DSII White Paper

PPOL 565 Data Science II: Applied Statistical Learning

Shiaw-Shiuan (Sharon) Chuang

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

I. EXECUTIVE SUMMARY

This research paper investigates partisan bias in political news reporting during election periods in Taiwan and the United States. The study addresses three key research questions: the extent of conservative and liberal media bias, identification of biased keywords, and a comparison of media bias between the two countries.

The highly polarized political climates in both countries necessitate an examination of news bias's role in shaping public opinion during crucial elections. While political polarization has risen in both the United States and Taiwan, the underlying factors and extent of polarization differ. In Taiwan, polarization centers around national identity and cross-strait relations with China, while in the United States, it focuses on race, immigration, and economic inequality.

The 2020 United States presidential election and Taiwan's 2020 presidential election were significant events characterized by political polarization and contentious issues. In the United States, the election revealed deep political divisions and debates on topics such as immigration, healthcare, the environment, and economic policy. Media bias in news coverage played a role in influencing election outcomes and shaping public opinion. This research offers valuable insights into how media bias can impact public opinion during elections in countries with varying degrees of political polarization.

II. BACKGROUND

The 2020 United States presidential election and Taiwan's 2020 presidential election were both significant events marked by political polarization and charged atmospheres. In the United States, the election exposed deep political divisions and hotly contested issues such as immigration, healthcare, the environment, and economic policy (Time, 2020). Media bias in news coverage also influenced the election outcomes, shaping public opinion (CEPR, 2021).

Similarly, Taiwan's political climate during the 2020 presidential election was highly polarized, strongly influenced by the 2019 protest in Hong Kong and emphasizing the values of

democracy and the China challenge (Hass, 2022). The election in Taiwan witnessed allegations of electoral malpractice and revealed tensions between pro-independence and pro-China forces within the country.

Given the highly polarized political climates in both countries, it becomes essential to examine the role of news bias in shaping public opinion during these pivotal elections. While both the United States and Taiwan have experienced rising political polarization in recent years, the nature and extent of polarization differ. In Taiwan, polarization is largely driven by issues related to national identity and cross-strait relations with China, whereas in the United States, polarization is more focused on issues related to race, immigration, and economic inequality (Clark C, Tan AC, Ho K, 2019). Consequently, understanding the interplay between media bias and the bipolar political status quo is crucial in comprehending the factors that influenced the election outcomes in each country.

Political news media plays a vital role in shaping public opinion and influencing electoral outcomes. However, the presence of partisan bias within political news coverage can have significant consequences for democracy. Partisan bias can distort the perception of events, reinforce existing political polarization, and ultimately impact electoral results. Consequently, examining the role of partisan bias in shaping public opinion during the 2020 presidential elections in both the United States and Taiwan emerges as a crucial area of study.

The 2020 US presidential election, characterized by its significance and polarization, showcased deep political divisions within the country, with media bias playing a role in shaping public opinion (CEPR, 2021). Similarly, Taiwan's 2020 presidential election witnessed political polarization, with democracy and the China challenge being central issues (Hass, 2022). In both cases, allegations of electoral malpractice and a charged atmosphere prevailed. However, the nature and extent of polarization varied between the two countries. Taiwan's polarization revolved largely around issues concerning national identity and cross-strait relations with China, while the United States' polarization focused more on race, immigration, and economic inequality (Clark C, Tan AC, Ho K, 2019). Understanding how media bias and the bipolar political status quo interacted to influence the election results is essential in both contexts.

III.DATA

The objective of this project is to classify news articles from various media sources in the United States and Taiwan during the election period. To achieve this, the project requires news-

based data with labeled sources, titles, release dates, and categories. For the United States, the project utilizes the "NELA-GT-2020" dataset from Harvard Dataverse (Horne, Benjamin; Gruppi, Mauricio, 2021), which contains nearly 1.8 million news articles from 519 sources collected between January 1st, 2020 and December 31st, 2020. I choose to use the election-related news subset and use all articles from CNN (2020), Foxnews(2020), and PBS(2020) in the subset. In the case of Taiwan, a dataset is generated through web scraping political news articles from three media sources: Liberty Times (LTN, 2019-2020), the pro-independence and pro-DPP media; CTI News (2019-2020), the pro-China and pro-KMT media; and a collection of Central News Agency(2019-2020), Public Television Service(2019-2020), and The Reporter(2019-2020) for central position news sources.

