The "One Belt One Road" in Chinese Public Sphere: An Empirical Study using Topic Models

Zichang Ye (zy1545@nyu.edu)

CDS, New York University

May 2020

(Word Count: 2683)

1 Introduction

The diplomatic speeches can be viewed as the signal the government sends to the *international* community. Meanwhile, the news articles, when strictly censored, are often seen as the message that the government wants the *domestic* community to perceive. If such interpretation is indeed accurate, then these two sources should have similar trends of discussing certain topics over time. Furthermore, we can expect the newspapers to react to the change in tactics from the officials, and modify their coverage strategy accordingly. Is such an argument true?

In this paper, we used Structural Topic Models (Roberts 2014) to capture the change in the attention on the "One Belt, One Road" (OBOR) initiative in the Chinese official diplomatic speeches and newspapers. Our analysis suggests that in the official speeches, the mentions of the OBOR generally fluctuated after its launch. Moreover, there is evidence suggesting that the trends of spotlights on OBOR are different in the official speeches and the news.

2 Background and Literature

2.1 Public Sphere in China

In general, the main principles of Chinese government's actions in the public sphere are: (1) oversee the general discussions on the regimes, (2) allow the media to discover local problems, (3) prevent any activities on the ground, and (4) guide the public opinion (Tai 2016; King et al. 2017; Lorentzen 2013).

The government interacts with the newspaper industry through multiple channels. Firstly, there are state-owned newspapers such as People's Daily, voice of which explicitly represents the intents of the Party (Remin Ribao, 2013). Secondly, the privately-owned newspapers are more or less supervised by a censorship program (Tai 2016; Lorentzen 2013). Thirdly, scholars speculate that privately-owned newspapers can be involved in the propaganda, as "watchdog journalism", and publish news according to the signals from the governments out of financial (Esarey 2005) or political incentives (Lorentzen 2013).

2.2 One Belt, One Road

"One Belt, One Road" is the idea of facilitating the economic corporations between China and its Euro-Asia neighbors. The current initiative includes an intensive plan of investment in infrastructures, and setting up financial institution to facilitate the transaction between China and its partners (Ferdinand 2016).

Although there are substantial analyses on the rationale of the OBOR initiative (Yu 2016; Ferdinand 2016; Aoyama 2017), our primary focus is on the spotlight that this initiative has received. The media coverage of the OBOR initiative were widespread globally, and reached its peak in 2017 (Morales, 2019; Google Trends), with an understandable disagreement on the perceptions of the initiative. For example, while Chinese media usually address the opportunities that the OBOR can lead to, media in Vietnam and Czech Republic pay more attention to the controversial issues such as a territorial dispute(Morales, 2019; Tung 2018; Matura 2018).

3 Research Questions

In this paper, we are interested in how the proportions of topics related to OBOR change over time in both government speeches and newspapers. We divided this task into two following questions.

- **Research Question 1.** Is it true that recently, Chinese diplomatic officials seemed to mention less about the OBOR in the public sphere? If so, what can be the reasons driving fewer spotlights on OBOR now?
- **Research Question 2.** Will we observe similar trends of the proportions of some topics in the news article dataset as we saw in the speeches dataset? Moreover, will the trends in speeches predate the trends in the media, as the media needs time to respond to the change in the official narrative?

4 Data and Methods

4.1 Data

The speeches are selected and published by the Ministry of Foreign Affairs of China from April 7th, 2011 to April 20th, 2020, addressed by Chinese diplomatic officials in various circumstances[1]. The preprocessed corpus consists of 999 documents and 17,318 tokens.

The news articles are crawled news articles from September 1st, 2016 to November 31st, 2016, by Webhose.io[2]. The preprocessed document-feature matrix consists of 316,003 documents and 92,375,660 tokens. The vocabulary of the News Articles Dataset covers the entire vocabulary of the speeches.

For both corpora, we removed the built-in Chinese stop-words set in Quanteda, the numbers, and punctuation. The numbers and punctuation are removed because they are relatively irrelevant to the topics contained in the corpus. We also trimmed the document-term matrix by setting the minimal document frequency to 20, and minimal term frequency to 30[3].

4.2 Methods: Structural Topic Models

4.2.1 Research Question 1

To test the first hypothesis, we fitted an STM model on the speeches corpus and estimated the effects of time on the proportions of the relevant topics. The first step is to find the relevant topics. The package used spectral methods to search for the optimal number of topics, and selected 81 topics.

