From: Shiaw-Shiuan (Sharon) Chuang

Subject: Partisan bias in political news reporting during election periods in Taiwan and the US.

Date: April.7.2023

A. Background

Political news media plays a crucial role in shaping public opinion and influencing electoral outcomes. However, the extent to which media coverage is biased towards one political ideology or party can have significant consequences for democracy. Partisan bias in political news can create a distorted view of events, reinforce existing political polarization, and ultimately affect electoral outcomes. Thus, understanding the role of partisan bias in shaping public opinion during the 2020 presidential election in both the United States and Taiwan is a crucial area of study.

The 2020 US presidential election was a significant and polarizing event, revealing deep political divisions in the country, with media bias playing a role in shaping public opinion (CEPR, 2021). Taiwan's 2020 presidential election was also characterized by political polarization, with the values of democracy and the China challenge being crucial (Hass, 2022). Both countries experienced a charged atmosphere with allegations of electoral malpractice. However, the nature and extent of polarization differ, with Taiwan's largely driven by issues related to national identity and cross-strait relations with China, and the United States' more focused on issues related to race, immigration, and economic inequality (Clark C, Tan AC, Ho K, 2019). Understanding how media bias and bipolar political status quo interacted to affect election results is important in both cases.

B. Data

The goal of this project is to classify news articles from different media sources in the United States and Taiwan leading up to the elections. To accomplish this, I require news-based data with labeled sources, title, release date, and category. For the United States, I use the "NELA-GT-2020" dataset from Harvard Dataverse (Horne, Benjamin; Gruppi, Mauricio, 2021),

which includes election-related news articles from CNN (CNN, 2020), Foxnews (Fox News. 2020), and PBS (PBS, 2020). For Taiwan, I generated a dataset through web scraping of political news articles from three media sources: Liberty Times (LTN, 2020) (pro-independence), Cti News (CTI, 2020) (pro-China), and a collection of Central News Agency (中央社 CNA, 2020) and The Reporter (報導者 the reporter 2020) for the more neutral media news.

Both datasets contain the same variables: title, release date, content, one hot encoded keyword (e.g., keyword_trump, keyword_biden, keyword_election, etc.), and source (Table 1 & 2). I use the title and content as the feature and the source as the target.

Variable name	Meaning	Data type
title	A string variable representing the title of the news article.	string
date	A date variable representing the date the news article was published.	datetime64[ns]
content	The preliminary **feature**, A string variable representing the content of the news article.	string
source	The preliminary **target**,A categorical variable representing the source of the news article. The three possible values for the US dataset are "cnn", "fox news", and "pbs".	categorical
keyword-xxx	A boolean variable representing whether the news article contains the certain keyword. Ex ‘keyword-trump’ represent if a article contain the keyword "Trump".	boolean

Table 1: the definition of variables

content (feature)	source (target)
MADISON Wis Wisconsin finished a recount of its presidential results on Sunday confirming Democrat Joe Biden s victory over President ….	pbs
Those things that are required that the President has directed us to do in compliance with the decision that the GSA made yesterday I ll do all of those things Pompeo said …	cnn
DEARBORN Mich The Democratic presidential primary is down to two major candidates and it shows Former Vice President Joe Biden and Vermont Sen Bernie …	pbs

Table 2: sample feature and target of the US data

1. US Data:

The "NELA-GT-2020" dataset contains nearly 1.8M news articles from 519 sources collected between January 1st, 2020 and December 31st, 2020. For election-related news, there are 7663 observations from CNN, 3926 observations from Foxnews, and 2280 observations from PBS, which were combined to form a dataset with over 10,000 observations.

To simplify the analysis, we created one-hot coded keyword records for a pandas dataframe. This involved searching the "content" column of the dataframe for certain keywords related to Trump, Republican, Biden, Democratic, and the presidential election and creating new boolean columns "keyword-trump", "keyword-republican", "keyword-biden", "keyword-democratic", and "keyword-election," respectively. The new columns are set to True if the corresponding keyword appears in the "content" column and False otherwise, providing a more structured and easy-to-use representation of the keyword information.