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

Political Bias	Media Name	# of observations
Lean Right	FOX	3775
Lean Left	CNN	7327
Centre	PBS	2135

Table 3: News sources of the US dataset

1. US Data:

The "NELA-GT-2020" dataset comprises approximately 1.8 million news articles collected from 519 sources between January 1st, 2020 and December 31st, 2020. Specifically for election-related news, there are 7,663 observations from CNN, 3,926 observations from Foxnews, and 2,280 observations from PBS, which were combined to create a dataset with over 10,000 observations.(Table 3)

To simplify the analysis, a pandas dataframe was utilized to create one-hot coded keyword records. This involved searching the "content" column of the dataframe for specific keywords associated with Trump, Republican, Biden, Democratic, and the presidential election. Boolean columns such as "keyword-trump," "keyword-republican," "keyword-biden," "keyword-democratic," and "keyword-election" were created accordingly. These new columns are set to True if the corresponding keyword appears in the "content" column and False otherwise, providing a more structured and user-friendly representation of the keyword information.

Additionally, the dataset was subset to ensure an equal amount of data for three sources, resulting in a final balanced dataset with 6,381 rows. Each source contributes 2,127 observations to this balanced dataset.

Figure 1. Word Cloud of the Whole dataset

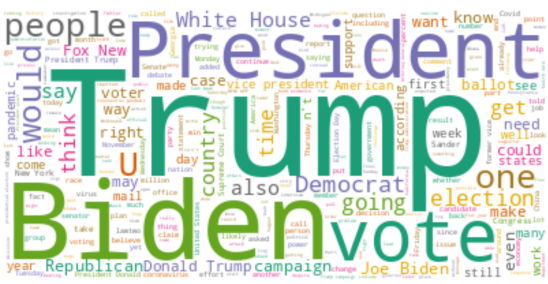


Figure 2. Bar chart of keywords from each source

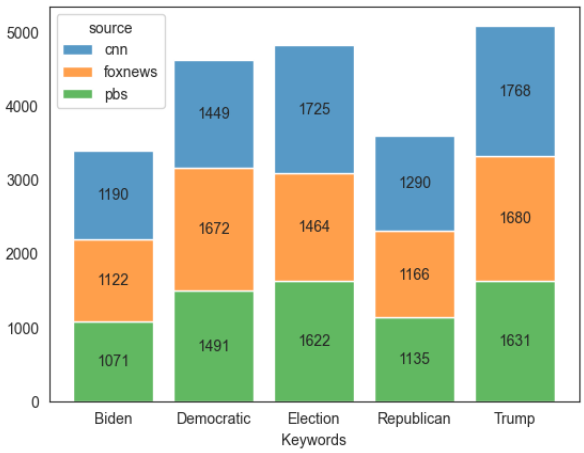


Figure 3. Weekly frequency by Keywords of the whole US dataset

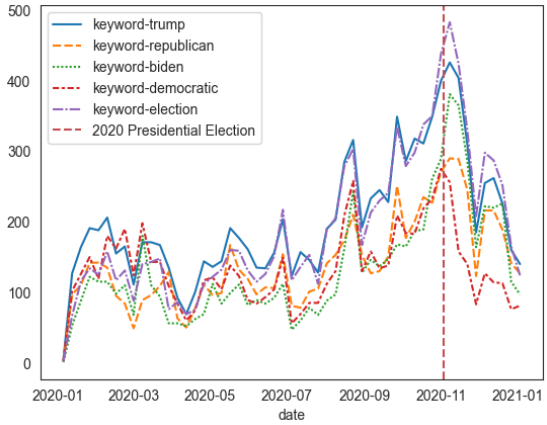


Figure 4. Weekly frequency by keywords of the balanced US dataset

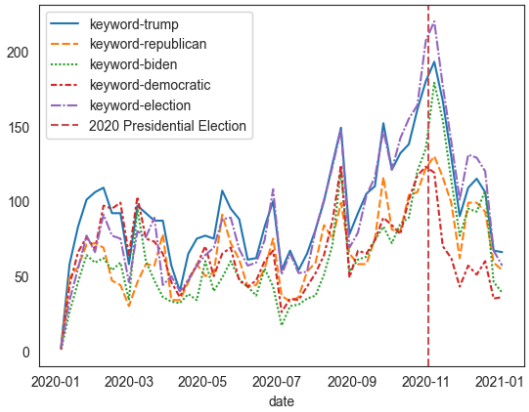


Figure 5. Weekly frequency by sources of the whole US dataset

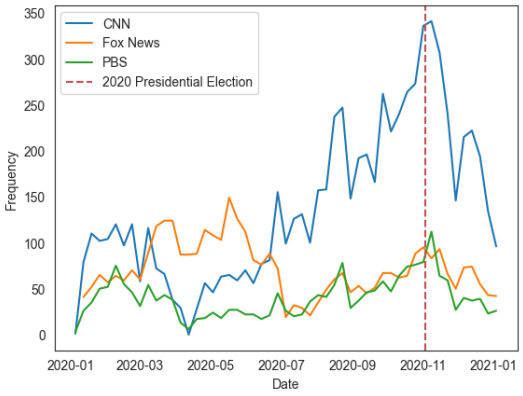
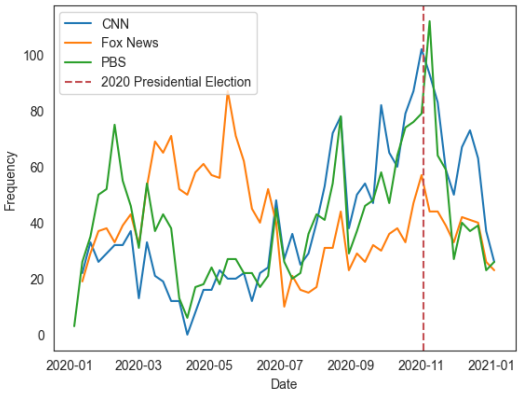


Figure 6. Weekly frequency by sources of the balanced US dataset



