How China's Narratives toward Taiwan Evolve: A Topic Modeling and Linkage Network Analysis of Newspaper Coverage [PFO]

Charity S. Jacobs

CASOS, Institute for Software Research
Carnegie Mellon University, Pittsburgh, PA 15213
csking@andrew.cmu.edu

Abstract. This white paper presents a novel analysis of how China's newspaper narratives towards Taiwan have evolved, using topic modeling and co-occurring linkage networks. By applying these techniques to a corpus of newspaper articles, we identify key topics and themes in China's discourse towards Taiwan, as well as the relationships between these topics and how they change over time. Our analysis reveals both static topic themes in addition to geo-political shifts in China's narratives around Taiwan, reflecting changes in political, economic, and cultural factors. This paper contributes to a better understanding of the dynamics of China-Taiwan relations and provides insights into how China's discourse towards Taiwan has evolved in response to changing circumstances.

Keywords: China · Global Times · Taiwan · Network Analysis

1 Introduction

As China's influence on the international stage continues to expand, its efforts to control the narrative through media outreach have intensified. However, this effort contrasts with the traditional notion of independent media, which is crucial to democracy and the checks on government power. China has developed its own brand of democracy, dubbed "Chinese characteristics," which prioritizes impact over the process, but a free press does not fit within this model [23].

To address what it perceives as biased coverage of China by Western media, China launched its "Right to Speak" campaign in the early 2000s. This effort involved significant funding and organizational changes to create international media companies such as Xinhua and the Chinese Global Television Network, and to expand existing outlets like the Global Times. The objective was to produce English-language content that could project China's national interests through a "discourse war" [6]. Under Xi Jinping's leadership, the Chinese Communist Party has tightened its control over state-affiliated media outlets. During a visit to key outlets, Xi emphasized that "the media run by the party and the government are the propaganda fronts and must have the party as their family name" [10]. China's centralized power and authoritarian government structure have enabled it to inform and manage information campaigns more efficiently than democratic countries [4]. We adopt the definition of propaganda proposed by philosopher Jason Stanley, which is the use of political, economic, aesthetic, or rational ideals to prop up or erode some ideal for political purposes [21].

The growing global influence of China has significantly impacted the geopolitical landscape of the international community. A balance of military tension and cooperation characterizes China's relationships with neighboring countries in Asia. Among China's geopolitical goals in the region, Taiwan occupies a central position due to China's goals of reunification. China maintains a view of sovereignty over Taiwan, despite Taiwan's complete autonomy and independently elected government. This policy also provides context for why China considers reunification with Taiwan, whether peaceful or otherwise, an "internal issue" [16]. China's stance on any foreign governments legitimizing Taiwan as an autonomous country has become increasingly hostile, characterized by threats of countermeasures and an escalation in show of force, such as China's drills around Taiwan following U.S. House Speaker Nancy Pelosi's visit to Taiwan in 2022 [24]. Taiwan has become China's litmus test for strong geopolitical ties with other countries, and a point of contention with rivals.

China's state-sponsored media plays a crucial role in shaping its foreign policy narrative, particularly regarding its neighboring countries. Analyzing China's state-sponsored news articles is critical to gaining insights into China's foreign policy, its strategic priorities, and the potential implications for the international community. In this paper, we aim to use topic modeling and linkage networks to analyze China's state-sponsored propaganda news articles to shed light on how China portrays its neighboring countries, focusing on its discourse toward Taiwan and how it has shifted over time. Due to China's growing militarism towards Taiwan, this analysis provides novel insight into the shifting of priorities and narratives over the last 15 years. We seek to answer how China's key themes or topics are connected in regards to Taiwan, and how this network structure changes over time.

2 Literature Review

Previous research has highlighted the significance of media and communication in shaping foreign policy, particularly in authoritarian regimes [20]. China's ability to extend its soft power influence at least domestically through its media outlets has worked well to reinforce messaging while placating the public with a sense of comprehensive news reporting [22]. In recent years, there has been a growing interest in analyzing China's media strategy towards Taiwan [25].

