**Abstract**

As a student focusing on East Asian Studies, the topic of the project will be the Chinese's people attitude toward the government's birth-encouraging policies online. China has been loosening its birth constraint and launching birth-encouraging policies for the past few years. Although it is not hard to imagine that most of the general opinion might not favor these, I am interested in what people on particular topics complain about or discuss regarding the policies. My conclusion for the analysis is that there is not much different for people's attitude toward birth-rate related policies across 2021~2023, and the discussions were most likely to be response to government news.

**How to read the notebooks**

The analysis is consisting of three notebooks: Data Collection, Text Analysis, and Appendix, along with some folders, files, graphs for use. Please read the notebook with the below flow:

Script1: Data Collection – Weibo web scarping

This notebook includes the introduction to the project, and the data collection part of the analysis.

Script2: Text Analysis - Weibo comments

This notebook is the main analysis, starting with stating the analysis framework, then go through the detail codes with conclusion for each step, and coming to the main conclusion with next steps and limitations of the research.

Script3: Appendix - Zhihu web scarping and text analysis

This notebook contains the Zhihu web scraping and text analysis I did but decided not to use as for project. The main reason is that the web scraping results were not ideal as I lost 30%~70% of comments for each post for technical reasons. Due to bias concerns, I decided only using Weibo and put the part of Zhihu as appendix.

**Script 1: Data Collection**

* Intro

China is one of the countries facing low birth rate issue, meanwhile, China's population decreased by 850,000 in 2022, marking the first negative growth in over 60 years despite years of work trying to avoid the situation. China has repealed One-Child policy in 2016 and launched birth-encouraging policies since even earlier years. However, birth rate has not increased as the government anticipated. (<https://www.npr.org/2023/01/17/1149453055/china-records-1st-population-fall-in-decades-as-births-drop>)

The topic of this project will be understanding the attitude toward the government's birth-encouraging policies online of the Chinese people. Intuitively, it is not hard to imagine that the general opinion may not be very supportive, but I am interested in the degree on dissatisfaction across time and what aspects people discussed about most when commenting on the topic. To do so, I went through the process of data collection, data cleaning, and text analysis. For text analysis, I performed sentimental analysis to answer my question of people’s degree of satisfaction/ dissatisfaction from 2021-2023, used word clouds to demonstrate the overall most-frequent word, and lastly used tf-idf to identify year-specific key words and further interpret possible reasons behind.

* Framework of web scraping:

Unfortunately, the main data sources I wish to use for the analysis do not provide API for researchers, and hence required me to scrape data. Below is a simple framework to demonstrate the main steps of web scarping:

1. Identify the data to be scraped: The first step is to decide what data you want to obtain. For this project I selected to candidates and extract the text only.
2. Select a web scraping tool: The second step is to choose a web scraping tool. In class we used BeatuifulSoup as example. However due to the structure of the website, I used Regular Expression (regex/re) to extract the content.
3. Identify the website's structure: The third step is to investigate the website structure and source data to identify where the content I want is located. I performed this by leveraging Chrome’s developer tools to view the HTML and CSS>
4. Write a web scraping script: The fourth step is to create a web scraping script based on my understanding from step 3. We did similar thing with the Met website by using requests, parsing, etc. in class.
5. Clean and store the data: Clean and process the extracted data to make it usable for your analysis or application. Store the data in a structured format, such as a CSV or database, for easy retrieval and analysis.

* Determine the data source: Weibo vs Zhihu

To obtain the data representing public’s opinion most, I selected Weibo (China’s twitter) and Zhihu (China’s Quora) to compare and eventually used Weibo. The main difference between the two is that Weibo is quantitatively more advantageous in comments while Zhihu is better in quality of response. In other words, Weibo may contain more comments from more people, but may just be a short lol. However, for Zhihu, there is like to have more discussions with key words such as “not practical”, “should do XX instead of YYY.” As I am not sure which data source will provide a better result, I scrape both data sources, ran tokenization and sentimental analysis to compare and decided to proceed with Weibo as it has more in quantity and also more balanced amount of data across the three years. I will describe only the process of how I selected which Weibo to scrape below. If interested in the process of scaping Zhihu and the data quality check, please refer to the appendix.

* Main Reference for this script

I used and adapted the code from this link script from this link <https://zhuanlan.zhihu.com/p/443802888> to create the scraping script.