The articles frequently feature keywords such as "Trump," "Biden," and "President" (Figure 1). Although the candidates' names have a high frequency, the number of news articles containing the keyword "Biden" is the lowest (Figure 2). There are more news articles with the keyword "Democrat" compared to "Biden." The media appears to have no specific preference for particular keywords. The distribution of keywords across time does not exhibit significant changes between the entire dataset and the balanced dataset (Figure 3 & 4). Throughout the entire period, keywords related to candidates' names remain popular, while party names are less frequently used. Moreover, after subsetting the dataset to a balanced one, the weekly frequency by source becomes more apparent (Figure 5 & Figure 6). The distribution of frequency for CNN and PBS aligns with the distribution of keywords, gradually increasing and reaching its peak during the week of the election. However, Fox News had the highest amount of election news from March to July, albeit fewer news articles compared to the other two sources.

2. Taiwan Data:

To get data for this project, news articles related to the election period in Taiwan were scraped from a popular Taiwanese news website. The presidential election took place on January 11, 2020, and news articles were collected from September 1, 2019, to January 18, 2020. I choose CTI news and LTN news as the biased media sources. The major shareholders of CTI, Tsai Eng-meng (蔡衍明) have good relations with the Chinese government (HÉRAIT, 2022), and the founder of LTN, Lin Rong-san (林榮三) is thought to take a Pan Green (DPP) pro-independence political stance. For the centre position, I choose three sources, The Central News Agency (CNA) is a semi-official news media, Public Television Service (PTS) is an independent public broadcasting institution, and The Reporter is an independent non-profit digital media.

Summaries of Variables

The dataset comprises over 10,000 news articles gathered between September 1, 2019, and January 18, 2020. It includes 4269 observations from CTI, 6387 observations from LTN, and 962 observations from CNA, PTS, and The Reporter (Table 4).

