Decoding Government Tweets' Impact on Policy Evolution in China Over a Decade

Jason He, Valentina Simon
Economics, Statistics and Data Science
Yale University
May 6th 2024
jason.he@yale.edu, valentina.simon@yale.edu

1 Introduction

China has the second largest economy in the world and a 300 billion dollar trade deficit with the United States, making it a key market for investors worldwide ("Analysis: China's Economy and Its Influence on Global Markets | U.S. Bank," 2024). Due to its government-led capitalism, characterized by state-owned enterprises, massive infrastructure spending, and expansive industrial policy, the "Visible Hand" holds tremendous sway over economic trends in the People's Republic (Yang et al., 2019). Predicting government policy before its announcement would help private corporations mitigate risk and acquire early mover advantage in the next focal industry. However, there is limited research on quantitative indicators for the Chinese government's policy priorities.

While it is not interesting that Chinese propaganda moves concurrently with its policy, a recent study indicates that the rhetoric in state media could indicate political shifts months in advance (Chan & Zhong, 2019). The method described in the paper takes the text of the *People's Daily* front page from 1951Q1 to 2019Q1. Like the Soviet *Pravda*, *People's Daily* is the Chinese Communist Party's press for publicizing its theories, opinions, policies, and ambitions. The authors vectorize the corpus through word embedding and employ machine learning metrics to compute an index that tracks the changes in the paper's policy priorities ahead of the official announcements of related policies. The resulting predictive metric is called the Policy Change Index or PCI.

A literature review shows that similar methods have incorporated only official newspapers and decrees as their data or used pre-defined word lists. However, the readership of legacy papers and public government documents has dwindled over the past two decades. Social media posts of state-affiliated accounts now constitute a crucial source of information. Word frequency analysis may be adequate for press releases and State Council reports, which have a narrow and fixed context and a small and stable lexicon of bureaucratic jargon. They become less robust for newspapers and especially social media because of the fluid and complex context and constant invention of new expressions. In a policy context, these new catchphrases often serve a propagandistic purpose, hence it is important to capture their emergence and usage (Link, 2012).

For the current study, we concentrate on the Weibo tweets of *People's Daily*. Launched in August 2009 after Chinese authorities banned Twitter, Sina Corporation's Weibo experienced a meteoric rise. By the end of 2012, the platform hosted around half a billion registered accounts with 46.3 million daily active users (Chao, 2013). The tremendous impact of social networks initially caught CCP propagandists off guard, yet its media ecosystem soon entrenched itself online. Over a decade later, the three outlets under the State Council's direct control, People's Daily, Xinhua News Agency, and China Media Group, each boast over 100 million followers on the platform (Shelley, 2023).

With the sum of likes, forwards, and comments, one may construct a sample of Weibo tweets simulating a traditional paper's front page. Unlike the top-down selection of front-page articles or prime-airtime programs, engagement with tweets contains a message's reception by the public beyond its emphasis by the administration. A tweet having high total engagement could suggest its importance to users, its prioritization by recommendation algorithms under political demand, or its mandate for government employees to like and share. An imbalance between the number of comments and likes may highlight a tweet's controversy, while an outsized amount of shares may display its popularity. Though lacking in democratic feedback, the Chinese state can revise its policies based on these online interactions.

In this paper, we investigate whether changes in the topics of Weibo posts by the *People's Daily* account are predictive of future governmental policy decisions, months in advance.

2 Related Works

Reading China: Predicting Policy Change with Machine Learning
(Chan & Zhong, 2019): This work developed the first quantitative indicator of the Chinese government's priorities over a long period, which they named the Policy Change Index (PCI) for China. The policy index was trained on the Communist Party of China's official newspaper *People's Daily*, but not tweets on Weibo which we are using in our using in our analysis. The success of PCI in predicting Chinese policy changes months in advance was the inspiration for this project. We follow a similar technique to this article but employ more advanced machine learning techniques (modern Embedding packages and LSTM instead of GRU). We hope to improve upon the results of this paper.

Predicting Authoritarian Crackdowns: A Machine Learning Approach
(Chan & Zhong, 2019): This work employs the PCI as a quantitative indicator of political change in China to predict if and when protests in Hong Kong will meet a Tiananmen-like crackdown. This is an example of how to use a quantitative measure of propaganda article severity to predict policy change, political unrest, and economic implications.