Topic modeling has emerged as a popular method for analyzing large corpora of text data and identifying latent topics and themes [3]. Previous studies have used topic modeling to analyze China's media coverage of various issues, including domestic COVID-19 coverage, political reforms, and generative text prediction [14] [13]. Co-occurring linkage network analysis has also been used to identify the relationships between topics and how they evolve over time [2]

[17]. However, to the best of our knowledge, no previous studies have used topic modeling and co-occurring linkage network analysis to analyze China's state-sponsored media coverage of Taiwan.

Overall, the literature suggests that China's media strategy towards Taiwan is an important aspect of China's foreign policy, and analyzing state-sponsored news articles can provide insights into China's strategic priorities and potential implications for the international community. Past research on this particular corpus compared the sentiment of the Ministry of Foreign Affairs data with the state-affiliated newspaper articles and found linguistic differences but did not examine the content of the articles[9]. By using topic modeling and co-occurring linkage network analysis, this paper aims to contribute to the existing literature by providing an empirical analysis of how topics are related within China's state-sponsored media coverage of Taiwan and how those topics change over time.

3 Methods

Data We exported our dataset from the FOCUSdata project at the National Security Studies Department at New Jersey City University, a collaborative project with the Rutgers University Center for Critical Intelligence Studies under a grant from the U.S. Office of the Director of National Intelligence. Due to a data scarcity of the Ministry of Foreign Affairs' official press releases, we decided to use the Global Times corpus, which consists of 677,532 English-language articles from the Chinese state-affiliated Global Times outlet [8]. This dataset spans from April 9, 2009, to December 31, 2022. Additionally, we chose not to include the People's Daily News which consisted of over 500,000 articles, to avoid capturing linguistic differences between the two news outlets.

Unsupervised Topic Modeling using LDA We use the Latent Dirichlet allocation (LDA) topic modeling algorithm to capture the semantic context of our data. LDA is a simple yet powerful tool for uncovering latent variables in observed data [3]. These probabilistic models assume that a dataset comprises some number of topics and that each word in each document is mapped to a topic, effectively extracting word and phrase patterns and clustering them to best describe document groupings.

We filtered the articles that contain the word *Taiwan* and checked to ensure there were at least 1,000 articles per year. Two years had approximately 700 articles each. We used multi-year binning of the data to mitigate the skewness of data across certain years. We then filtered the news articles to remove stop-words, punctuation and artifacts from the scraping process. We used the Python open-source library Gensim to generate a 20-topic model across all years [18], which we then manually examined. We initially removed the top 300 words, but reverted back to the standard stop list to improve topic quality.

Linkage Networks After training a topic model, we can generate the *linkage* between topics, where a link indicates the extent that two topics co-appear within a document, weighted by the joint probability or how often two topics would co-appear by chance [17]. This linkage is called point-wise mutual information (PMI), and is a measure in information theory to quantify the association or dependency of two events [5][15]. PMI can be defined as the following:

$$PMI(x, y) = \log_2 \left(\frac{P(x, y)}{P(x) \cdot P(y)} \right)$$

We divided the data into three periods of time, where Period 1 includes the years 2009-2013, Period 2 covers 2014-2018, and Period 3 covers 2019-2022. Using our trained topic model, for each period of time, we extracted the topic distributions for each document within a given timeframe and computed the linkage between topics i and j for each period, where each period of time contains an adjacency matrix of PMI values between all pairwise topics. Lastly, we use ORA-Pro to visualize the topic linkage networks [1], fixing the nodes across all periods in place to analyze how PMI values between topics change over time. We then analyze the resulting linkage networks of all links in the top 25% of PMI values, where Period 1 had a threshold of .383 bits, Period 2 has a threshold of .277 bits, and Period 3 has a threshold of .249 bits. Our code used for processing, model generation, and analysis is available in the following github repository [12].

4 Results

An initial overview of resultant topics in Table 1 indicates there is a wide spectrum of topics co-associated with discussion on Taiwan to include other neighboring countries such as Japan, South Korea, and Singapore, topics related to public relations and photography, COVID-19 related topics and a multitude of foreign relations topics. Of note, Topic 5 has high co-occurrence around themes of defense and the United States, while Topic 14 maintains key themes of cooperation and Russia.