* Decide what Weibo comments to scrape

The second step of the analysis is to decide the particular posts and their comments to use for analysis. For this step, I leverage a website (Link: <https://weibo.zhaoyizhe.com/#>) that recorded most popular Weibo (ranked by view times in 24 hours) and searched birth-policy related words to find the most popular posts and select the ones with most comments for web scraping, the URLs I selected to scrape as below:

1. https://m.weibo.cn/detail/4727660435213482

2. https://m.weibo.cn/detail/4635622721458463

3. https://m.weibo.cn/detail/4643035767642831

4. https://m.weibo.cn/detail/4726596705715427

5. https://m.weibo.cn/detail/4660999765890124

6. https://m.weibo.cn/status/4635609970509108

7. https://m.weibo.cn/detail/4660990257666817

8. https://m.weibo.cn/status/4875327685790516

9. https://m.weibo.cn/detail/4881790165059616

10. https://m.weibo.cn/detail/4734289247207643

11. https://m.weibo.cn/detail/4734283263247434

12. [https://m.weibo.cn/detail/4734299180368580](https://m.weibo.cn/detail/4734299180368580%3cbr)

After selecting the posts, I create below script to scrape the comments and save them as excel file for ease of eyeballing the results and perform preliminary data cleaning and combination through excel. Then, I convert the excel files into txt format through below code, and upload them to a folded name “weibo by year” in jupyternotebook.

Weibo is known to be hard to scrape as their data is extremely valuable. I tried to battle with the website for a couple of days and then realized that the main trick: use mobile url but not web url. Mobile link is somehow not well-protected as web url, all I need to do is to provide my login infos. However, there are also drawbacks of using mobile url to scrape, it is much slower in speed and often breaks down. I did my best to collect as many comments as my computer and time allowed.

**Script 2: Text Analysis**

* After finishing data collection, I now proceed on text analysis with the following framework.
* Analysis Framework:

After finishing data collection, I now proceed on text analysis with the following framework:

1. Hypothesis:

First, I hypothesize that the most mentioned words when people comment on the birth policies will change at some level over time, despite the fact that the reasons for the low birth rate could be assumed to be similar.

2. Data selection and collection:

To select data for analysis, for the timeframe, I chose 2021, 2022, and 2023 as the timeframe for recency. As for the text content to analyze, I chose Weibo as the source as it is the most popular social network in China (basically twitter in China) to use as a proxy of people's opinions online. To further narrow down what to analyze on, I leverage this website (Link: https://weibo.zhaoyizhe.com/#) by first typing in different keywords regarding a birth policy or raising children and selected the most popular posts for each year recorded on the "hot topic billboard" with related contents and performed web scrapping to obtain the comments under the original post. For the data collection methods, I created a script based on existing Weibo crawler to deal with Weibo's anti-crawling algorithm which updates often.

3. Data cleaning and analysis:

To test my hypothesis, I performed sentimental analysis, word count ranking with word cloud, and TF-IDF to

First, for sentimental analysis, I compared the distribution plot of sentimental scores for the 3 years to see whether there were observable changes across time. Secondly, I conducted a simple word count ranking with word cloud as the visualization method to see overall what popular words were for the 3 years, whether there were changes, and infer possible reasons for the changes. Lastly, I used TF-IDF to filter out the most relevant words of each year, comparing the difference, exploring the causes, and interpreting the implications.

Data cleaning will be along the process while performing word count and TF-IDF. I leveraged the same package used for these analyses to perform tokenization, stop word removal, etc. As I was analyzing Chinese, I used snownlp package for sentimental analysis, jieba package for tokenization, and further used sklearn.TfidfVectorizer package to vectorize the tokens.

4. Conclusion:

I summarized my finding in each step and also wrote a comprehensive conclusion of all finding and then state the limitations and next steps for further analysis.

* Step1: Load Data
* Step2: Sentimental Analysis

**Conclusion**: The average score is around 0.52 when combining three years, which is a bit higher than neutral (slightly positive), contradicting the intuition of people being not supportive to the policies. However, when we take a look at the distribution chart above, we can see that despite a high peak around 0.5, meaning most comments are scored as neutral, the other smaller but obvious peak is around 0.35~0.4. Meanwhile, positive comments are more scattered comparing to negative ones. The implication here is that although people might not be satisfied with the policies, but no "harsh words" were really used to criticize the government policies as snownlp was trained on a Taobao (China’s Amazon) dataset for sentimental analysis, where people are more direct when complaining or criticizing online shop owners.