Who is leading China's family planning policy discourse in Weibo?
(Deng et al., 2023): This work identifies a correlation between Weibo postings about family planning policy by the general public and government-run news networks, indicating a bottom-up agenda-setting effect. We used it as an example of how to mine Weibo posts, and of an indication of the importance of social media in Chinese policy-making decisions.

The softening of Chinese digital Propaganda Evidence from the People's Daily Weibo account during the pandemic
(Shao et al., 2023): This longitudinal study fills a critical gap by empirically analyzing the transformation in communication styles of Chinese official media, focusing on the People's Daily's Weibo account from 2019 to 2021. The findings reveal a notable shift towards softer communication strategies, characterized by increased use of soft news, positive messaging, and emotional content, despite occasional spikes in hard news during the initial stages of the COVID-19 pandemic. We used this paper as an example of how to evaluate the sentiment of Weibo posts by the People's Daily and analyze their connection to policy events. However, in addition to sentiment, our study also examines the direct content of Weibo posts to quantify the potential for policy changes.

Predicting judicial decisions of the European Court of Human Rights: a Natural Language Processing perspective
(Aletras et al., 2016): This paper presents the first systematic study on predicting outcomes of cases tried by the European Court of Human Rights solely based on textual content, achieving strong predictive accuracy (79 percent on average). We used it to demonstrate the predictive power of textual content, which we capture from tweets via word embedding. It was a valuable example of using natural language processing to extract key information for analysis.

Measuring economic policy uncertainty in China
(Huang & Luk, 2020): Using a list of keywords and 10 major Chinese newspapers, Huang and Luk constructed a revised Economic Policy Uncertainty (EPU) index based on the seminal paper by Baker et al. (2016). The authors argue that press from mainland China accurately reflects public sentiment about economic uncertainty despite censorship concerns. Their EPU is weakly correlated with the realized daily volatility of the Shanghai Composite Index, but it only picks up financial uncertainty rather than social-political events. Although EPU is uncorrelated with PCI, we adapt this paper's technique of using SSE Composite volatility as a dependent variable.

3 Objective:

In this paper, we investigate two questions:

(1) What is the predictive capacity of viral tweets by Chinese state media on Weibo in determining major changes in Beijing's policies? Can policy shifts be predicted from Weibo texts months in advance?

(2) Does the positive or negative engagement with a tweet on Weibo impact whether and how the government chooses to implement a policy?

We will answer these two questions for the ten-year historical period under Chairman Xi Jinping's rule in China and the specific case study of the COVID-19 Policy in China.

4 Dataset

We obtain the dataset using a crawler to collect the official Daily News tweets on Weibo for the last ten weeks. To avoid bot detection, the crawler periodically sleeps for 5 to 10 seconds randomly for every 1 to 5 pages. As a result, the crawler took nearly 70 hours to run. For every post, we collected: the text of the tweet itself, the number of likes, the number of shares, and the number of comments. Likes, shares, and comments were summed to create a total engagement number column used for analysis. For any given day, we keep the tweets with the highest engagement. We define virality as 1.5 times the inter-quartile range (IQR) above from the third quartile. Over the decade covered by the study, on average there are 28.9 tweets per day, and 2.9 of those tweets (or around 10.1%) are kept in the database.

text	publish_time	likes	forwards	comments	total_engagement
#我支持新疆棉花# 原图	3/24/2021 21:44	2822004	40318437	148596	43289037
#汶川大地震10年# 生死不离，生生不息。十年了，你好吗？...	5/11/2018 23:59	894892	20711529	442583	22049004
【速扩！#河南全省救援电话汇总#】#河南多地暴雨致灾#，...	7/20/2021 21:19	380552	18001583	54196	18436331
【盼英雄归故乡！#老志愿军回忆牺牲战友掩面落泪#】老志...	8/28/2021 12:21	396494	17289451	31066	17717011
今天，转发接力，把最美好的祝福都给你！生日快乐，#14...	10/1/2019 0:00	1162309	14786160	258847	16207316
【此时此刻，无论你在何地，请为中国传递这条微博】对你...	10/1/2018 7:00	308567	14879380	140748	15328695
【转发为#武汉加油#！让武汉人民知道，全国人民和你们在...	1/23/2020 7:17	1174535	12846485	106111	14127131
【#汶川地震13周年#，永不忘却的记忆】缅怀逝者，勇毅前...	5/12/2021 0:00	394346	12743602	52488	13190436
【如果你记得今天，请发条微博：#牢记九一八#】#九一八8...	9/18/2020 0:00	511201	12450987	51687	13013875
【转发祝贺！#神十二发射圆满成功#】北京时间6月17日9时...	6/17/2021 9:46	168231	12766444	17243	12951918