To prepare the data, the same preprocessing steps as those applied to the US dataset were followed. Initially, one-hot coded keyword records were created for a pandas dataframe, incorporating keywords such as the names of the two major candidates (韓國瑜 - Han and 蔡英文 - Tsai), the parties they represented (國民黨 - KMT and 民進黨 - DPP), and the presidential election. This process facilitated a structured and easily usable representation of the keyword information.

Subsequently, new boolean columns were generated, set to True if the corresponding keyword appeared in the "content" column and False otherwise. These columns were employed

Political Position	Media Name	# of observations	
Pro-China (Pro-KMT)	CTI News	4269	
Pro-Independent (Pro-DPP)	LTN News	6387	
Centre	PTS (公共電視)	144	962
	CNA (中央社)	702	
	The Reporter (報導者)	116	

Table 4: News sources of the Taiwan dataset

in the analysis to identify patterns in the data. Additionally, the dataset was subsetted to ensure an equal number of data samples from the three sources, thus achieving a more balanced classification model. The final balanced dataset consisted of 2,700 rows, with 900 rows from each source.

During the study period, the most prominent news topic in Taiwan was found to be the KMT party. This could be attributed to the KMT party primary election, which involved multiple candidates and generated controversies, taking up considerable time. Moreover, neutral media sources tended to focus on the election as a whole rather than individual candidates. A similar trend was observed in the US, where conservative candidates garnered more attention than their respective parties, while liberal parties received more coverage than their candidates.

The distribution of LTN news exhibited a non-normal pattern. Additionally, the distribution of keywords over time was significantly affected by the absence of LTN news before November 2019 (Figure 8) . This absence resulted in a surge in the dataset in December 2019

