Xiao Sun

July 30, 2024

My research focuses on China’s parental leave policies, specifically investigating the main themes in media coverage of these policies. To address this question, I developed a dataset comprising news coverage on China's parental leave policies from 2015-2024, sourced from three mainstream news outlets: Global Times, China Daily, and Xinhua News Agency, obtained from the NewsBank database (English version).

I extracted relevant text data from the NewsBank PDF files using pdfplumber in Python. The dataset includes the full text of each news article, the date, and the media source. After extracting the data, I cleaned and saved it as a CSV file. My unit of analysis is the article, excluding the title and author but including the date and media source.

For preprocessing, I utilized stopwords tools from gensim and created my own custom stopwords list to remove unimportant words, such as "said," "year," "time," and "people." Additionally, I performed lemmatization for simplicity, manually merging similar words not included in gensim, like "woman" and "women," "child" and "children," etc. This customization aimed to improve the results.

I vectorized the text data by converting it into a numeric format using the Bag of Words (BoW) method. First, I created a dictionary that maps each unique word to an integer ID. Then, I converted each document into a list of tuples, where each tuple contains a word ID and its frequency in the document. Finally, I used the BoW representation of the documents to perform LDA modeling.

I experimented with LDA models using 3, 4, 5, 6, 7, and 10 topics, tuning the models by calculating coherence scores. The scores ranged from 0.24 to 0.38, indicating that the words within a topic were not highly coherent or meaningful. I also found it challenging to label each topic consistently (see Figure 1), likely due to the limited dataset of only 19 documents. According to Figure 2, five topics appeared optimal, but several topics were quite similar. Therefore, I chose four topics, resulting in a coherence score of 0.37.

A graph with blue and black text

Description automatically generated

Figure 1. Top words within 4 topics

A graph with a line

Description automatically generated

Figure 2. Coherence scores for 1-10 topics

Despite the limitations of my data, the unsupervised method was helpful in identifying dominant topics. Although some words lacked coherence, I was able to label the four topics as follows:

1. Parental Leave Policies and Employment
2. Family Support and Parental Policies
3. Demographics and Policy Development
4. Role of Fathers and Gender Dynamics

By plotting the distribution of dominant themes across different media outlets, I observed that the three media outlets had different focuses (see Figure 3). China Daily predominantly covered parental leave policies and their impact on employment. Global Times heavily focused on the role of fathers and gender dynamics of parental leave. Xinhua News Agency emphasized population change and policy development.

In summary, the unsupervised method enabled me to explore and identify the main themes in media coverage of China's parental leave policies. Despite the challenges and limitations due to the small dataset, the findings provide valuable insights into how these policies are portrayed by different media outlets.

A graph with different colored squares

Description automatically generated

Figure 3. Distribution of dominant themes