To label the topics, we adapted the qualitative principles described in Catalinac (2016). The first step is to obtain a candidate set of topics that are related to the "One Belt One Road" tactics. We first read the top 10 words of the highest probabilities of occurring under each topic to filter four topics, which are directly related to the keywords "One Belt One Road" and "Silk Road." Because the policy of the "One Belt One Road" first appeared in September 2013, there could be tactics before September 2013 that have similar natures but are named differently. To find these topics, we found the topics of the highest cosine similarity with each of these four topics. Doing this adds two more topics to our candidate set (cosine similarity = 0.148, and 0.173).

The next step is to examine the documents related to these topics. For each topic, we read the top 2 documents with the highest proportion of words from this topic and determined its content and themes. We noticed that in two topics, the top 1 document talk about the OBOR directly, while the top 2 documents are about the peaceful use of nuclear power. Top 2 documents in another two topics talk about international

[1] https://www.fmprc.gov.cn/web/

[2] https://webhose.io/free-datasets/chinese-news-articles/

[3] In topics model, the model estimates the probability that a word occurs under a given topic. Less likely words are in the lower end of the distribution of words, and are not likely to affect the fitted models.

cooperation and the OBOR [4]. Finally, we fitted a regression model using the function `estimateEffect()`[5].

(a) Selected Keywords For Topic 71

(b) Selected Keywords For Topic 34

Figure 1: Research Question 1. Selected Topics

4.2.2 Research Question 2

To explore the second question, we experimented with three strategies. Firstly, we used the built-in `fitNewDocuments()`function in STM package. We first preprocessed the `news` corpus so that its vocabulary matches the vocabulary of the `speeches` corpus. This function then used the topic models we had trained for the `speeches` corpus to predict the estimated proportions of topics in the new documents.

The second method was to combine the two corpora. Because the resulted corpus are imbalanced, we have a few options. Upsample the speeches will lead to a document-features matrix too large to manage. We adapted the idea of downsampling but sampled 1% from the news data. The resulting corpus contains 999 `speeches` documents and 3159 `news` documents. We wanted the corpus to reflect the public sphere in reality, where speeches are a relatively small portion in the public sphere. We then trained a new STM model, and estimated the effects of the `source` and the interaction of `source` and `time` (N = 4158).

The third method we attempted is intuitive: we simply counted the occurrences of the keywords generated by the topic model on the `speeches` corpus in the `news` corpus and normalized them by the length of each document.

5 Results

5.1 Research Question 1: Change in Official Narratives

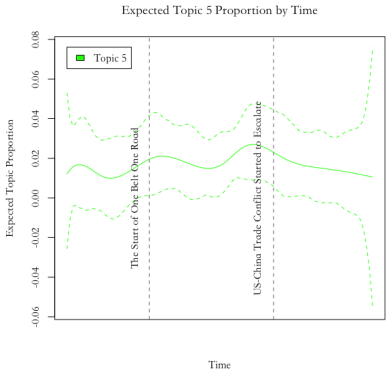
We found that the expected proportions of topics related to OBOR generally increased after its launch, fluctuating, and reached a historically low point recently (April 2020). The effect of time on the expected

[4]The themes of documents align with the well-known fact that China usually phrased the OBOR as global cooperation.

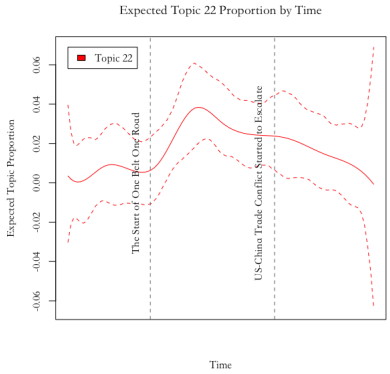
[5]We noticed that there is an error in the function `summary.estimateEffect()`, which is supposed to give us the standard errors and p-value of the coefficients of the covariates, in our case "date". We then looked at the source codes and re-implemented the summary function.

proportions of the topics are not statistically significant (N = 999), which implies that there is no evidence that the topics related to OBOR keep increasing over time.

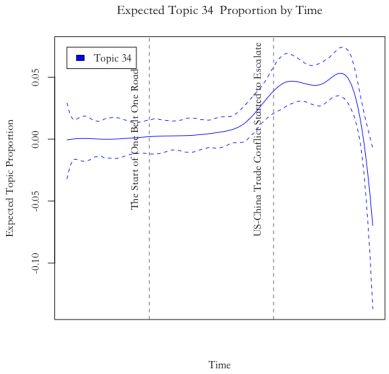
One reason that may drive the drop in spotlights that OBOR is the pressure from the United States and Japan because OBOR will inevitably compete with these countries in the Euro-Asia economic sphere. We thus plotted the date on which the U.S. and China signed the first of a series of trade deals, and considered that as a time when the tension between the U.S. and China started to escalate. The topics 5 and 71 reached a peak before the signing of the trade deal and started to drop after that. However, Topic 22 started to decline earlier than that, and Topic 34 even reached a historical peak after the cutoff. As a result, there is not solid evidence on whether the diplomatic pressure drives China to conceal the regional ambition[6].