Figure 7. Bar chart of keywords from each source w/balanced dataset

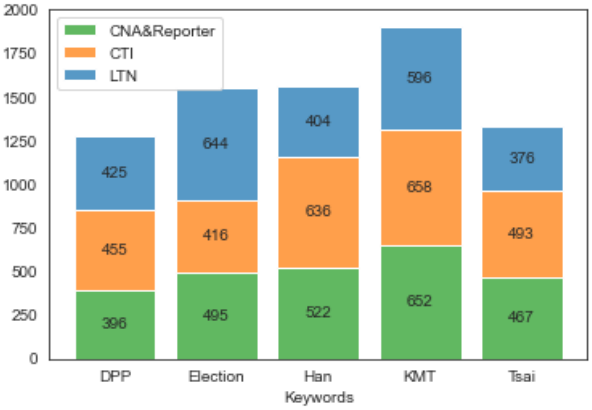


Figure 8. Weekly frequency by sources of the whole Taiwan dataset

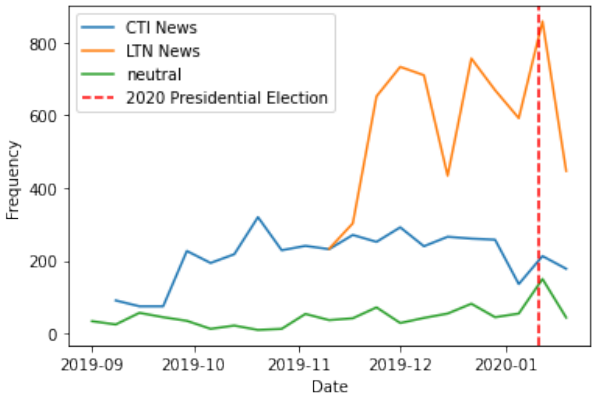


Figure 9. Weekly frequency by keywords w/ the balanced Taiwan dataset

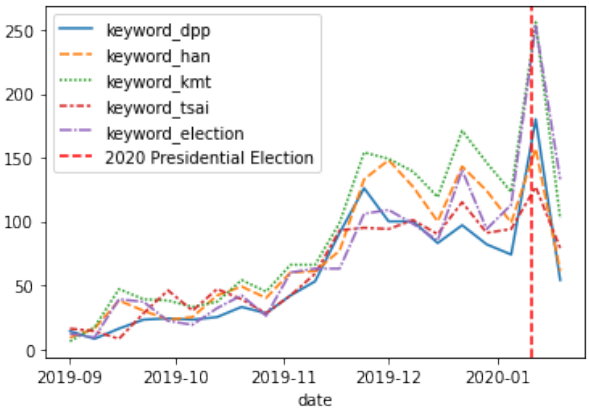
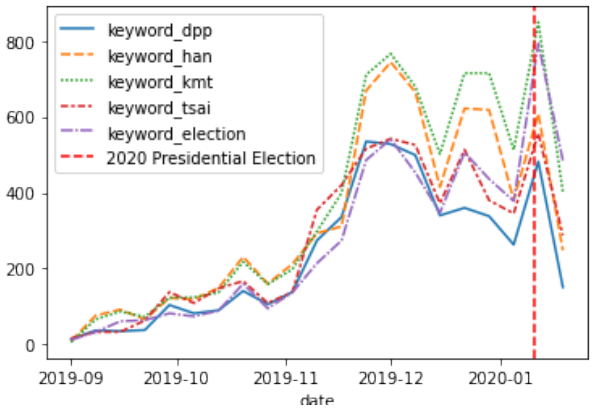


Figure 10. Weekly frequency by Keywords w/ the whole Taiwan dataset



(Figure 10). The balanced dataset reached its peak on the day of the election (January 11, 2020), (Figure 9), aligning with expectations. Despite the limited data, there was a decrease in coverage approximately one week before the election day. This decline could be attributed to election silence regulations that prohibit the release of opinion polls starting from ten days before the election day (Wikipedia, 2021).

IV. METHODOLOGY

The project mainly conducted prediction analysis. The methodology employed in this study aims to provide a comprehensive understanding of news bias during the presidential election in the United States and Taiwan, utilizing quantitative methods. The research questions focus on identifying the extent of bias exhibited by conservative and liberal media outlets, determining the most likely biased keywords in news coverage, and comparing media bias between Taiwan and the United States.

To classify the political news articles, a combination of parametric and non-parametric classification models will be employed. The selected models for this study include Multinomial Naive Bayes (MNB), K-Nearest Neighbor (KNN), Support Vector Machine (SVM), and Decision Tree.

Multinomial Naive Bayes (MNB) is well-suited for text classification tasks, as it handles a high number of features and can accommodate both numerical and categorical data. MNB has been successfully used in various domains, including spam filtering and sentiment analysis.

Support Vector Machine (SVM) is a powerful classification model that works well with linear and non-linear datasets. It aims to find the hyperplane that maximizes the margin between classes. SVM has demonstrated effectiveness in text classification tasks and has been applied in domains such as finance and bioinformatics.

K-Nearest Neighbor (KNN) is a non-parametric classification model that assigns a class to an unknown data point based on the majority vote of its nearest neighbors. KNN is simple to implement and interpret, making it suitable for small datasets.

Decision Tree is a classification model that utilizes a hierarchical structure of decision nodes and leaf nodes to classify instances. It is known for its interpretability and has been applied in various domains, including healthcare and finance.

Through the application of these methodologies, this study aims to provide valuable insights into the presence and extent of news bias during the presidential election in the United

States and Taiwan. By utilizing a combination of classification models, statistical analysis, and data preprocessing techniques, the research will contribute to a deeper understanding of the complex dynamics of media bias and its impact on public perceptions in polarized societies.

V. ANALYSIS

For both datasets, I divide the data into different kinds of subsets, including groups based on different keywords. The classification models are also trained using articles from only two specific news sources. This approach allows us to analyze which keywords are more heavily biased or which news media are more different from each other by comparing the accuracy of the models trained from different subsets.

A. Results of US models

From the accuracy results presented in Table 5, it is evident that decision tree models consistently outperformed other models in terms of classification accuracy. Across all decision tree models, the accuracy exceeded 70%, indicating their effectiveness in accurately classifying news articles. Notably, the models that compared only two classes achieved even higher accuracy rates, reaching approximately 90%. This suggests that distinguishing between specific pairs of media sources was relatively easier for the classifiers.