Additionally, we subset the dataset to allow three sources to have the same amount of data, resulting in a final balanced dataset with 6381 rows, containing 2127 observations from each source.

Frequent keywords in the articles include "Trump", "Biden", and "President" (Figure 1). The frequency of candidates' name are high, but the amount of news contain keyword "Biden" is the least(Figure 2). The are more news have keyword "Democrat" than keyword "Biden". The media seems don't have preference on specific keywords. The distribution of keywords across

Figure 1. Word Cloud of the Whole dataset

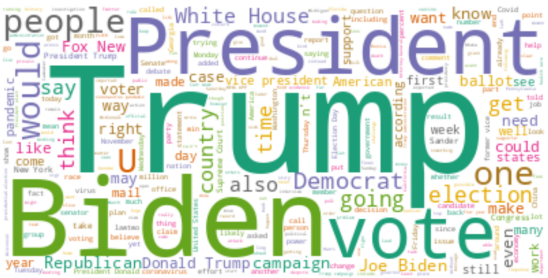


Figure 2. Bar chart of keywords from each source

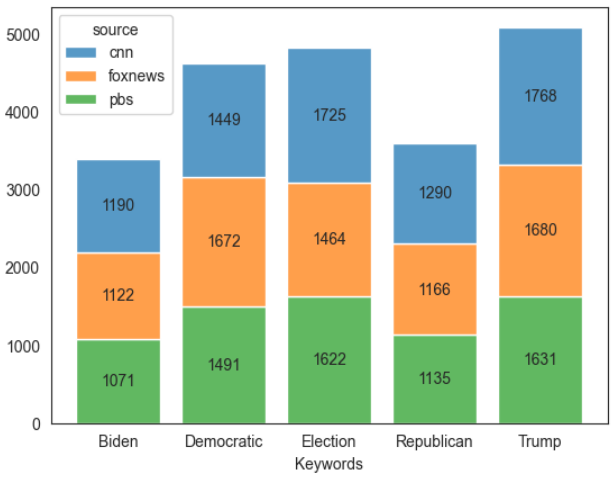


Figure 3. Weekly frequency by Keywords of the whole US dataset

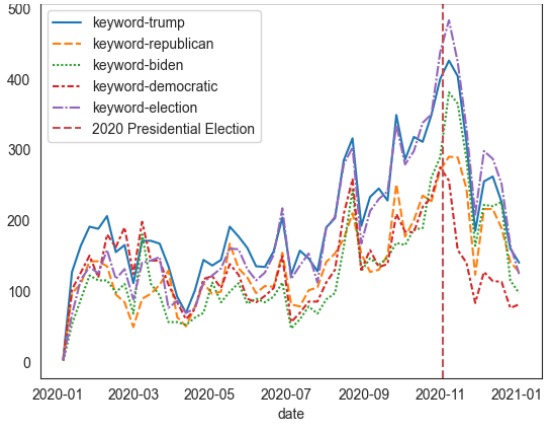


Figure 4. Weekly frequency by keywords of the balanced US dataset

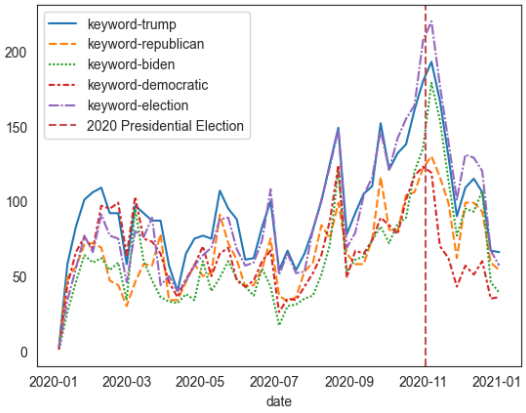


Figure 5. Weekly frequency by sources of the whole US dataset

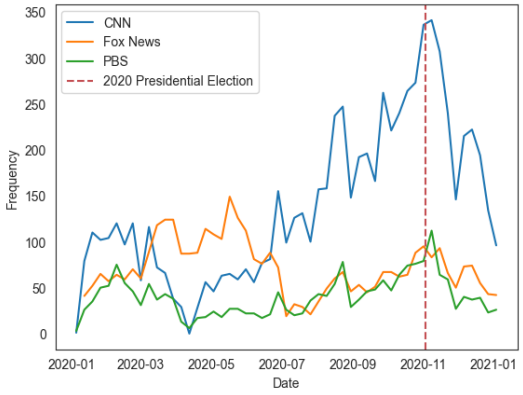
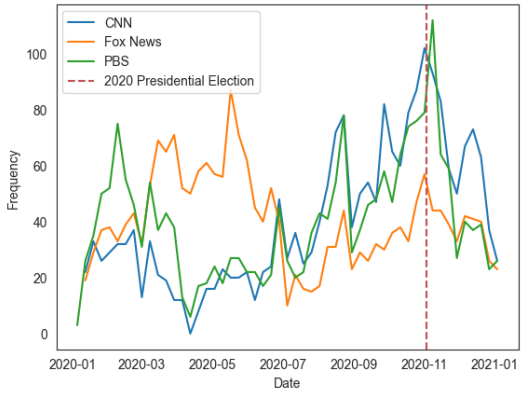


Figure 6. Weekly frequency by sources of the balanced US dataset



time does not show a significant change between the whole dataset and the balanced data (Figure 3 & 4). The keywords related to candidates' names are more popular throughout the whole period, while the names of parties are less used. Furthermore, the weekly frequency by source becomes clearer after subsetting the dataset to a balanced one (Figure 5 & Figure 6). The distribution of frequency of CNN and PBS aligns with the distribution of keywords, gradually increasing and achieving the peak at the week of the election. However, Fox News had the highest amount of election news from March to July, but less news compared to the other two sources.

2. Taiwan Data:

a. Data Collection