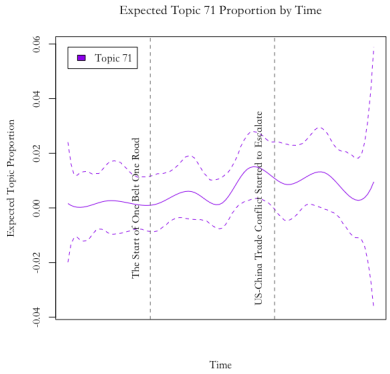
(a) Topic 5 Keywords: OBOR, International Cooperation



(b) Topic 22 Keywords: Silk Road, Nuclear



(c) Topic 34 Keywords: OBOR, Diplomacy



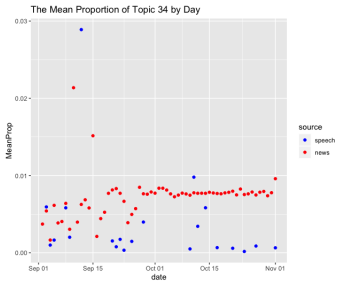
(d) Topic 71 Keywords: OBOR, Nuclear

Figure 2: Research Question 1: Expected Proportion of Selected Topics by Time

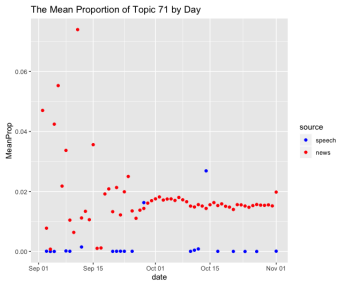
[6]Interestingly, as the Google Trends suggests, the interest index on the OBOR peaked during May 2017. See Appendix for an illustration.

5.2 Research Question 2: Comparison between Official and Media Narratives

The first method, however, classified the new documents incorrectly[7]. We might lose too much information about the texts in the preprocessing steps, forcing the topic model to "overfit" the new data. In a supervised learning framework, such error occurs when the training and test sets are highly heterogeneous. Because the classification is inaccurate, the statistical test is not reliable[8].



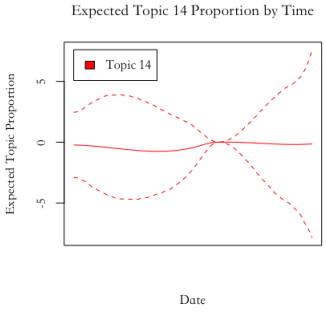
(a) Topic 34 Prevalence by Source and Time



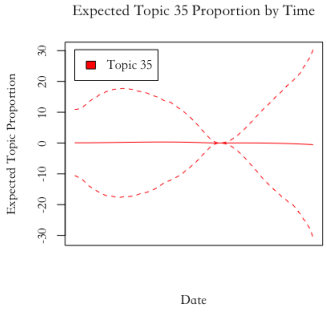
(b) Topic 71 Prevalence by Source and Time

Figure 3: Research Question 2 Method 1: Mean Proportion of Topics 34 and 71 by Source and Time

The second method tried to remedy the loss of information in the preprocessing step. As expected, because the tactics of the OBOR plays a relatively small role in the public sphere, the mean of the estimated proportions are all close to zero. The result, however, shows a large variation, making the inference untrustworthy as well[9]. The reason is mainly because there are not enough data outside of the period from September 2016 to November 2016.



(a) Topic 14 Expected Prevalence by Time

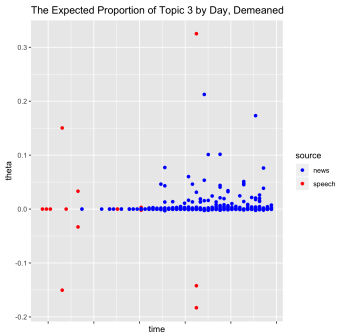


(b) Topic 35 Expected Prevalence by Time

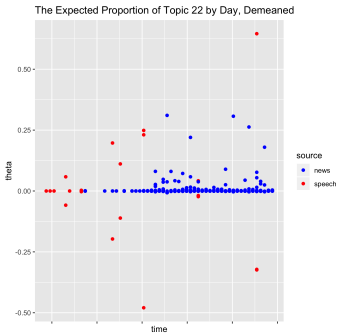
Figure 4: Research Question 2 Method 2: Expected Proportion of Topics 14 and 35 by Time, All