	KNN	MNB	SVM	Decision Tree
Whole dataset	68.50%	55.14%	68.58%	77.45%
Undersampled dataset	55.04%	72.05%	66.66%	70.88%
keyword democratic	67.27%	54.72%	69.53%	77.68%
keyword Biden	67.73%	59.15%	69.79%	79.50%
keyword republican	70.84%	58.31%	68.23%	80.02%
keyword Trump	70.07%	57.51%	69.22%	80.94%
keyword election	69.72%	59.77%	68.72%	77.97%
CNN vs FOX	81.63%	68.35%	82.04%	95.14%
PBS vs CNN	81.46%	77.65%	81.04%	88.54%
PBS vs FOX	75.30%	66.41%	74.45%	92.56%

Table 5: Accuracy Results of Prediction Models w/ The US news articles

Figure 11. Average accuracy of classifier for different media subset in in the United States

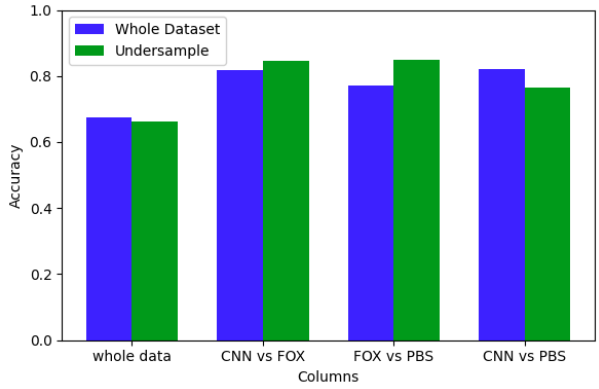
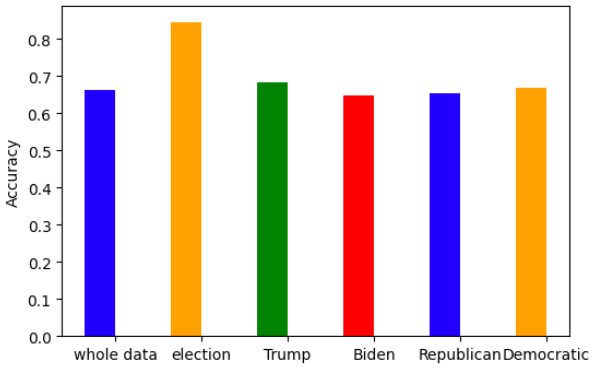


Figure 12. Average accuracy of classifier for different keywords subset in the United States



When examining the models that compared two selected media sources, it is evident that the classifier for FOX and CNN demonstrated high accuracy in both the fall dataset and the balanced under-sample. This implies that the model was successful in identifying distinct patterns and characteristics associated with these particular media sources. The high accuracy of these models indicates the presence of notable differences in news reporting between FOX and CNN.

Among the models focused on specific keywords, the election model stood out with the best performance. This finding suggests that discussions related to elections tend to exhibit a higher degree of bias in news articles. It is likely that media coverage during election periods is influenced by political affiliations and interests, leading to more pronounced biases in reporting.

These accuracy results shed light on the prevalence of biased news reporting when it comes to election-related topics. The high accuracy of the decision tree models, especially in comparing two media sources, highlights the distinctiveness of these sources and their unique biases. Furthermore, the performance of the election keyword model reinforces the notion that political events and processes can be particularly susceptible to biased reporting.

B. Results of Taiwan models

The analysis of the Taiwan models revealed interesting findings. Similar to the results obtained for the US datasets, the decision tree model performed exceptionally well in classifying news articles, achieving an accuracy rate of nearly 80%. This highlights the effectiveness of decision tree models in capturing patterns and making accurate predictions.

However, when comparing the models for different news sources in Taiwan, distinct patterns emerged. Contrary to the US model, where the classifier that distinguished between two biased media sources had the highest accuracy, the Taiwan model exhibited a different trend

(Figure 13). The model that classified central news articles demonstrated a relatively larger difference in accuracy when compared to the biased media sources.