The data collection process for this project involved scraping news articles related to election periods in Taiwan from a popular Taiwanese news website. And the Presidential election happened in January 11, 2020. I collected news articles from September 1st, 2019 to January 18, 2020. Python code was used to collect links to news articles, which were then used to obtain the contents of the articles themselves. Multi-thread was utilized in order to improve the efficiency of the data collection process, with a thread pool executor being created to submit tasks and collect futures. The use of threading allowed for multiple requests to be made simultaneously, reducing the amount of time needed to collect the data.

In order to handle missing or null data, the code was designed to filter out any articles that could not be scraped successfully. This was done by checking for null or missing data and dropping those articles from the dataset. Additionally, the code was set up to handle cases where no articles were found on a given page, in which case the data collection process would be terminated. The data collection process was limited to articles from Taiwan only.

b. Summaries of Variables

The dataset contains over 10,000 news articles collected between September 1st, 2019 and January 18, 2020. There are 4269 observations from CTI, 5387 observations from LTN, and 420 observations from CNA and The Reporter.

To prepare the data, I followed the same preprocessing steps as I applied to the US dataset. First, I created one-hot coded keyword records for a pandas dataframe, using keywords that include the names of two major candidates - 韓國瑜 (Han) and 蔡英文 (Tsai) - as well as the parties they represented - 國民黨 (KMT) and 民進黨 (DPP) - and the presidential election. This process provides a more structured and easy-to-use representation of the keyword information. Next, I created new boolean columns that are set to True if the corresponding keyword appears in the "content" column and False otherwise. These columns will be used in the analysis to better understand patterns in the data. Additionally, I subset the dataset to allow three sources to have