Figure 1: Top 10 engaged tweets from People's Daily

With over 40 million forwards alone, the first tweet supports the Xinjiang cotton industry against allegations of Uighur slave labor. The 2nd and 8th tweets memorialize the anniversaries of the 2008 Sichuan earthquake. The 4th one celebrates Korean War veterans; the Korean War has been played up since tension with America heightened over Taiwan in August 2021. The 8th one is when Wuhan entered lockdown after the government acknowledged the COVID-19 outbreak.

Figure 2 plots the total number of tweets *People's Daily* sends each month and the fraction of viral tweets. As the plot illustrates, the number of tweets varies greatly from under 500 in August 2023 to nearly 2000 in March 2020, yet the fraction stays around 10% from 2016 onwards. Unlike the newspaper issues of *People's Daily* (Chan & Zhong, 2019), the Weibo account exhibits less fluctuation in these two summary statistics, ensuring roughly uniform sample sizes across time.

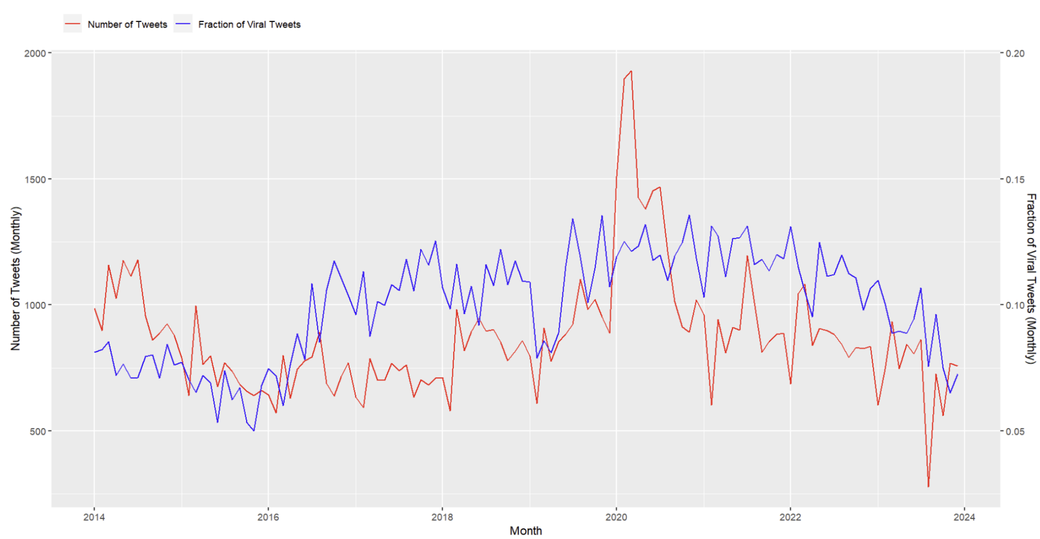

Figure 2: Fraction of Viral Tweets and Number of Tweets in the Data Set

To model economic trends, we downloaded the daily price and volatility of the Shanghai Composite Index from Yahoo Finance. The aggregate index from the largest Chinese financial market is used as a continuous indicator of policy change. Studies have shown that when there is policy uncertainty in China the financial markets react with heightened volatility (Huang & Luk, 2020) and that policy announcements often result in significant movements in the stock indices (Yang et al., 2019). We average the price and volatility of each month and include monthly controls downloaded from the FRED database (CPI growth rate, 3-month bond yield, commodity export, Economic Policy Uncertainty Index, and RMB to USD exchange rate).

Additionally, we created a dataset of key Chinese policy events in the past ten years, as identified by major U.S. news outlets like the Associated Press and Reuters. These dated policy implements are shown below. For each policy change, we determined if it was predictable up to 3 months in advance from Weibo tweet activity on the CCP's official *People's Daily* account.

