. Artif. Intell. 4:664737. doi: 10.3389/frai.2021.664737