News Representations of Chinese International Students

By Susan Wang

01
Introduction
&
Research Question

Background

- ❑ China is the No.1 source of international students in US
- ❑ Chinese international students is popularly discussed in US news medias
- ❑ The COVID-19 pandemic had impacted Chinese international students and discussions about them,
  - ❑ Topics of Discrimination?
- ❑ So, how has the news representations changed since the COVID-19 outbreak?

Research Question

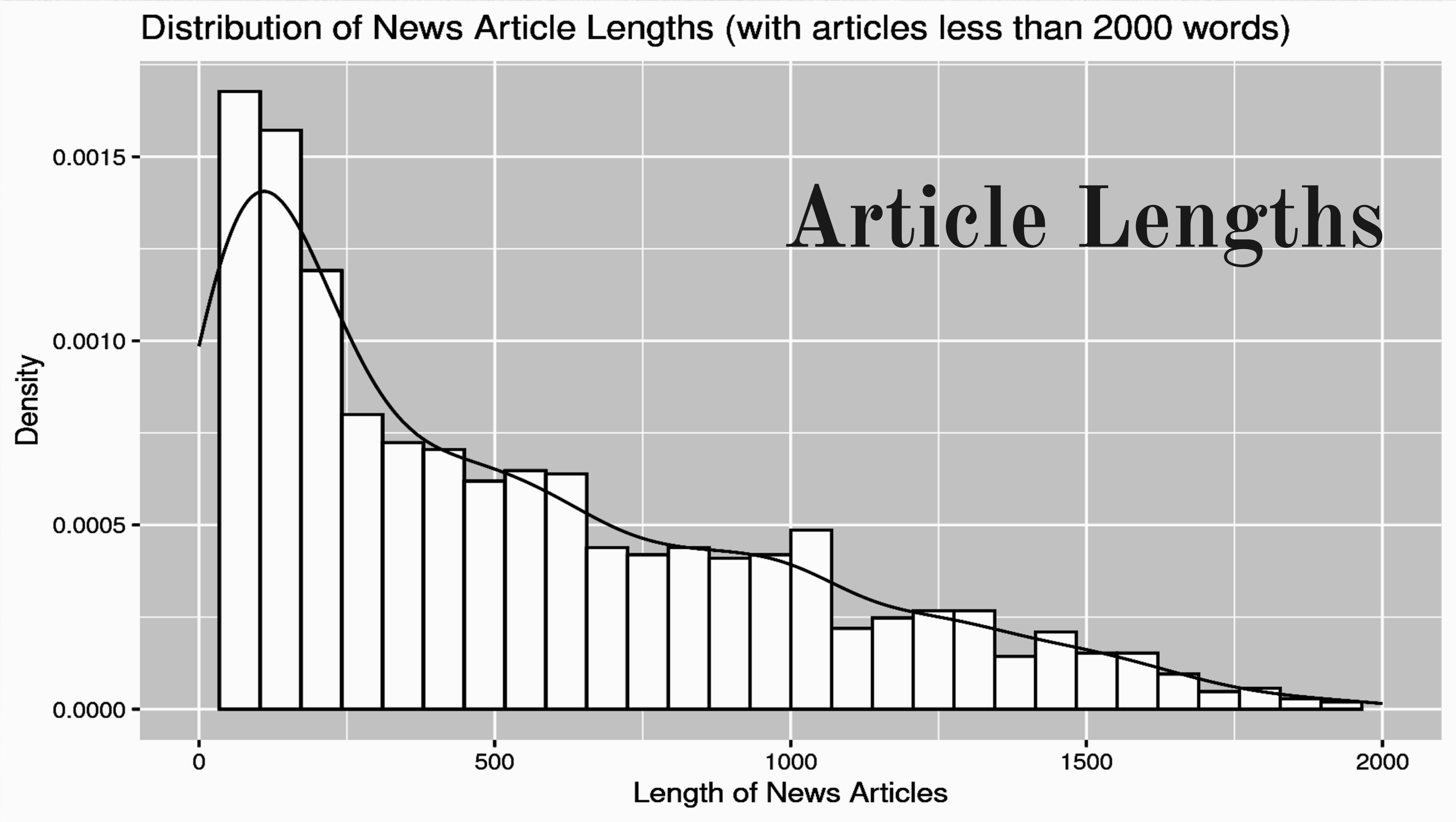
- ❑ Does the COVID-19 pandemic have a negative impact on popular US news media representations of Chinese international students?
  - ❑ Based on Structural Topic Modeling, what topics are significantly correlated with each time periods?
  - ❑ Does those topics have significantly different sentiment scores?