Figure 14 illustrates this disparity, showing that the accuracy of the central news model differs significantly from that of the biased media models. This suggests that the classifier was able to distinguish central news articles more effectively, potentially due to distinctive language or content characteristics associated with central news sources. On the other hand, the accuracy of the model classifying between two biased media sources was not the highest, indicating that differentiating between specific biased sources might be more challenging.

	KNN	MNB	SVM	Decision Tree
Whole dataset	69.36%	71.30%	75.59%	79.12%
Undersampled dataset	60.90%	62.22%	68.59%	81.28%
keyword KMT	68.88%	62.62%	67.01%	81.21%
keyword Han	71.17%	72.63%	71.09%	82.39%
keyword DPP	69.36%	71.30%	75.59%	79.16%
keyword Tsai	70.62%	64.04%	66.81%	79.03%
keyword election	66.78%	59.44%	68.27%	78.41%
CTI vs LTN	72.56%	68.39%	67.92%	83.49%
Central vs CTI	86.25%	83.00%	97.13%	95.33%
Central vs LTN	89.66%	87.55%	97.41%	95.31%

Table 5: News sources of the Taiwan dataset

Figure 8. Average accuracy of classifier for different data subset in Taiwan

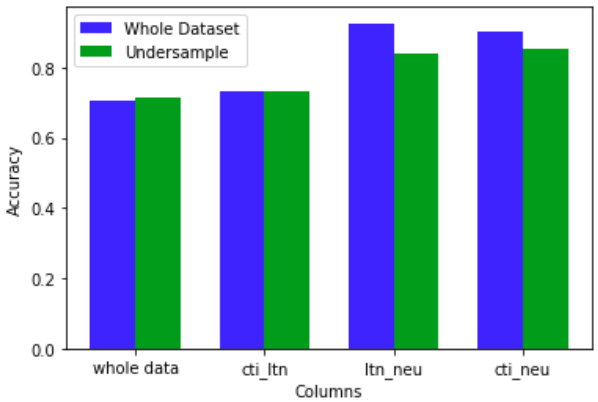
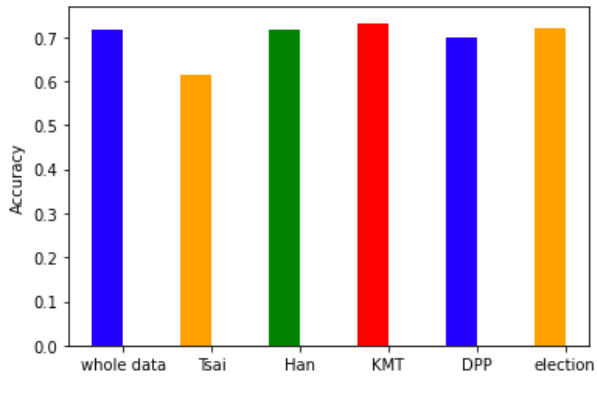


Figure 8. Average accuracy of classifier for different keywords subset in Taiwan



Moreover, the keyword model exhibited similar trends to the model trained with the full dataset. However, the accuracy of the model focusing on the keyword "Tsai" was lower compared to the accuracy of the other models. This suggests that news articles containing the keyword "Tsai" may be less biased. This result is different from what I expected, after all, Tsai Ing-wen is the president of the time.

VI. CONCLUSION

In conclusion, this research examined partisan bias in political news reporting during election periods in Taiwan and the United States using quantitative methods. The study aimed to answer three main research questions: the extent of conservative and liberal media bias, the identification of biased keywords, and a comparison of media bias between the two countries.

The findings revealed an interesting difference in the accuracy of the classification models between Taiwan and the United States. In Taiwan, the models performed well in distinguishing between news articles from central and biased media sources but struggled to accurately classify articles between two biased sources. On the other hand, in the United States, the models achieved the highest accuracy when categorizing articles from two biased media sources.

One possible explanation for this divergence is the variation in language and content used by media sources in the two countries. It is likely that Taiwan's biased articles contain more emotional and intense words, reflecting a polarized media landscape. In contrast, central media sources in Taiwan may use more neutral language to maintain a balanced perspective. This linguistic contrast could explain the higher accuracy in classifying central and biased media sources in Taiwan.