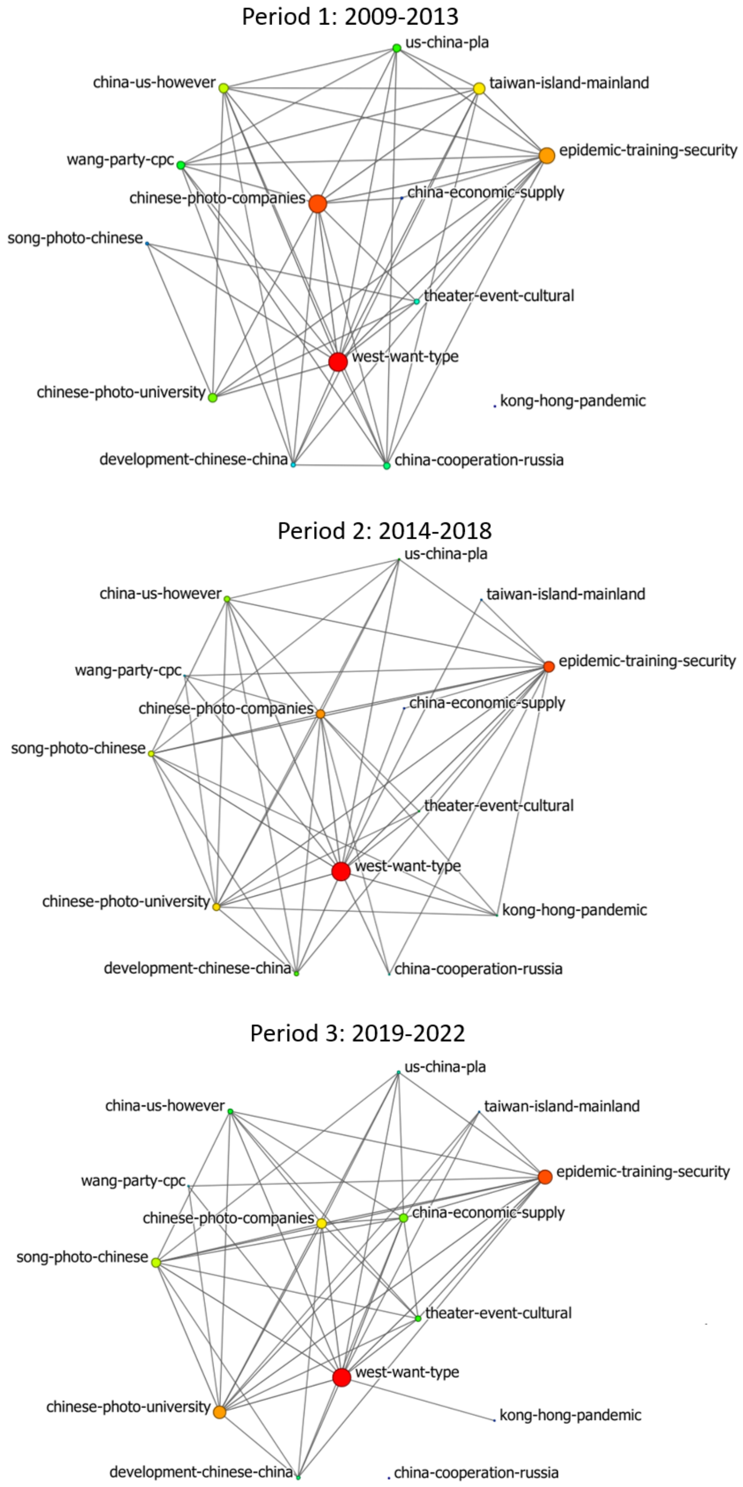
4.1 Linkage Changes

Our linkage networks display the most common co-occurring topics in Figure 1. For labeling, we use the top three weighted words underlying a topic to provide more context on the latent themes for a given topic. For all three periods, Topics 8 (arab-opposes-qin), 17 (south-korea-asia), 11 (australia-australian-myanmar), 12 (online-liu-yuan), and 16 (tsai-chen-election) are absent from our linkage networks due to not making the PMI threshold.