**Conclusion**: The shape of distribution is quite similar across year. Overall, there is no big difference be identified through the graph, which the statistical description also supported. However, if we look closely, we can see that 2022 has a small peak at a more negative score (around 0.3,) this is even clearer in the second graph “standardized distribution of sentimental analysis - by year.” The reason for such phenomenon is that one of the most discussed Weibo of the year is a report published by the government stating that fertile women were giving less birth. Some people considered the report as pointing fingers to women for low birth rate, ignoring other social factors, which I further deep dived in the discussed content in the TF-IDF part.

* Step3: Create stop word list
* Step4: Tokenize and remove stop word
* Step5: Word Count Ranking
* Step6: Word Cloud

**Conclusion**: The most appeared words across the three years are straight forward and similar: children, second/third child, work, marriage, money, etc. These are all common words when discussion life plans. In short, the most frequent words with regard birth-policy are similar among the three years. However, I continued with TF-IDF to reflect the important words for respective years.

* Step7: TF-IDF

**Conclusion**:  
For 2021: If we define “hot word” as TF-IDF score over 0.0015 as significant, the significant word of the year is Chengdu. The reason is that during the period, another news about a Chengdu high school student fall from a building (reason unspecified to this date.) This incident held a lot of mystery with conspiracies. People felt that birth rate related topics should not take over the public’s focus of this.

Beyond the word, the other words are population census, private sector, age, and other work-related words. For this year, people’s discussions about giving child were more about their concerns of raising a child, especially for women. People’s main concern was that private sector companies were more inclined to not hire women at certain age with family responsibility.

For 2022: China's National Health Commission published a news about "the annual average reduction of women in their peak reproductive years is 3.4 million."  
As "women" is specified in the news title, people felt it was insinuating that women were responsible for low birth rate and population decrease, which induced discussion on women rights. In response, the government of Zhejiang Province conducted a survey on people's willingness to raise another child if the government provide 1000 RMB monthly for each child since birth for 3 years, which is 36k (3.6万) in total, and people responded that it is just a drop in the bucket(杯水车薪) which clearly classified as negative attitudes.

For 2023: The most discussed birth-rate related policies in 2023 (till March) is the government of Shandong province is planning on providing free high-school education. (China’s mandatory education is to middle age.) People are sarcastic about the policy as families can only enjoy the policy benefit after a decade(十几年) when the new-born turn 15 . Some even say that they don't think their future children can pass the high school entrance exam and prefer cash right now, or else they will just go bankrupt first. Also, some are concern whether policies with such as time gap could be implemented, feeling like it might be a promise that cannot be fulfilled.

* Conclusion & Next Step

The main conclusion of the analysis is that there are not much different across the three years for people overall attitude, the key words are similar as well. Looking into year-significant word, we can see that the discussions are initiated from governmental news and trilled down to the daily life worries of career plan, family plan, whether the society to support family-oriented decisions despite intense competition in the job market. Weibo serves as a quick reflection spot for people commenting on new topics but lacking discussion with depth. Hence, the next step for the analysis to be more informative and solid. I plan to increase the amount of Weibo comments for better quantitative analysis by adding more years, and restructure analysis method for Zhihu comments to obtain people’s deeper thoughts on the topic.

* Limitation

This analysis is limited in several aspects. First, the selection of Weibos could be improved by making sure that there are relatively same amount from each month through the year, making the comments smoothly distributed within each year. Currently, due to time and computer limitation, I focused on most popular birth-related Weibo in each year, not necessarily collect data to represent general attitude toward the topic. Second, as mentioned in the next step, comments in Weibo tend to be shorter and shallower, which I consider could be ameliorate by including data from Zhihu due to its structure of post and replies. Lastly, the sentimental analysis model of snownlp was trained on a Taobao dataset, so might be more accurate in identifying online selling/buying attitude but might not be tailorized for social topic discussions like birth rate. If time and skills allow, I would like to train my own model on sentimental analysis on people’s attitude toward these topics and taking into account on how the Chinese tend to be more conservative in words when discussing about the government by through scaling.