Date	Policy Change	Details
11/9/2014	"New Normal"	Government announced new policies to address concerns over Chinese economic slowdown
5/19/2015	Made in China 2025	Calls for reducing excess industrial capacity, esp. steel and coal. Relies on direct government action, not a market-oriented reform.
12/27/2015	End of one-child policy	Two-child policy implemented as a result of demographic concerns
10/27/2016	Xi declared the "Core" leader	Xi was elevated to the same status as Mao, Deng, and Jiang
10/25/2017	Enshrining of Xi Jinping Thought	After the 19th Party Congress, Xi was reelected Xi's thought reached the same status as that of Mao Belt & Road written into the Party Constitution
7/6/2018	Beginning of Trade War	China retaliates against Trump tariffs
12/13/2019	Truce over Trade War	Truce reached & tariff revocation
1/24/2020	COVID-19 Outbreak Announced	Wuhan finally entered lockdown despite COVID having been around for over a month
5/24/2020	Hong Kong National Security Law	Fearing further riots in HK like those between Mar and Nov 2019, the government increased its control over the city dramatically
8/20/2020	Three Red Lines	Regulation on the real estate market to prevent further indebtedness
11/7/2020	Jack Ma Affair	Beginning of crackdowns on the tech, financial, crypto, and education sectors
11/11/2021	Historical Resolution	Celebrating centenial of the CCP and building a new narrative for Party history
8/3/2022	Pelosi visits Taiwan	Escalation of tensions over Taiwan ahead of ROC elections
12/7/2022	End of Zero-COVID	Zero-COVID ends after widespread protests
7/7/2023	Military corruption scandal	Mass arrests of the Rocket army, Defense Dept., and Aerospace Admin leadership

Figure 3: Policy changes we used to test the predictive power of Weibo People's Daily tweets

5 Methods

Our model takes in textual information in the form of Weibo tweets and outputs a Policy Change Index (PCI) score. This involves three steps: word embeddings, an LSTM machine learning model, and evaluating F1 scores. These three steps are described in detail below. Assumptions are described in detail in section 7.1

5.1 Word Embedding

When text is modeled as data, the curse of dimensionality kicks in quickly. This can lead to data sparsity, increased computational complexity, and overfitting. In Chinese, there are tens of thousands of distinct characters that can be used within a given tweet. Dimension reduction, consequently, is critical to machine learning analysis.

We utilize word embedding, a technique in natural language processing introduced by Mikolov et al. (2013), to decrease the complexity of our dataset. This method translates words or phrases into numerical vectors within a vector space, effectively reducing the dimensionality of the text while preserving semantic relationships between them.

One key aspect of word embedding is its preservation of the linguistic essence of words by capturing their relationships with other words. For instance, the cosine distance between the embeddings of two words gauges their similarity. Hence, the cosine distance between "the United States" and "China" is smaller than, for instance, the cosine distance between "banana" and "umbrella." Additionally, the orientation of a vector also carries linguistic significance. For example, the disparity between "the United States" and "Washington, DC" — represented as a vector in the same space — resembles the contrast between "China" and "Beijing."

For this project, we used BCEmbedding, a BERT-based embedding model developed by Youdao for text generation. We found that the traditional word2vec did not perform well in preserving the context of words in the Chinese language and most Weibo-specific embeddings date back to 2018 when it was easier to download data from the site.

5.2 LSTM

Long Short-Term Memory (LSTM) is a type of recurrent neural network (RNN) architecture designed to effectively capture and process sequential data while addressing the vanishing gradient problem commonly encountered in traditional RNNs. LSTM networks consist of memory cells that maintain a cell state over time and three gating mechanisms: the input gate, forget gate, and output gate. These gates regulate the flow of information into and out of the memory cells, allowing LSTMs to selectively remember or forget information from previous time steps. This enables them to capture long-term dependencies in sequences by preserving relevant information over extended periods while discarding irrelevant or redundant information. Consequently, LSTMs are widely used in various tasks such as natural language processing and time series prediction, where understanding and modeling sequential patterns is crucial, making them uniquely suited to our needs (Staudemeyer & Morris, 2019).