Figure 7. Bar chart of keywords from each source

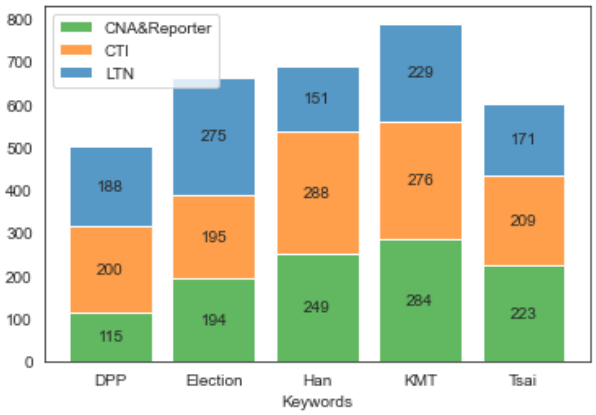


Figure 8. Weekly frequency by sources of the whole Taiwan dataset

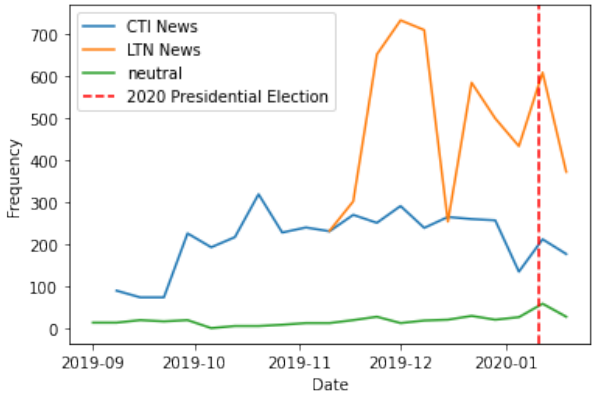


Figure 9. Weekly frequency by Keywords of the whole Taiwan dataset

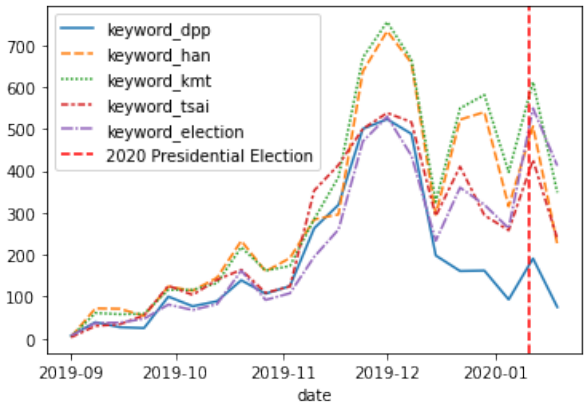
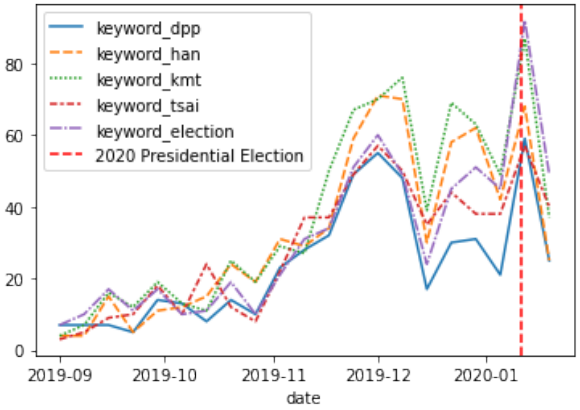


Figure 10. Weekly frequency by keywords of the balanced Taiwan dataset



the same amount of data, making the classification model more balanced. The final balanced dataset contains 1260 rows, with 420 from each source.

The most popular news topic in Taiwan during the study period was found to be the KMT party (Figure 7). One possible reason for this is that the KMT party primary election took a lot of time, involved many candidates, and generated several controversies. Additionally, neutral media sources tended to focus on the election rather than individual candidates. A similar trend was observed in the US, where conservative candidates attracted more attention than their party, and liberal parties received more coverage than their candidates.

The distribution of LTN news follows a non-normal pattern. Additionally, the distribution of keywords over time was significantly impacted by the absence of LTN news before November 2019, as shown in Figure 8. This absence led to a surge in the dataset in

December 2019, as depicted in Figure 9. The balanced dataset peaked on election day (January 11, 2020), as illustrated in Figure 10, which aligns with the expected norm. Despite the limited data, there was a decrease in coverage around a week prior to the election day. This decrease may have been due to election silence regulations that forbid releasing opinion polls starting from 10 days before the election day (Wikipedia, 2021).