Topic 18 (west-want-type) remains the largest topic (16,194 contributing articles) with the most connections to other topics across time. However, we characterize this topic as an entertainment cluster. It is likely the most connected because the themes are widespread and contain sports, movies, and cultural themes. Some article headlines include “Jackie Chan’s remarks backfire", “Karaoke goes

Table 1: Topic Modeling Output which contains topics and the highest weighted words for each topic, based on the frequency of the word in documents linked to a given topic.

Topic	**Underlying Words**
0	china, us, however, chinese, europe, european, times, cannot, anti, toward
1	song, photo, chinese, cui, tv, show, vaccine, aerial, video, shows
2	japan, china, japanese, war, minister, chinese, pacific, prime, islands, tokyo
3	kong, hong, pandemic, macao, travel, xia, special, mainland, administrative, tourism
4	chinese, photo, companies, times, business, industrial, industry, company, cn, announced
5	us, china, pla, defense, washington, chinese, war, aircraft, question, sea
6	epidemic, training, security, china, times, public, observers, law, health, control
7	development, chinese, china, politicians, democracy, social, economic, public, leadership, rights
8	arab, opposes, qin, significantly, vehicles, gains, rose, suspension, dollars, divisions
9	theater, event, cultural, museum, beijing, shanghai, da, china, yu, province
10	china, economic, supply, india, trade, economy, us, industry, data, largest
11	australia, australian, myanmar, poll, pork, nepal, transparent, minority, advances, endeavors
12	online, liu, yuan, internet, accounts, cn, platform, video, manipulation, qian
13	pakistan, taliban, afghanistan, islamabad, militants, khan, india, security, military, peace
14	china, cooperation, russia, foreign, chinese, development, diplomatic, economic, trade, security
15	wang, party, cpc, affairs, principle, taiwan, committee, chinese, central
16	tsai, chen, election, wu, apple, funeral, leader, court, supreme, dozen
17	south, korea, asia, korean, pacific, asian, singapore, north, indonesia, thailand
18	west, want, type, see, western, beijing, times, right, get, think
19	chinese, photo, university, author, times, life, young, students, family, school



Period 1: 2009-2013

Period 2: 2014-2018

Period 3: 2019-2022

Fig. 1: Linkage Networks over time. Nodes are topics connected by pointwise mutual information links that measure the co-occurrence of topics within a given document. Nodes are sized and colored by their Total-Degree Centrality within the network, measuring how frequently they are connected to other topics.

on the record in China", and "China's Cinderella story". The prominence of this topic and its connectedness also underlines the close cultural ties that China and Taiwan share.

Topic 0 (china-us-however) is a constant and stable topic.s We use a probability threshold greater than .35 to examine key documents for a given topic. This topic contained geopolitical themes of the west Versus China, represented by article titles such as "Can China, US avoid tragedy of great power politics?", "US malice to fail to drive wedge between China and Europe", and "How should we view rising tough voices against China in Europe: Global Times editorial?"

Topic 2 (japan-china-japanese) represents a wholly anti-Japanese cluster revolving mainly around territorial and historical claims. This might actually be one of the more consistent topics highlighting the fact that China really does not like Japan. Notable headlines extracted with a probability $>$ .5: "China lashes Japanese attempt to mark seizure of Diaoyu Islands", "History proves Diaoyu Islands are China's territory", "Japan's theft of Diaoyu Islands tramples on anti-fascist victory: expert", "Japanese fabrications cannot deny history".

Topic 3 (kong-hong-pandemic) is primarily a tourism topic regarding Taiwan and Macau. Topic 4 (chinese-photo-companies) discusses business themes and trade deals.