To circumvent this defect of data, we zoomed into the period between September 2016 to November 2016 and ran a regression. There is evidence that there is a significant difference between the trends of proportion of Topics 1, 3, 22 and 35 in the `speeches` and `news` corpus[10]. To visualize whether one of the

[7] For example, one of the documents that the model predicted to be the most relevant to the OBOR was an advertisement of a Taiwanese village.
[8] We included the results of the statistical test in the Appendix.
[9] Full regression table in Appendix.
[10] Full regression table in Appendix.



(a) Topic 3 Expected Prevalence by Time, Demeaned



(b) Topic 22 Expected Prevalence by Time, Demeaned

Figure 5: Research Q. 2 Method 2: Expected Proportion of Topics 1 and 59 by Time, Demeaned and Zoomed

trends predate another, we demeaned the expected proportions by group, so that we can see the relative increase and decrease in the prevalence in this period. We did observe that the predicted proportions of OBOR-relevant topics in `speeches` increased earlier than the `news`, but we should also be aware that the `news` data is highly skewed to the later time. As a result, one can argue that the lag in an increase in the topics proportions in `news` is due to the lack of news data in earlier time, not the fact that the newspaper is actually "watchdogs".

Finally, in our third method of counting keywords in both corpus, we observed that the `news` discusses the keywords related to OBOR in a more stable manner (`Std.Dev=0.02`) than the `speeches` (`Std.Dev=0.1`). The beauty of this method is that we can have an intuitive comparison between how the topics were being touched upon during this period without investing too much computational effort.[11]

One Road	Cooperations	Development	One Belt	East
Economy	Cooperations	Partner	Asia	Construct

Table 1: Reserach Question 2 Method 3: Selected Keywords

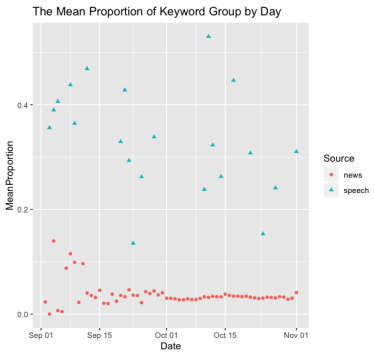


Figure 6: Reserach Question 2 Method 3: The Mean Proportion of Keywords Group by Day

[11] As Boellstorff (2012) points out, counting can be a powerful method in answering quantitative social science problems.

6 Discussion

This case study started with a vague, personal feeling that the mentions of OBOR seemed to cool down recently, and extends its curiosity to the public sphere in China. We attempted to answer our questions by analyzing the speeches corpus given by Chinese diplomatic officials and a publicly accessible news corpus using the Structural Topics Models. We noticed that in diplomatic statements, the mentioning of topics related to OBOR generally climbed up after its launching, and have reached a historically low level recently. Additionally, we found evidence that the trends of OBOR-related topics may be different in these two corpora.

There are four major limitations to our current methodology. Firstly, the regressions we ran did not control for coufounding variables, but the topics discussed by diplomats are highly contingent on the circumstances. For example, the low mentions of OBOR in the diplomatic speeches in April 2020 can be due to COVID-19, instead of the diplomats' dampening interests in OBOR. As a result, we need to control variables more carefully to make formal inference on the evolutions of topic proportions over time.

Secondly, the idea of applying the trained topic models on new documents has its inherent difficulty. In addition to the challenge of losing information during preprocessing we discussed, it is likely that the new documents are from a different word-generating process, especially when the sphere of the new documents has a greater variety of topics and vocabulary. Fitting a new topic model on the larger corpus can already be costly; however, even when computational resources is not a constraint, a new topic model fitted on the larger, combined corpus is not guaranteed to group the topics we want.

Thirdly, our news corpus is a web-crawled dataset covering a relatively small period, and the academic community has used Chinese news corpus with more extended coverage for their studies. For example, Li and Hovy (2014) used People's Daily spanning from 1950 to 2010, and Morales (2019) used the LexisNexis Academic Universe[12]and GDELT corpus[13].

Fourthly, we notice that our second question can be framed as whether two time series have the same shape, or whether one predates another. There are statistical and geometric tests designed for such purposes. For example, we can run a time series model on both series of data and see whether the parameters of trends are similar[14], or we can make an analogy from spatial-temporal questions, and use metrics such as Fréchet Distance to compare the similarity of the shapes of these two time series.