02
Data

Data Sources

- **ProQuest.com.**
- **Publications:**
  - New York Times
  - Wall Street Journal
  - The Washington Post
  - Chicago Tribune
  - Los Angeles Times
  - Boston Globe
- **Time Ranges:**
  - **Before**: 2017/01/01 — 2020/01/01
  - **After**: 2020/01/02 — 2023/01/02
- **Total Articles**: 1525
  - Before: 741
  - After: 784

Article Lengths



Metadata

Subject	Keywords and topics for the article (from ProQuest)
Title	News article title
Publication	Which news publication is the article from
Publication Date	Which day were the article published
Place of Publication	The location of publication (city + state)
Time_COVID	"Before" or "after" the COVID-19 outbreak

03
Methodology

Overview

Pre-processing

Lowercase etc. and add columns

01

Structural Topic Modeling

Get topics that are significantly associated with each time period

02

Sentiment Analysis Using LIWC

Extract sentiment scores for each article

03

Two sample t-tests

Determine which sentiment features are significant for each topics

04

Structural Topic Modeling

= LDA (or other TM) + Covariates

- Each document is mixture of topics — **Document-Topic Distribution**

- Each topic is a mixture of words — **Topic-Word Distribution**

$$\vec{\theta}_d | \boxed{X_d}\gamma, \Sigma \sim \text{LogisticNormal}(\mu = \boxed{X_d}\gamma, \Sigma)$$

$$\beta_{d,k} \propto \exp(m + \kappa_k^{(t)} + \kappa_{\boxed{y_d}}^{(c)} + \kappa_{\boxed{y_d},k}^{(i)})$$

$$z_{d,n} | \vec{\theta}_d \sim \text{Multinomial}(\vec{\theta}_d)$$

$$w_{d,n} | z_{d,n}, \beta_{d,k=z_{d,n}} \sim \text{Multinomial}(\beta_{d,k=z_{d,n}})$$

STM Procedure

Optimize

Optimize the number of topics using `searchK`

Estimate

Estimate the model using `stm` with the optimal number of topics

Distribution

Generate the Document-Topic Distribution matrix

Significance

Only keep topics that are significantly associated with each time period

Label

Label significantly associated topics

Remove

Remove topics about specific events

Linguistic Inquiry and Word Count (LIWC)

Dictionary-based

Articles — Vectors with sentiment scores

Calculate sentiment scores mainly based on word counts

Dimension	*Abbreviation*	*Example*	*# Words*	*Mean*
I. Standard linguistic dimensions				
Word Count	WC			238.87
% words captured, dictionary words	Dic			73.67
% words longer than six letters	Sixltr			13.57
Total pronouns	Pronoun	I, our, they, you're	70	12.76
First-person singular	I	I, my, me	9	3.97
Total first person	Self	I, we, me	20	4.72
Total third person	Other	she, their, them	22	4.04
Negations	Negate	no, never, not	31	2.98
Articles	Article	a, an, the	3	7.30
Prepositions	Preps	on, to, from	43	11.93
II. Psychological processes				
Affective or emotional processes	Affect	happy, ugly, bitter	615	3.54
Positive emotions	Posemo	happy, pretty, good	261	2.14
Negative emotions	Negemo	hate, worthless, enemy	345	1.39
Cognitive processes	Cogmech	cause, know, ought	312	8.75
Causation	Cause	because, effect, hence	49	1.39
Insight	Insight	think, know, consider	116	2.16