- Input Gate (i_t): The input gate controls the flow of information into the memory cell at each time step. It consists of a sigmoid activation function that takes as input the current input x_t and the previous hidden state h_{t-1}. It outputs a vector of values between 0 and 1, representing how much of the new input should be added to the cell state.
- Forget Gate (f_t): The forget gate regulates the information retained in the memory cell from previous time steps. It uses a sigmoid activation function to decide which information in the cell state should be discarded. This gate considers both the current input x_t and the previous hidden state h_{t-1}, producing a vector of values between 0 and 1, where 0 means "completely forget" and 1 means "completely remember."
- Output Gate (o_t): The output gate determines the output of the LSTM cell at each time step. It controls how much of the current cell state should contribute to the output and the next hidden state. The output gate employs a sigmoid activation function along with the hyperbolic tangent (tanh) activation function to produce the final output. The sigmoid function regulates the flow of information, while the tanh function squashes the values to the range $[-1, 1]$, ensuring that the output is normalized.

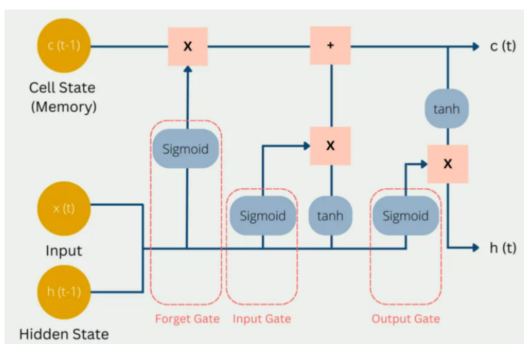

Figure 4: Diagram of an LSTM cell, many of which are strung together to form a machine learning model. (“Long Short-Term Memory Networks (LSTM)” 2022).

5.3 F1 Scores: Metric

The F1 score is a metric commonly used to evaluate the performance of classification models, particularly when dealing with imbalanced datasets. It is the harmonic mean of precision and recall. Precision measures the proportion of true positive predictions among all positive predictions made by the model, while recall measures the proportion of true positive predictions among all actual positive instances in the dataset. The F1 score combines these two metrics to provide a balanced assessment of a model's performance, especially when there is an imbalance between the classes being predicted.

$$F1 = \frac{2 \times \text{precision} \times \text{recall}}{\text{precision} + \text{recall}} \tag{1}$$

The Policy Change Index (PCI) is computed as the absolute difference in the F1 scores between the training window (T_s) and a specific month (s):

$$\text{PCI}(s) = |F1(T_s) - F1(s)| \tag{2}$$

Where: - $F1(T_s)$ represents the F1 score calculated using the training data. - $F1(s)$ represents the F1 score calculated using the specific year-quarter's data.

We used training windows of one year and compared those against one month of tweets. We hypothesize that a large difference in F1 scores can be used as an indicator of future policy change.

This formula ensures that the PCI captures any significant deviations in classification performance between the training period and a given year-quarter, indicating potential shifts in China's policy priorities.

5.4 Alternative Approaches

If infinite data and computing power were available, we could consider leveraging predictive-focused topic modeling via feature selection (Ren et al., 2020). This approach involves optimizing the Evidence Lower Bound (ELBO) objective, which consists of two terms: the log-likelihood of the observed data and the Kullback-Leibler (KL) divergence between the variational distribution and the true posterior distribution of the latent variables.

$$\text{ELBO}(\theta, \phi) = \log p_\theta(y, w) - \text{KL}(q_\phi(\Theta)||p_\theta(\Theta|w, y)) \tag{3}$$

Here, θ represents the model parameters, ϕ represents the variational parameters, and Θ represents the latent variables (e.g., topic proportions, topic assignments).

To perform feature selection, we focus on minimizing the KL divergence term. By choosing a variational family $q_\phi(\Theta)$ that encourages sparsity or selectivity in the latent variables, we can effectively perform feature selection. This is achieved by setting certain components of the variational parameters to zero or near zero, effectively excluding corresponding features from the model.

For example, in the context of pf-sLDA, we use a variational family that includes variational parameters for topic proportions (ϕ), topic assignments (z), and channel switches (π). By setting certain components of the variational parameters (e.g., π) to zero, we effectively exclude certain topics or features from the model. This acts as a form of feature selection, allowing us to focus on the most relevant topics for prediction.