With all the limitations, the value that this study adds to the academic discussion is its efforts to experiment and summarize three strategies of analyzing the relationship between the government and the media in the public sphere press using computational methods. Based upon our findings, we can doubt the naive speculations about the the press industry being simply "watchdog" in China[15]. The relationship

[12]https://guides.nyu.edu/az.php?q=NYU02479
[13]https://www.gdeltproject.org/
[14]https://www.researchgate.net/post/How_do_I_compare_two_time_series_trends_in_terms_of_magnitude_of_decline_or_increase
[15]A bias that the author used to hold.

between the government and the press in China may better be interpreted on two levels. First of all, political speeches, state-owned newspapers, and private newspapers have different functions in the public sphere, and there is no evidence that they all speak in an identical tone. Secondly, the relationship between the censorship program and the press is more like chess players, instead of parents and children. With the evolution of new media, the measures of the program have switched from mere coercion to more diverse strategies such as distracting and incentivizing (Esarey 2015; Lorentzen 2013; King et al. 2017).

Can we use text data to measure the strictness of these measures over time? This is another interesting question, and we will now leave it for future works.

7 Acknowledgement

I would like to thank David Cai [16] for a brainstorming conversation that helps setting the second research question.

8 Reference

- Aoyama, R. (2016). "One belt, one road": China's new global strategy. Journal of Contemporary East Asia Studies, 5(2), 3-22.
- Boellstorff, T., Nardi, B., Pearce, C., Taylor, T. L. (2012). Ethnography and virtual worlds: A handbook of method. Princeton University Press.
- Breslin, S. (2013). China and the global order: signalling threat or friendship?. International Affairs, 89(3), 615-634.
- Callahan, W. A. (2009). China: The pessoptimist nation. OUP Oxford.
- Catalinac, A. (2018). From pork to policy: The rise of programmatic campaigning in Japanese elections. In Critical Readings on the Liberal Democratic Party in Japan (pp. 882-917). Brill.
- Chen, X., Shi, T. (2001). Media effects on political confidence and trust in the People's Republic of China in the post-Tiananmen period. East Asia, 19(3), 84-118.
- Esarey, A. (2005). Cornering the market: state strategies for controlling China's commercial media. Asian Perspective, 37-83.
- Ferdinand, P. (2016). Westward ho—the China dream and 'one belt, one road': Chinese foreign policy under Xi Jinping. International Affairs, 92(4), 941-957.

[16] https://lawecon.ethz.ch/group/people/cai.html

- King, G., Pan, J., Roberts, M. E. (2017). How the Chinese government fabricates social media posts for strategic distraction, not engaged argument. American political science review, 111(3), 484-501.

- King, G., Pan, J., Roberts, M. E. (2017). How the Chinese government fabricates social media posts for strategic distraction, not engaged argument. American political science review, 111(3), 484-501.

- Li, J., Hovy, E. (2014, October). Sentiment analysis on the people's daily. In Proceedings of the 2014 conference on empirical methods in natural language processing (EMNLP) (pp. 467-476).
- Morales,D. (2019). The Belt and Road Initiative in global media:Comparing news coverage in 30 (28) nations. *Presentation*
- Ribao, R. (2013). People's Daily Profile. Retrieved from http://www.people.com.cn/GB/50142/104580/index.html.
- Roberts, M. E., Stewart, B. M., Tingley, D. (2014). stm: R package for structural topic models. Journal of Statistical Software, 10(2), 1-40.
- Sidaway, J. D., Woon, C. Y. (2017). Chinese narratives on "One Belt, One Road"() in geopolitical and imperial contexts. The Professional Geographer, 69(4), 591-603.
- Wang, H., Sparks, C., Huang, Y. (2018). Measuring differences in the Chinese press: A study of People's Daily and Southern Metropolitan Daily. Global Media and China, 3(3), 125-140.
- Yu, H. (2017). Motivation behind China's 'One Belt, One Road'initiatives and establishment of the Asian infrastructure investment bank. Journal of Contemporary China, 26(105), 353-368.