LIWC & t-tests Procedure

Extract

Using LIWC software to extract sentiment scores for each article

Select

Negative and Positive Emotion

Sentiment t-test I

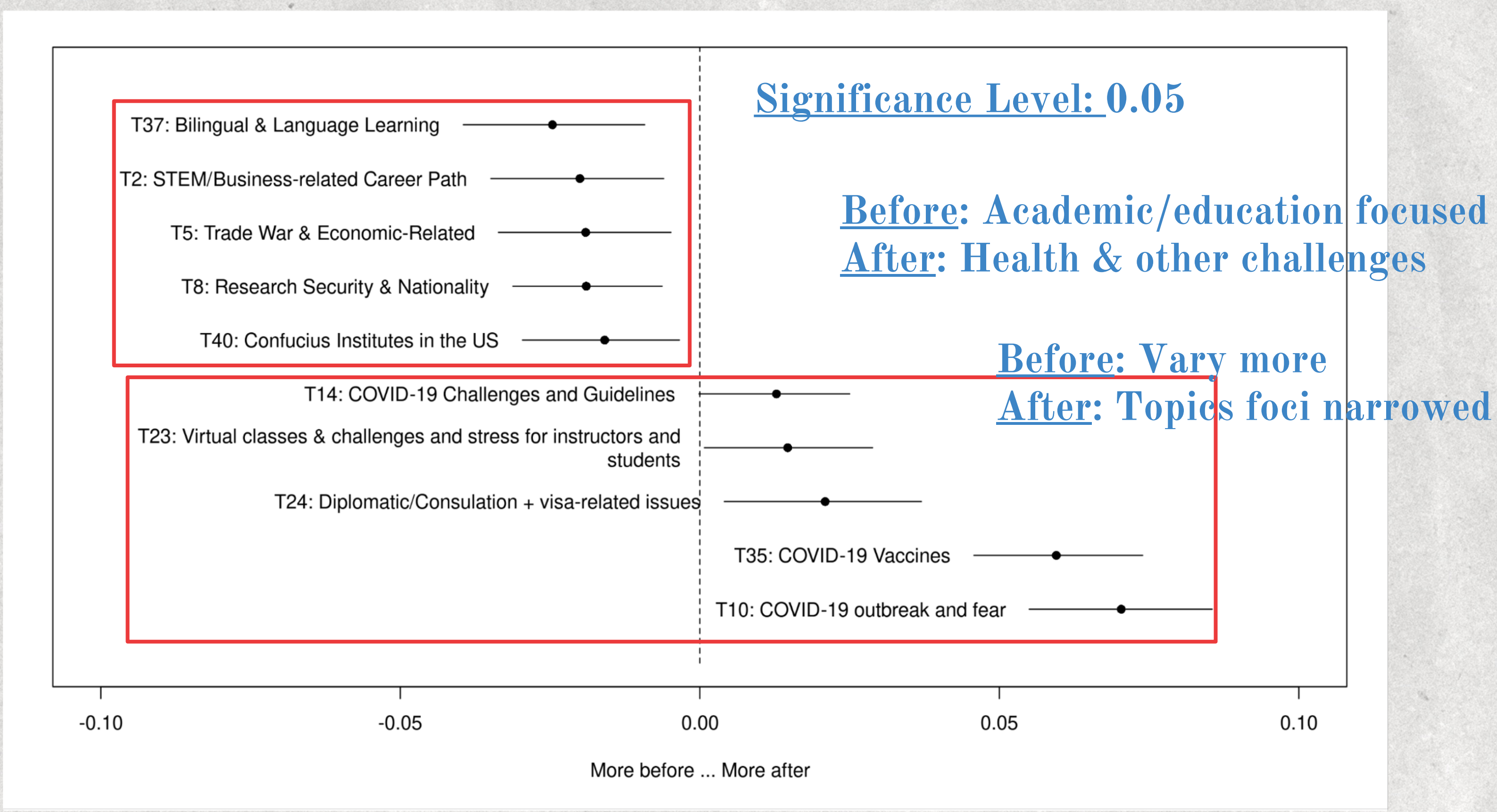
How are the sentiment scores different according to the significance of STM