In summary, predictive-focused topic modeling via feature selection involves optimizing the ELBO objective while using a variational family that encourages sparsity or selectivity in the latent variables. This allows us to identify and focus on the most relevant features for predicting Chinese policy changes using Weibo tweets from the official CCP account while discarding less relevant ones.

6 Results

We find that spot, 1-month, 2-month, and 4-month PCIs are significantly correlated with the average volatility of the Shanghai Composite Index. The 1-month lagged PCI has the most significant correlation (p = 0.01) and the lagged PCIs are all more significant than the spot value.

Unlike the Economic Policy Uncertainty (EPU) index, which weakly correlates with the monthly price of the SSECI, lagged PCI does not correlate with price at all. This is consistent with the result in existing literature (Huang & Luk. 2020) that EPU does not correlate with PCI. In addition, PCI approximates changes in the language of *People's Daily* Weibo but does not indicate the direction of policy change. It follows that shifts in state propaganda make investors uncertain and inject volatility into the market.

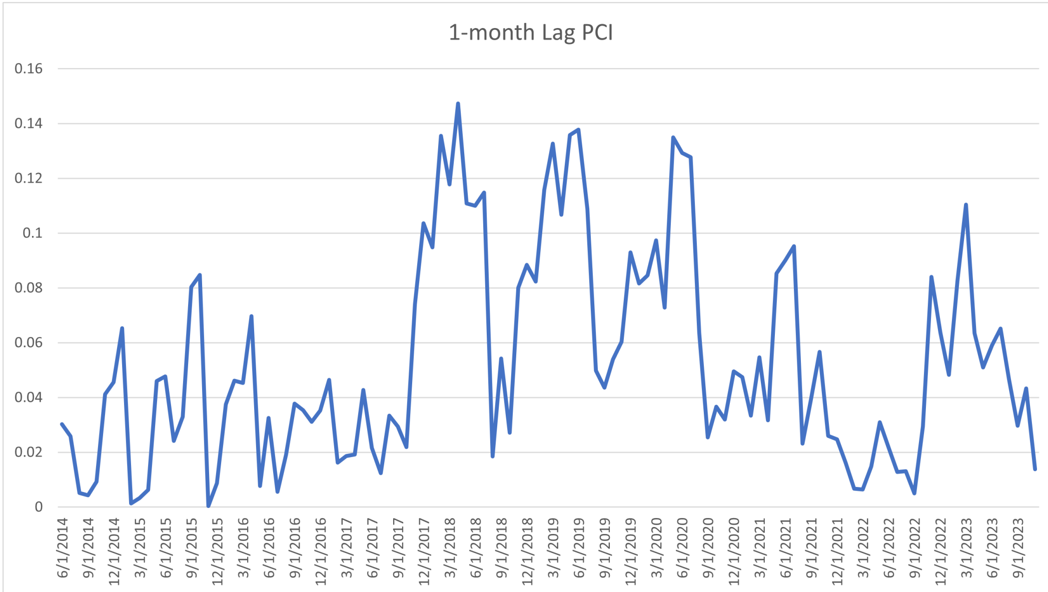

Figure 5: 1-month lagged PCI from People's Daily Weibo

7 Conclusion

Using new embedding techniques and a Weibo corpus, we demonstrate that the Political Change Index developed by Chan & Zhong has a significant correlation with volatility in the Shanghai stock market. Such a correlation is the strongest 1 to 2 months before a high volatility point. Since policy announcements or implementations cause significant movements in the financial markets, PCI could predict Chinese government policies before they enter the official bureaucratic channels.

7.1 Assumptions and Limitations

Numerous assumptions were made in this project. Some of the key ones include:

(1) The central assumption to our model is that social media posts associated with the Chinese Communist Party of China will provide clues to future Chinese policy before it is officially made into law. From the success of similar models examining newspapers from state-run media, this seems to be the case (Chan & Zhong, 2019). However, if it is false, the entire premise of this project fails.

(2) We used a limited dataset for this analysis. We only analyzed Weibo tweets from *People's Daily*, and of those tweets we only looked at the ones that were highly engaged with, or above 1.5 standard deviations more engagements than the mean. This limited our study. Instead, using a comprehensive analysis of CCP tweets, general public tweets, and multiple news sites, could culminate in a better predictive model. Additionally, since China exists in an international market, looking at China related trends internationally in Taiwan or the United States could be valuable as well.