9 Appendix

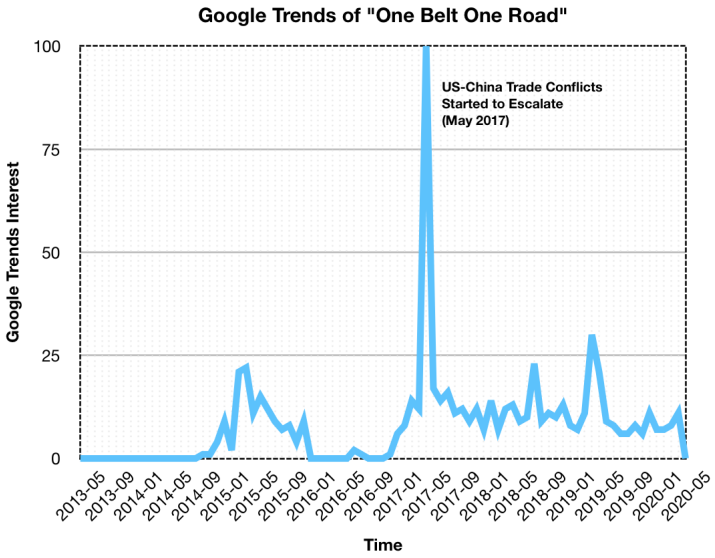


Figure 7: Google Trends of OBOR from May 2013 to May 2020

	Topic 5	Topic 22	Topic 34	Topic 71
(Intercept)	0.08	0.16	0.02*	−0.00
	(0.09)	(0.08)	(0.01)	(0.03)
total_source1	−0.08	−0.16	−0.01	0.03
	(0.09)	(0.09)	(0.01)	(0.03)
s(total_date)1	−0.26	−0.27	−0.03	0.00
	(0.15)	(0.14)	(0.01)	(0.05)
s(total_date)2	0.52***	−0.09	−0.01	−0.00
	(0.08)	(0.08)	(0.01)	(0.03)
s(total_date)3	0.06	0.23	0.02	0.01
	(0.15)	(0.15)	(0.01)	(0.05)
s(total_date)4	−0.15	−0.35***	−0.04***	−0.01
	(0.09)	(0.09)	(0.01)	(0.03)
s(total_date)5	−0.00	0.03	0.01	0.04
	(0.14)	(0.13)	(0.01)	(0.04)
s(total_date)6	−0.06	−0.11	−0.01	−0.01
	(0.10)	(0.09)	(0.01)	(0.03)
s(total_date)7	−0.08	−0.04	−0.01	0.02
	(0.11)	(0.10)	(0.01)	(0.03)
s(total_date)8	−0.08	−0.27*	−0.02	−0.02
	(0.12)	(0.11)	(0.01)	(0.04)
s(total_date)9	−0.08	−0.07	−0.01	0.02
	(0.13)	(0.12)	(0.01)	(0.04)
s(total_date)10	−0.08	−0.16	−0.02	0.00
	(0.09)	(0.09)	(0.01)	(0.03)
total_source1:s(total_date)1	0.26	0.29	0.02	−0.02
	(0.16)	(0.15)	(0.01)	(0.05)
total_source1:s(total_date)2	−0.51***	0.10	0.01	0.02
	(0.10)	(0.09)	(0.01)	(0.03)
total_source1:s(total_date)3	−0.06	−0.22	−0.02	−0.04
	(0.16)	(0.15)	(0.01)	(0.05)
total_source1:s(total_date)4	0.16	0.36***	0.04***	−0.00
	(0.10)	(0.10)	(0.01)	(0.03)
total_source1:s(total_date)5	0.01	−0.02	−0.01	−0.06
	(0.15)	(0.14)	(0.01)	(0.05)
total_source1:s(total_date)6	0.06	0.12	0.02	−0.00
	(0.11)	(0.10)	(0.01)	(0.03)
total_source1:s(total_date)7	0.09	0.05	0.01	−0.04
	(0.11)	(0.11)	(0.01)	(0.04)
total_source1:s(total_date)8	0.08	0.28*	0.02*	0.00
	(0.13)	(0.12)	(0.01)	(0.04)
total_source1:s(total_date)9	0.09	0.08	0.01	−0.04
	(0.14)	(0.13)	(0.01)	(0.04)
total_source1:s(total_date)10	0.08	0.17	0.02*	−0.01
	(0.10)	(0.10)	(0.01)	(0.03)
R^2	0.79	0.55	0.55	0.44
Adj. R^2	0.72	0.39	0.40	0.24
Num. obs.	82	82	82	82
RMSE	0.04	0.03	0.00	0.01