These findings have important implications. They highlight the complex nature of media bias and its manifestation in different socio-political contexts. The identification of biased keywords provides valuable insights into language patterns used in news coverage during elections, aiding in the evaluation of news sources and public perception.

This research contributes to a better understanding of partisan bias in political news reporting. By using quantitative methods and classification models, the study offers insights into the challenges of classifying news articles and the impact of language and content choices. These findings have implications for media literacy, public perception, and critical evaluation of news sources in polarized societies.

Future research could explore additional factors influencing media bias, such as the role of social media platforms and the influence of political ideologies. Furthermore, studying the effects of biased news coverage on public opinion and democratic processes would provide further insights into the consequences of media bias.

Ultimately, this study aims to promote informed public discourse, media transparency, and a nuanced understanding of media bias. By unraveling the complexities of partisan bias, we can strive towards a more balanced and informed democratic process, encouraging critical thinking and facilitating the exchange of ideas.

VII.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. (2019 - 2020). Election news. Retrieved from https://www.ltn.com.tw/
9. CTI News. (2019 - 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. (2019 - 2020). Election news. Retrieved April 9, 2023, from https:// www.cna.com.tw/
11. PTS News. (2019 - 2020). Election news. Retrieved from https://news.pts.org.tw/
12. 報導者 the reporter. (2019 - 2020). Election news. Retrieved April 9, 2023, from https:// www.twreporter.org/
13. Wikipedia contributors. (2021, September 28). Election silence. In Wikipedia. Retrieved October 6, 2021, from https://en.wikipedia.org/wiki/Election_silence
14. HÉRAIT, A. (2022, July 20). In Taiwan, freedom of expression faces Chinese disinformation. PRESS FREEDOM. https://ijnet.org/en/story/taiwan-freedom-expression-faces-chinese-disinformation

I. IMPLEMENTATION APPENDIX

A. Data Collection

For the US dataset, existing data was utilized for this project. However, for the Taiwanese data, a Python code was developed to collect links to news articles, which were then used to extract the article contents. The data collection process incorporated multi-threading to improve efficiency, employing a thread pool executor to submit tasks and collect results. By utilizing multi-threading, concurrent requests were made, significantly reducing the data collection time.

To handle missing or null data, the code implemented a filtering mechanism to remove articles that couldn't be successfully scraped. This involved checking for null or missing data and excluding those articles from the dataset. Additionally, the code was designed to handle situations where no articles were found on a particular page, thereby terminating the data collection process. The data collection focused exclusively on articles from Taiwan.

B. Data Preprocessing:

Prior to the classification task, the news articles underwent preprocessing to transform the raw text into a suitable format for analysis. The following key steps were involved in the data preprocessing stage.

Stopword Removal

To enhance the quality of the textual data, a set of stopwords was employed to filter out common and non-informative words. Initially, widely used English stopwords from the NLTK library were incorporated using the 'stopwords.words('english')' function. Additionally, a custom set of domain-specific stopwords was created for the US data, including terms such as 'cnn,' 'fox,' and 'news.' This extended set was converted into a list and combined with the initial stopwords.

Vectorization

The next crucial step was transforming the raw text data into a numerical representation using the Term Frequency-Inverse Document Frequency (TF-IDF) scheme. This was accomplished using the 'TfidfVectorizer' class from the scikit-learn library. The 'TfidfVectorizer' was initialized with the 'stopwordset' as the stop words parameter. The 'fit_transform' method of the 'TfidfVectorizer' was then applied to the 'x' data, performing tokenization, stopword removal, and calculating TF-IDF weights for each term in the text. This resulted in a transformed feature matrix.

By completing these preprocessing steps, the dataset was refined and prepared for training the classification models. The textual data was converted into numerical features using TF-IDF vectorization, irrelevant stopwords were eliminated, and the dataset was divided into separate training and testing subsets. This allowed for subsequent analysis and evaluation of the models' performance.