Topic 5 (us-china-pla) represents another geo-politically adversarial cluster. We increased the document probability to $>$ .55 to find the highest probability documents. All documents at this threshold are about US arms sales to Taiwan or military operations by either the US or China. Article titles include "US arms sale to Taiwan in violation of commitment to Beijing", "China urged to expand nuclear arsenal to deter US warmongers", and "PLA expels US warship trespassing South China Sea."

Topic 6 (epidemic-training-security) covers natural disasters, logistic disasters, and health crises (to include the COVID-19 pandemic).

We can potentially see China's decreased discussion of cooperation as a whole when discussing Taiwan by analyzing 14 (china-cooperation-russia). This topic encompasses strategic partnerships and meetings with other countries. Top headlines using a probability threshold greater than .5 include: "China vows to advance military cooperation with Uzbekistan", "State Councilor vows to advance Sino-African strategic partnership", "China, Pakistan pledge to strengthen ties."

4.2 Betweenness Centrality: Key Facilitating Topics

For each period's linkage network, we analyzed how the shortest path connecting topics shifts by analyzing the topic with the highest betweenness centrality which is at least one standard deviation above the mean of all other topics within a linkage network. A topic with high betweenness centrality indicates that it serves as a connector between topics and groups of topics. Within the first period, the topic with highest betweenness centrality is Topic 4 (chinese-photo-companies). By Period 2, Topic 6 (epidemic-training-security) contains the highest centrality. By Period 3, Topic 5 (us-china-pla) has the highest betweenness centrality.

This shift in key topics that have the highest betweenness centrality indicates which topics are key bridging nodes within the topic landscape for a given period. During Period 1, Topic 4 was the primary linking topic, which focuses on business and trade deals. By Period 2, Topic 6 which entails articles on the COVID-19 pandemic, was the primary facilitator. Most recently, we see Topic 5, which covers the adversarial relationship between the United States and China over Taiwan policy, bridging the most topics. This metric, in a simple way, covers the lens aperture through which China focuses its narratives.

5 Conclusions

Overall, China's pop culture coverage was the most prominent topic in this corpus and maintained the most connections with all other topics over the duration of time. This is unsurprising as China and Taiwan share many cultural and historical aspects in addition to trade and tourism [7]. Additionally, the topics that remain the most connected all pertain to non-geopolitical themes and are more oriented towards entertainment and trade.

China's negative nationalist sentiment can be seen in the prominent themes around the United States, Japan, and Taiwan. These findings coincide with past research on these same themes also being prominent in mainland Chinese media coverage due to the domestic appetite of Chinese citizens for coverage of these countries [20]. Additionally, we note how China's adversarial dialogue towards the United States transitions in the last period of our data to connect the most topics by having the highest betweenness centrality, indicating a transition in the way that China discusses this topic and how it's linked to other topics.

Further research using these methods would benefit from more finetuning of the topic models. There was noise with the topic model that would benefit from further iterations. Precisely improving the quality of the topic quality output is tricky and perhaps more of an art than a science as algorithmically optimizing output does not necessarily match human quality ratings [11]. A manual and iterative process for data cleaning, topic merging, and parameter tuning is the best way to increase topic quality [19]. From a more policy-based viewpoint, expanding this analysis to more countries could potentially capture more interactions in how China's dialogue has shifted, as this study only captured topics pertaining strictly to Taiwan. Additionally, comparing China's news article narratives to its social media themes may establish how China uses different types of media platforms for different purposes.

References

1. Altman, N., Carley, K.M., Reminga, J.: Ora user's guide 2020. Carnegie-Mellon Univ. Pittsburgh PA Inst of Software Research International, Tech. Rep **2**, 2 (2020)
2. Barron, A.T., Huang, J., Spang, R.L., DeDeo, S.: Individuals, institutions, and innovation in the debates of the french revolution. Proceedings of the National Academy of Sciences **115**(18), 4607–4612 (2018)