$^{***}p < 0.001$, $^{**}p < 0.01$, $^{*}p < 0.05$

Table 2: Reserach Question 2 Method 1: Statistical Test

	Topic 1	Topic 3	Topic 22	Topic 59	Topic 14	Topic 35
(Intercept)	−198.32	−75.26	−58.72	−55.51	−152.44	−100.59
	(362.72)	(529.12)	(551.95)	(387.57)	(91.69)	(435.08)
s(date_v2)1	−44.96	1864.17	2208.58	−1745.60	32.92	−1842.97
	(11648.79)	(16992.69)	(17725.90)	(12446.80)	(2944.69)	(13972.59)
s(date_v2)2	198.60	67.58	49.51	69.19	147.83	107.63
	(367.76)	(536.47)	(559.62)	(392.96)	(92.97)	(441.13)
s(date_v2)3	198.32	75.29	58.75	55.46	152.49	100.57
	(362.71)	(529.10)	(551.93)	(387.56)	(91.69)	(435.06)
s(date_v2)4	198.32	75.26	58.72	55.52	152.45	100.59
	(362.72)	(529.12)	(551.95)	(387.57)	(91.69)	(435.08)
s(date_v2)5	198.32	75.26	58.72	55.51	152.45	100.59
	(362.72)	(529.12)	(551.95)	(387.57)	(91.69)	(435.08)
s(date_v2)6	198.32	75.26	58.72	55.52	152.45	100.59
	(362.72)	(529.12)	(551.95)	(387.57)	(91.69)	(435.08)
s(date_v2)7	198.32	75.26	58.72	55.51	152.44	100.59
	(362.72)	(529.11)	(551.94)	(387.56)	(91.69)	(435.07)
s(date_v2)8	198.82	75.40	58.80	55.72	152.89	100.84
	(363.54)	(530.32)	(553.20)	(388.45)	(91.90)	(436.06)
s(date_v2)9	0.13***	−0.07*	0.17***	0.02	0.01	−0.04
	(0.02)	(0.03)	(0.04)	(0.03)	(0.01)	(0.03)
s(date_v2)10	0.01	0.02	−0.06*	0.12***	0.01	−0.08***
	(0.01)	(0.02)	(0.02)	(0.02)	(0.00)	(0.02)
source_v2speech	198.32	75.30	58.75	55.51	152.45	100.71
	(362.72)	(529.12)	(551.95)	(387.57)	(91.69)	(435.08)
s(date_v2)1:source_v2speech	44.96	−1864.21	−2208.58	1745.65	−32.92	1843.05
	(11648.79)	(16992.69)	(17725.90)	(12446.80)	(2944.69)	(13972.59)
s(date_v2)2:source_v2speech	−198.59	−67.51	−49.52	−69.16	−147.82	−107.81
	(367.76)	(536.47)	(559.62)	(392.96)	(92.97)	(441.13)
s(date_v2)3:source_v2speech	−198.27	−75.29	−58.72	−55.43	−152.49	−100.62
	(362.71)	(529.10)	(551.93)	(387.56)	(91.69)	(435.06)
s(date_v2)4:source_v2speech	−198.28	−75.14	−58.52	−55.50	−152.44	−100.70
	(362.72)	(529.12)	(551.95)	(387.57)	(91.69)	(435.08)
s(date_v2)5:source_v2speech	−198.30	−75.07	−58.88	−55.51	−152.43	−100.67
	(362.72)	(529.12)	(551.95)	(387.57)	(91.69)	(435.08)
s(date_v2)6:source_v2speech	−198.36	−75.47	−58.61	−55.55	−152.47	−100.70
	(362.72)	(529.12)	(551.95)	(387.57)	(91.69)	(435.08)
s(date_v2)7:source_v2speech	−198.27	−75.24	−58.60	−55.46	−152.44	−100.69
	(362.72)	(529.11)	(551.94)	(387.56)	(91.69)	(435.07)
s(date_v2)8:source_v2speech	−198.70	−75.33	−58.95	−55.66	−152.89	−100.90
	(363.54)	(530.32)	(553.20)	(388.45)	(91.90)	(436.06)
R^2	0.21	0.11	0.14	0.11	0.15	0.26
Adj. R^2	0.20	0.11	0.14	0.11	0.15	0.25
Num. obs.	4158	4158	4158	4158	4158	4158
RMSE	0.05	0.07	0.07	0.05	0.01	0.06