Sentiment t-test II

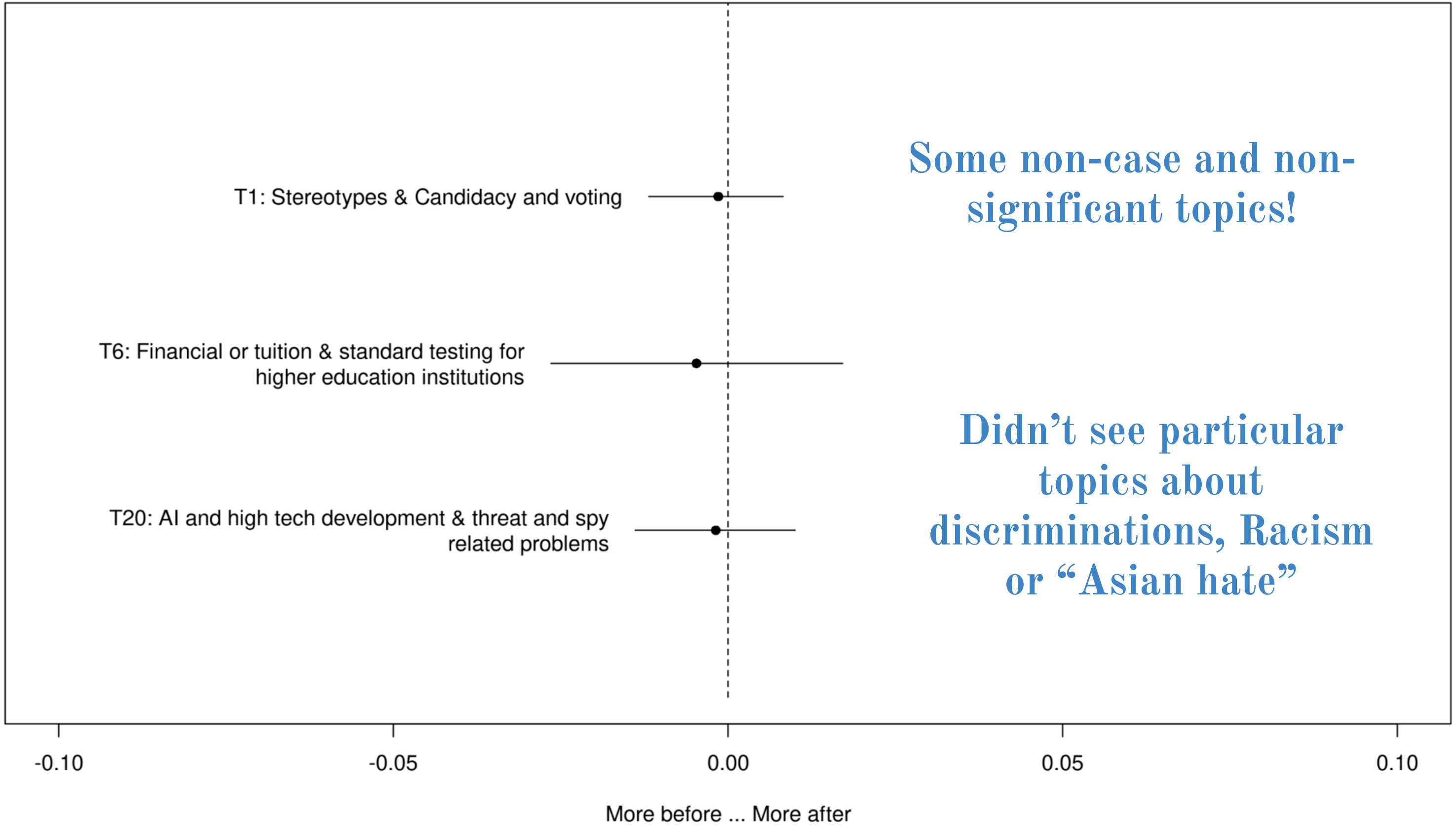
How are the sentiment scores different according to the significance of STM

04 Results

Structural Topic Modeling



Structural Topic Modeling



Some non-case and non-significant topics!

Didn't see particular topics about discriminations, Racism or "Asian hate"

Structural Topic Modeling

Document - topic Distribution Matrix

docnum <int>	Topic1 <dbl>	Topic2 <dbl>	Topic3 <dbl>	Topic4 <dbl>
1	1.641030e-04	3.176269e-03	6.731444e-04	1.163397e-04
2	7.530310e-04	4.121942e-03	1.290759e-03	1.179355e-03
3	1.249003e-04	1.959227e-01	2.134501e-04	4.604302e-05
4	1.249003e-04	1.959227e-01	2.134501e-04	4.604302e-05
5	2.828875e-05	1.786612e-03	7.087004e-05	1.751064e-05
6	2.842451e-05	1.798084e-03	7.122003e-05	1.760658e-05
7	3.706931e-04	8.367938e-02	1.188493e-03	5.168312e-04

Top documents for Topic 1

docnum <int>	Topic1 <dbl>
536	0.9968618225
537	0.9968348168
1421	0.9910672829
1420	0.9910099029
1441	0.9903104084
1442	0.9903104084
633	0.9850753945
634	0.9850753945
1136	0.9573765175

docnum <int>	Topic1 <dbl>	Topic2 <dbl>	Topic3 <dbl>	Topic4 <dbl>
536	0.9968618225	1.138108e-04	9.858072e-05	1.427415e-04
537	0.9968348168	1.149334e-04	9.939461e-05	1.437666e-04

Although each document is consist of topics, it generally only represent 1 or 2 topics