3. Limitation:

One potential limitation of this project is the limited number of media sources used in the datasets. The datasets for both the US and Taiwan only include news articles from a small number of selected media sources, which may limit the generalizability of the findings to other media sources. Additionally, the datasets only cover the year 2020, which may not provide a comprehensive understanding of the news coverage leading up to the election.
Furthermore, the keyword selection and one-hot encoding approach may oversimplify the complexity of the news articles and may not capture all relevant information for classification purposes. This could result in some important contextual information being overlooked in the analysis. Additionally, as the data was web-scraped, there may be issues with the reliability and accuracy of the data.

In the case of the Taiwan dataset, there were specific challenges that further limit the scope of the study. The extremely unbalanced data resulting from the lack of an organized dataset and the absence of some period data for one of the news sources (LTN) could limit the robustness of the findings. Moreover, there were only two media sources in Taiwan that could be considered neutral, one funded by a public consortium corporation and the other being a nonprofit media, and their data only accounted for one tenth of the other two sources. Furthermore, the LTN news did not have data before November due to technical issues while scraping news.

Overall, these limitations should be taken into account when interpreting the results of this project, and future studies should consider expanding the number of media sources and exploring other approaches to analyze news coverage leading up to an election.

C. Methodology

The aim of this project is to classify political news articles from different media sources during election periods in Taiwan and the US. To achieve this, the target variable will be the news source, and the features will include the content of the news articles. Preprocessing of the text data will be done by tokenizing, stemming, and removing stop words to extract meaningful features.

As the outcome of interest is categorical, a classification model will be appropriate for this task. Therefore, I plan to use two parametric and two non-parametric techniques, including Multinomial Naive Bayes (MNB), K-Nearest Neighbor (KNN), Decision Tree, and Support Vector Machine (SVM).

Multinomial Naive Bayes (MNB) is a popular choice for text classification tasks as it works well with text data, especially with a high number of features. It is simple to implement, fast, and can handle both numerical and categorical features. MNB is often used for spam filtering, sentiment analysis, and topic classification. Patel and Meehan (2021) conducted a study on detecting fake news on Reddit by developing and comparing supervised learning models with CountVectorizer and Term Frequency-Inverse Document Frequency methods, and found that the CountVectorizer and MultinomialNB model achieved the highest accuracy on the dataset.

Support Vector Machine (SVM) is a powerful classification model that works well with both linear and non-linear datasets. SVM works by finding the hyperplane that maximizes the margin between the classes. SVM has been used in various domains, including finance, bioinformatics, and image classification. Despite from Patel and Meehan mentioned above, SVM is applied on text classification on multiple languages, including Indonesian COVID news classification (Utomo & Prayoga, 2021) and Nepali news classification (Shahi & Pant, 2018). Miao et al,. (2018) designed a Chinese news text classification system model based on machine learning algorithm that includes text pretreatment, text representation, classifier training, and classification, and compared K-nearest Neighbor, Naive Bayes, and Support Vector Machine as classification algorithms, and concluded that the system can achieve satisfactory results.

K-Nearest Neighbor (KNN) is a non-parametric classification model that classifies an unknown data point by finding the K closest data points in the training data and assigning the class based on the majority vote. KNN is simple to implement, interpretable, and works well with small datasets. Chen et al., (2020) applied a KNN-based classification method for Lao news text, which includes preprocessing, feature extraction, adjusting parameters through KNN classifier, data normalization, and data dimensionality reduction to improve classification accuracy.

Decision trees are suitable for classification tasks and are easy to interpret. They have also been applied in various domains, including healthcare and finance. Keskar et al. (2020) proposed a method for detecting fake news on Twitter performing N-gram analysis for feature extraction and applying decision tree machine learning techniques to classify documents as fake or real, in order to address the challenge of detecting fake news with limited resources and datasets.