3. Blei, D.M., Ng, A.Y., Jordan, M.I.: Latent dirichlet allocation. In: Journal of Machine Learning Research. vol. 3, pp. 993–1022 (2003)
4. Bradshaw, S., Howard, P.N.: Challenging truth and trust: A global inventory of organized social media manipulation. The computational propaganda project **1**, 1–26 (2018)
5. Church, K., Hanks, P.: Word association norms, mutual information, and lexicography. Computational linguistics **16**(1), 22–29 (1990)
6. DiResta, R., Miller, C., Molter, V., Pomfret, J., Tiffert, G.: Telling China's Story: The Chinese Communist Party's Campaign to Shape Global Narratives. Stanford Internet Observatory (2020)
7. Ferle, C.L., Edwards, S.M., Lee, W.N.: Culture, attitudes, and media patterns in china, taiwan, and the us: Balancing standardization and localization decisions. Journal of Global Marketing **21**(3), 191–205 (2008)
8. Fisher, S., Klein, G.: China's global times articles from 9 apr 2009 to 31 dec 2022, `https://dataverse.harvard.edu/dataverse/focusdataproject`
9. Fisher, S., Klein, G.R., Codjo, J.: Focusdata: Foreign Policy through Language and Sentiment. Foreign Policy Analysis **18**(2) (02 2022). https://doi.org/10.1093/fpa/orac002, `https://doi.org/10.1093/fpa/orac002`, orac002
10. Hernández, J.C.: China deploys propaganda machine to defend move against hong kong `https://www.nytimes.com/2020/05/23/world/asia/china-hong-kong-propaganda.html`
11. Hoyle, A., Goel, P., Peskov, D., Hian-Cheong, A., Boyd-Graber, J., Resnik, P.: Is automated topic model evaluation broken?: The incoherence of coherence. https://doi.org/10.48550/arXiv.2107.02173, `http://arxiv.org/abs/2107.02173`
12. Jacobs, C.: Linking narratives. `https://github.com/charityking2358/Linking_Narratives` (2023)
13. Liu, J., Xia, C., Li, X., Yan, H., Liu, T.: A bert-based ensemble model for chinese news topic prediction. In: Proceedings of the 2020 2nd International Conference on Big Data Engineering. pp. 18–23 (2020)
14. Liu, Q., Zheng, Z., Zheng, J., Chen, Q., Liu, G., Chen, S., Chu, B., Zhu, H., Akinwunmi, B., Huang, J., et al.: Health communication through news media during the early stage of the covid-19 outbreak in china: digital topic modeling approach. Journal of medical Internet research **22**(4), e19118 (2020)
15. Manning, C., Schutze, H.: Foundations of statistical natural language processing. MIT press (1999)
16. Office, T.A.: China releases white paper on taiwan question, reunification in new era, `https://english.www.gov.cn/archive/whitepaper/202208/10/content_WS62f34f46c6d02e533532f0ac.html`
17. Perry, C., DeDeo, S.: The cognitive science of extremist ideologies online. arXiv preprint arXiv:2110.00626 (2021)
18. Rehurek, R., Sojka, P.: Gensim–python framework for vector space modelling. NLP Centre, Faculty of Informatics, Masaryk University, Brno, Czech Republic **3**(2) (2011)
19. Roberts, M., Stewart, B., Tingley, D.: Navigating the local modes of big data: the case of topic models. comput. Soc. Sci **4** (2016)
20. SHIRK, S.L.: Changing media, changing foreign policy in china. Japanese Journal of Political Science **8**(1), 43–70 (2007). https://doi.org/10.1017/S1468109907002472
21. Stanley, J.: How propaganda works, pp. 52–53. Princeton University Press (2015)

22. Stockmann, D., Gallagher, M.E.: Remote control: How the media sustain authoritarian rule in china. Comparative political studies **44**(4), 436–467 (2011)
23. Wong, B.: Why is china insisting it is a democracy?, `https://thediplomat.com/2021/12/why-is-china-insisting-it-is-a-democracy/`
24. Wu, H., Ng, E., Mascaro, L.: US house speaker pelosi arrives in taiwan, defying beijing, `https://apnews.com/article/china-asia-beijing-malaysia-a5a6acc391511c99b1b4c2d69e67b133`
25. Wu, J.m.: The china factor in taiwan: Impact and response. Routledge handbook of contemporary Taiwan pp. 426–446 (2016)