Topic-Sentiment t-test I

Significant Topics

Before

Topics: 37, 2,5,8,40

Articles with proportion more than 0.5 for each topic

140 articles

After

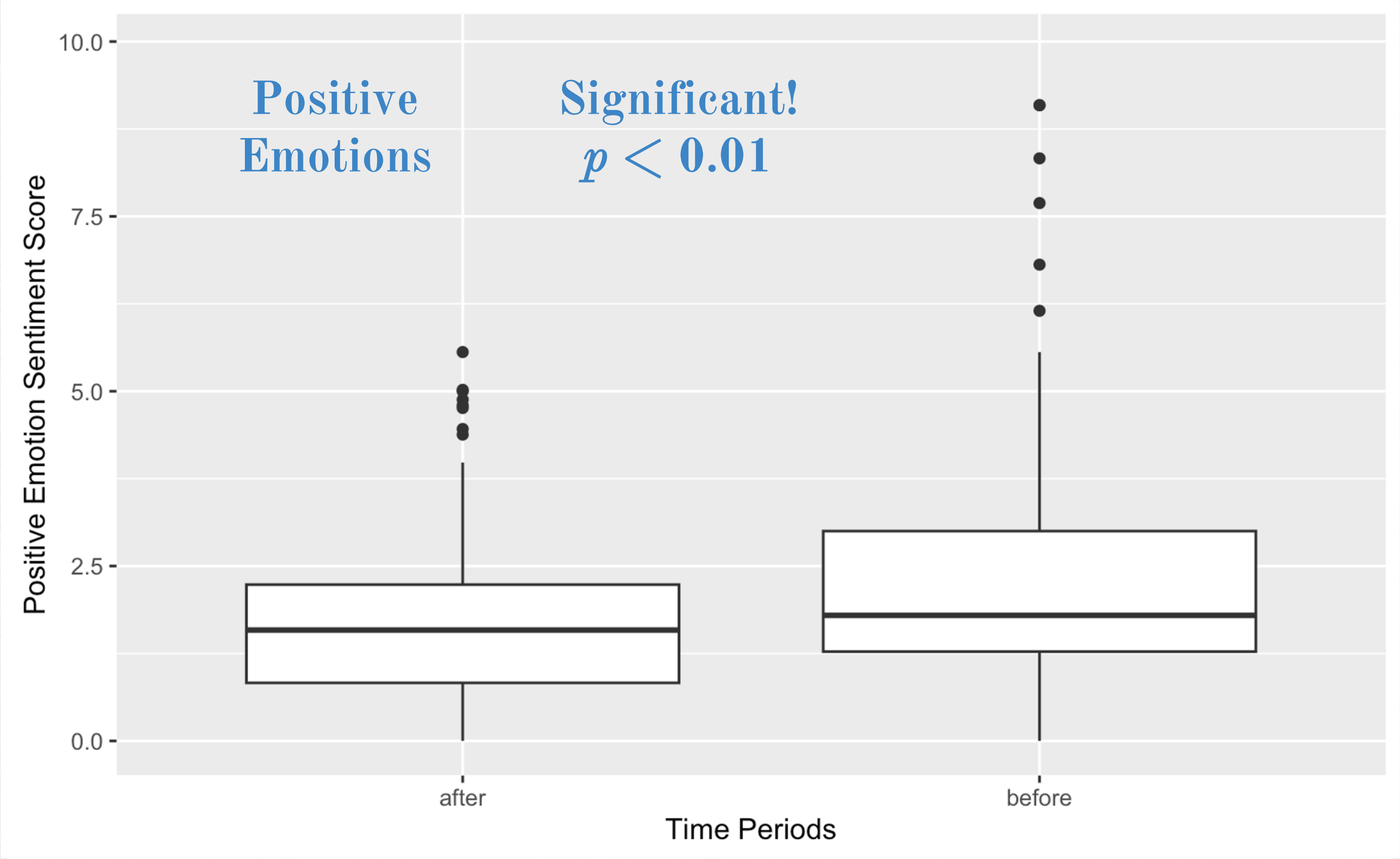