(3) We assumed that governmental policy change was associated with economic downturns to train our model. While these events are highly correlated, sometimes governmental policy change vastly stimulates the market in China or has no impact on the market at all. For instance, in 2022 China passed a 143 trillion dollar stimulus for its semiconductor industry, that jumpstarted the economy in that sector (Zhu, 2022).

8 Reproducibility

The code for this analysis can be found in the Github. A README file is included.

The crawler used to obtain tweets from Weibo can be found in this Github repository (Tao Xiao, 2019). One may use it to mine data from other Chinese state media Weibo accounts and test the robustness of PCI on a new corpus.

References

[1] Aletras, N., Tsarapatsanis, D., Preoţiuc-Pietro, D., & Lampos, V. (2016). Predicting judicial decisions of the European Court of Human Rights: a Natural Language Processing perspective. PeerJ. Computer Science, 2, e93. https://doi.org/10.7717/peerj-cs.93

[2] Baker, S. R., Bloom, N., & Davis, S. J. (2016). Measuring economic policy uncertainty. *The Quarterly Journal of Economics*, 131(4), 1593-1636. https://doi.org/10.1093/qje/qjw024

[3] Chan, J. T., & Zhong, W. (2018). Reading China: Predicting Policy Change with Machine Learning. Social Science Research Network. https://doi.org/10.2139/ssrn.3275687

[4] Chao, K. F. a. M. (2013, March 12). How Many People Really Use Sina Weibo? WSJ. https://www.wsj.com/articles/BL-CJB-17393

[5] Deng, W., Hsu, J., Löfgren, K., & Cho, W. (2021). Who is leading China's family planning policy discourse in Weibo? A social media text mining analysis. Policy & Internet. https://doi.org/10.1002/poi3.264

[6] Huang, Y., & Luk, P. (2020). Measuring economic policy uncertainty in China. *China Economic Review*, 59, 101367. https://doi.org/10.1016/j.chieco.2019.101367

[7] Kjell, O., Giorgi, S., & Schwartz, H. A. (2023). The *text*-package: An R-package for analyzing and visualizing human language using natural language processing and transformers. *Psychological Methods*. https://doi.org/10.1037/met0000542

[8] Link, P. (2012). *An Anatomy of Chinese: Rhythm, Metaphor, Politics*. Harvard University Press. https://doi.org/10.4159/harvard.9780674067684

[9] Long Short-Term Memory Networks (LSTM)- simply explained! | Data Basecamp. (2022, June 4). Data Basecamp. https://databasecamp.de/en/ml/lstms

[10] Shelley, L. (2023, October 19). China's six largest social media platforms – and how to engage its users. Campaign Asia. https://www.campaignasia.com/article/chinas-six-largest-social-media-platforms-and-how-to-engage-its-users/492217

[11] Staudemeyer, R. C., & Morris, E. R. (2019, September 12). Understanding LSTM – a tutorial into Long Short-Term Memory Recurrent Neural Networks. arXiv.org. https://arxiv.org/abs/1909.09586

[12] Yang, Z., Yu, Y., Zhang, Y., & Zhou, S. (2019). Policy uncertainty exposure and market value: Evidence from China. Pacific-basin Finance Journal, 57, 101178. https://doi.org/10.1016/j.pacfin.2019.101178

[13] Zhang, C., Zhang, D., & Shao, H. (2023). The softening of Chinese digital propaganda: Evidence from the People's Daily Weibo account during the pandemic. Frontiers in Psychology, 14. https://doi.org/10.3389/fpsyg.2023.1049671

[14] Zhu, J. (2022, December 14). Exclusive: China readying 143 billion package for its chip firms in face of U.S. curbs. Reuters. https://www.reuters.com

[15] Ren, J., Kunes, R., & Doshi-Velez, F. (2020). Prediction focused topic models via feature selection. Proceedings of the 23rd International Conference on Artificial Intelligence and Statistics (AISTATS) 2020 (Vol. Volume 108). Retrieved from https://proceedings.mlr.press/v108/ren20a/ren20a.pdf