Overall, using MNB, KNN, SVM, and decision tree models will provide a complementary approach to classify political news articles from different media sources. These models have been proven to be effective in text classification tasks and are suitable for our research question.

D. Bibliography

1. Boxell, L. (2021, October 13). Bias in news coverage during the 2016 US election: New evidence from images. CEPR. Retrieved April 2, 2023, from https://cepr.org/voxeu/columns/bias-news-coverage-during-2016-us-election-new-evi dence-images
2. Hass, R. (2022, March 9). *Democracy, the China Challenge, and the 2020 elections in Taiwan*. Brookings. Retrieved April 2, 2023, from https://www.brookings.edu/opinions/democracy-the-china-challenge-and-the-2020-ele ctions-in-taiwan/
3. Clark C, Tan AC, Ho K (2019). Political Polarization in Taiwan and the United States: A Research Puzzle. Taipei, Taiwan: 2019 TIGCR International Conference on "Political Polarization: Perspectives go Governance and Communication". 25/10/2019-25/10/2019.
4. Horne, Benjamin; Gruppi, Mauricio, 2021, "NELA-GT-2020", https://doi.org/10.7910/DVN/CHMUYZ, Harvard Dataverse, V3
5. CNN. (2020). Election news. Retrieved from https://www.cnn.com/
6. Fox News. (2020). Election news. Retrieved from https://www.foxnews.com/
7. PBS. (2020). Election news. Retrieved from https://www.pbs.org/newshour/
8. LTN News. (2020). Election news. Retrieved from https://www.ltn.com.tw/
9. CTI News. (2020). Election news. Retrieved from https://www.ctitv.com.tw/%E4%B8%AD%E5%A4%A9%E6%96%B0%E8%81%9E
10. 中央通訊社. (2020). Election news. Retrieved April 9, 2023, from https://www.cna.com.tw/
11. 報導者 the reporter. (2020). Election news. Retrieved April 9, 2023, from https://www.twreporter.org/
12. Wikipedia contributors. (2021, September 28). Election silence. In Wikipedia. Retrieved October 6, 2021, from https://en.wikipedia.org/wiki/Election_silence
13. A. Patel and K. Meehan, "Fake News Detection on Reddit Utilising CountVectorizer and Term Frequency-Inverse Document Frequency with Logistic Regression, MultinominalNB and Support Vector Machine," 2021 32nd Irish Signals and Systems

Conference (ISSC), Athlone, Ireland, 2021, pp. 1-6, doi: 10.1109/ISSC52156.2021.9467842.

14. Utomo, W. H., & Prayoga, K. J. (2021). Hoax classification Corona virus (COVID-19) news in Indonesian using the support vector machine (SVM) method. Journal of Computer Science, 17(8), 692–708. https://doi.org/10.3844/jcssp.2021.692.708
15. Shahi, T. B., & Pant, A. K. (2018). Nepali news classification using naïve Bayes, Support Vector Machines and neural networks. 2018 International Conference on Communication Information and Computing Technology (ICCICT). https://doi.org/10.1109/iccict.2018.8325883
16. Miao, F., Zhang, P., Jin, L., & Wu, H. (2018). Chinese news text classification based on machine learning algorithm. 2018 10th International Conference on Intelligent Human-Machine Systems and Cybernetics (IHMSC). https://doi.org/10.1109/ihmsc.2018.10117
17. Chen, Z., Zhou, L. J., Li, X. D., Zhang, J. N., & Huo, W. J. (2020). The Lao Text Classification method based on Knn. Procedia Computer Science, 166, 523–528. https://doi.org/10.1016/j.procs.2020.02.053
18. Keskar, D., Palwe, S., & Gupta, A. (2020). Fake news classification on Twitter using flume, N-gram analysis, and Decision Tree Machine Learning Technique. Proceeding of International Conference on Computational Science and Applications, 139–147. https://doi.org/10.1007/978-981-15-0790-8_15