Topics: 14, 23, 24, 35, 10

Articles with proportion more than 0.5 for each topic

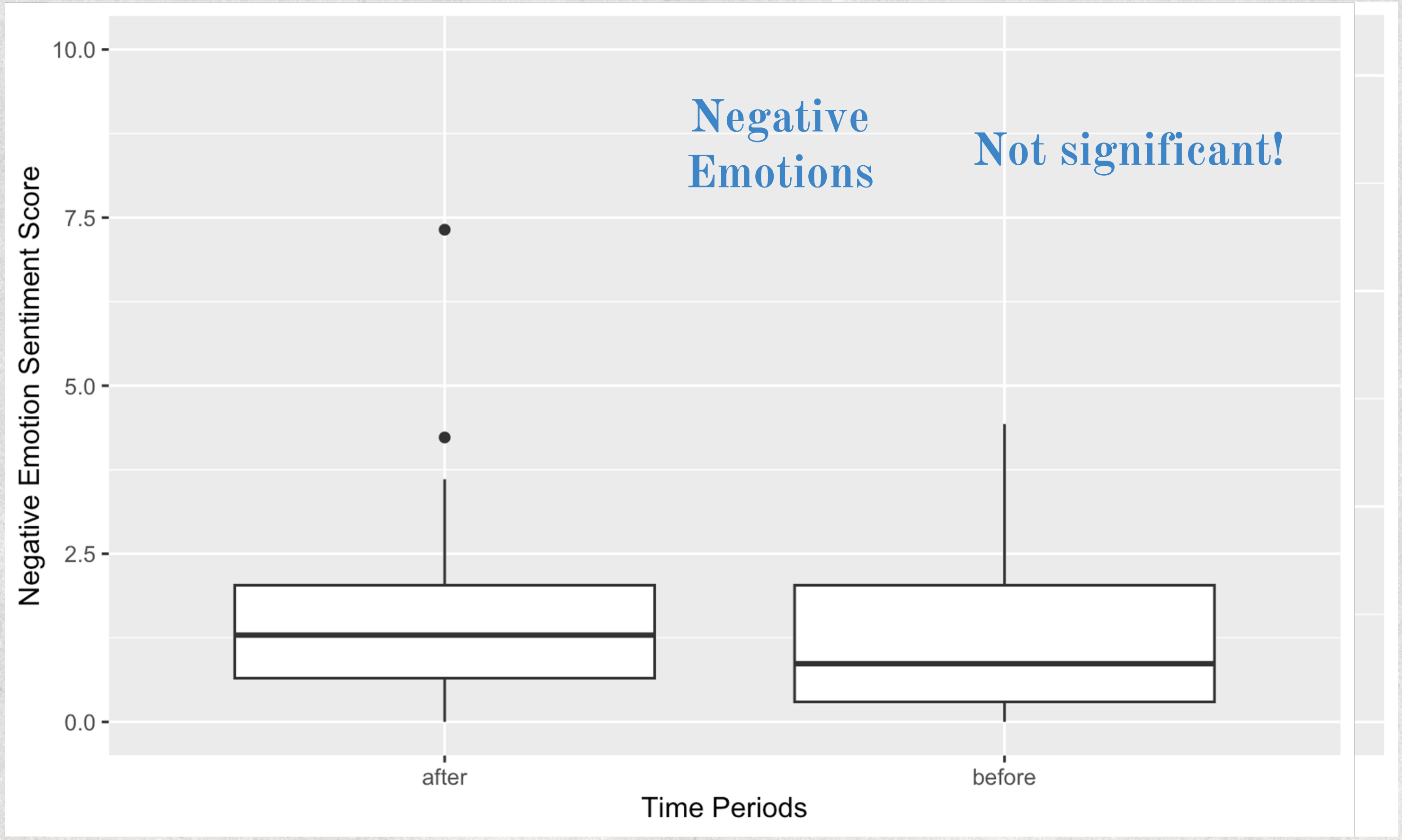
196 articles

Significant differences in average sentiment scores for those two groups?