$^{***}p < 0.001$, $^{**}p < 0.01$, $^{*}p < 0.05$

Table 3: Reserach Question 2 Method 2 ver.1: Statistical Test

	Topic 1	Topic 3	Topic 22	Topic 59	Topic 14	Topic 35
(Intercept)	−0.00	0.01	0.03	−0.01	−0.00	−0.01
	(0.04)	(0.05)	(0.08)	(0.07)	(0.03)	(0.03)
s(h2_v2_right_date)1	0.00	−0.03	−0.04	0.05	−0.00	0.02
	(0.05)	(0.08)	(0.12)	(0.10)	(0.05)	(0.04)
s(h2_v2_right_date)2	0.00	−0.01	−0.02	0.00	0.01	0.01
	(0.03)	(0.05)	(0.07)	(0.06)	(0.03)	(0.03)
s(h2_v2_right_date)3	0.00	−0.01	−0.03	0.02	0.01	0.01
	(0.04)	(0.05)	(0.08)	(0.07)	(0.03)	(0.03)
s(h2_v2_right_date)4	0.00	−0.01	−0.03	0.01	0.01	0.01
	(0.04)	(0.05)	(0.08)	(0.06)	(0.03)	(0.03)
s(h2_v2_right_date)5	0.00	−0.01	−0.03	0.02	0.01	0.02
	(0.04)	(0.05)	(0.08)	(0.07)	(0.03)	(0.03)
s(h2_v2_right_date)6	0.01	−0.01	−0.02	0.01	0.01	0.02
	(0.04)	(0.05)	(0.08)	(0.07)	(0.03)	(0.03)
s(h2_v2_right_date)7	0.00	−0.01	−0.03	0.02	0.00	0.01
	(0.04)	(0.05)	(0.08)	(0.07)	(0.03)	(0.03)
s(h2_v2_right_date)8	0.01	−0.01	−0.03	0.01	0.01	0.01
	(0.04)	(0.05)	(0.08)	(0.07)	(0.03)	(0.03)
s(h2_v2_right_date)9	0.00	−0.01	−0.02	0.02	0.01	0.02
	(0.04)	(0.05)	(0.08)	(0.07)	(0.03)	(0.03)
s(h2_v2_right_date)10	0.00	−0.01	−0.03	0.01	0.01	0.01
	(0.04)	(0.05)	(0.08)	(0.07)	(0.03)	(0.03)
h2_v2_right_sourcespeech	−0.01	−0.02	0.78***	0.03	0.01	0.11***
	(0.04)	(0.06)	(0.08)	(0.07)	(0.03)	(0.03)
s(h2_v2_right_date)1:h2_v2_right_sourcespeech	0.17**	1.59***	−1.82***	−0.15	0.01	−0.26***
	(0.06)	(0.09)	(0.13)	(0.11)	(0.06)	(0.05)
s(h2_v2_right_date)2:h2_v2_right_sourcespeech	−0.10*	−1.18***	0.31***	0.04	−0.01	0.15***
	(0.04)	(0.06)	(0.09)	(0.07)	(0.04)	(0.03)
s(h2_v2_right_date)3:h2_v2_right_sourcespeech	0.09*	0.48***	−0.71***	0.10	0.04	−0.23***
	(0.05)	(0.07)	(0.10)	(0.08)	(0.04)	(0.04)
s(h2_v2_right_date)4:h2_v2_right_sourcespeech	−0.01	0.25***	−0.78***	−0.11	−0.01	0.02
	(0.04)	(0.06)	(0.09)	(0.08)	(0.04)	(0.03)
s(h2_v2_right_date)5:h2_v2_right_sourcespeech	0.10	−0.30***	−0.80***	0.09	0.03	−0.30***
	(0.06)	(0.09)	(0.13)	(0.11)	(0.06)	(0.05)
s(h2_v2_right_date)6:h2_v2_right_sourcespeech	−0.18	0.59***	−0.13	−0.24	−0.10	0.44***
	(0.10)	(0.15)	(0.23)	(0.18)	(0.10)	(0.08)
s(h2_v2_right_date)7:h2_v2_right_sourcespeech	0.32	−0.89***	−1.89***	0.31	0.14	−1.02***
	(0.17)	(0.25)	(0.37)	(0.30)	(0.16)	(0.14)
s(h2_v2_right_date)8:h2_v2_right_sourcespeech	−0.93	2.66***	2.44*	−1.04	−0.38	2.55***
	(0.49)	(0.72)	(1.06)	(0.87)	(0.45)	(0.39)
s(h2_v2_right_date)9:h2_v2_right_sourcespeech	2.72	−7.69***	−8.45**	2.92	1.04	−7.88***
	(1.43)	(2.11)	(3.12)	(2.55)	(1.33)	(1.16)
R^2	0.05	0.65	0.53	0.02	0.02	0.15
Adj. R^2	0.04	0.65	0.53	0.01	0.01	0.14
Num. obs.	3185	3185	3185	3185	3185	3185
RMSE	0.01	0.02	0.03	0.02	0.01	0.01

$^{***}p < 0.001$, $^{**}p < 0.01$, $^{*}p < 0.05$

Table 4: Reserach Question 2 Method 2 ver.2: Statistical Test