Topic-Sentiment t-test I



Topic-Sentiment t-test I



Topic-Sentiment t-test II

Non-Significant Topics

1, 6, 13, 15, 16, 17, 20, 22, 26, 34

Articles with proportion more than 0.5 for each topic

Articles from before

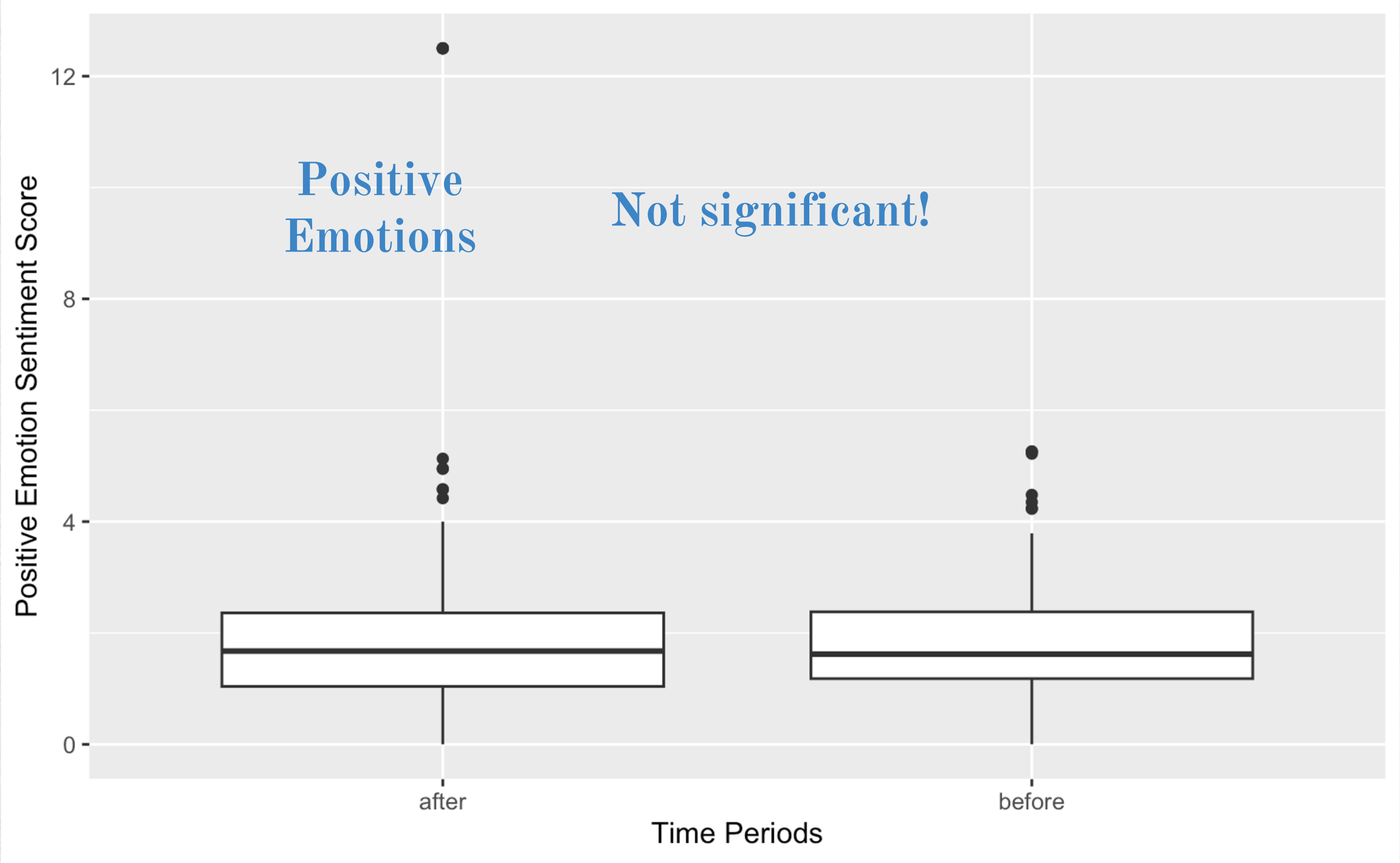
157 articles

Articles from after

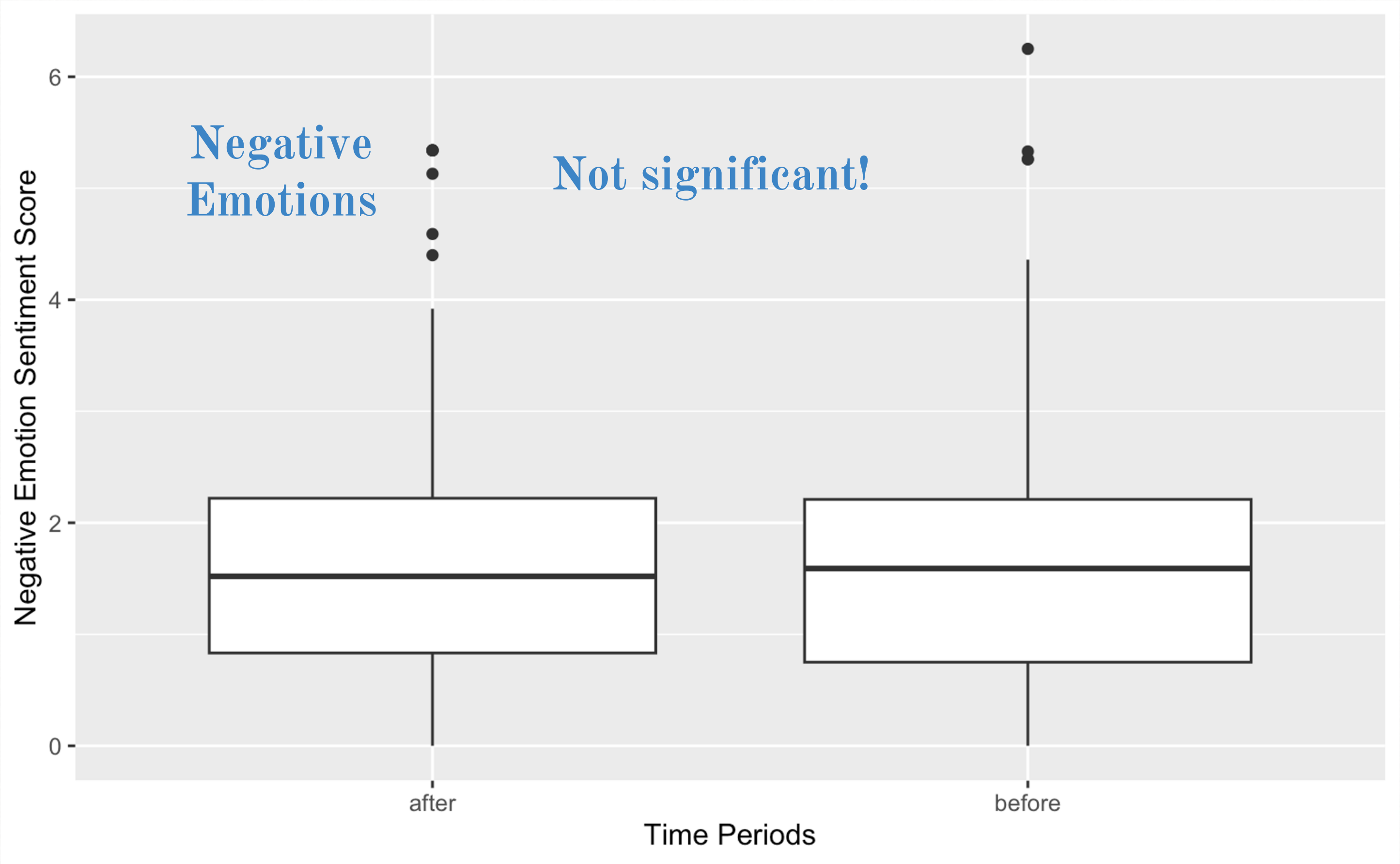
164 articles

Significant differences in average sentiment scores for those two groups?

Topic-Sentiment t-test II



Topic-Sentiment t-test II



05
Discussions
&
Conclusions

From Topic Modeling...

Significant Topics

Before

- More Education/Academic focused
- Generally more variety

After

- Health & Visa related issues focused
- Generally less variety

- ❑ **Health related issues concerns Chinese international students.**
- ❑ **More challenges about visa and travelling arise for Chinese international students.**
- ❑ **However, no topics related to discrimination and racism.**

From Sentiment Analysis & t-tests...

Topics	Sentiment Features	Whether significant difference between time periods
Significant Topics	Positive Emotions	Yes
	Negative Emotions	No
Non-significant topics	Positive Emotions	No
	Negative Emotions	No

Conclusion

For Chinese international students:

- ❑ No big changes on general US news media representations
- ❑ Concerns about discrimination and racism are not obvious in popular news media in US.
- ❑ Maybe don't let COVID-19 be the reason for not studying in the US :)

for a few cases)

06
Limitations
&
Future Directions

Limitations

- ❑ Not enough data!!
- ❑ The article lengths distribution is not the best one, and why removing articles over 2000?
- ❑ No gaps between two time periods!
- ❑ Methodology wise:
  - ❑ Blackboxes: Structural Topic Modeling and LIWC
  - ❑ Labeling and removing topics are human activities
  - ❑ Two sample t-test? Or other approaches?

Future Directions

- ❑ Comparative study of **US news representations vs. Chinese news representations** on Chinese international students
- ❑ **Social media data** rather than news media?
- ❑ Using other state-of-art **text mining methods**
- ❑ Impacts of **other significant events** on news/other media portrayals of Chinese international students

Thank you for listening!

It's an interesting project, and now I feel weird to be done!