Localized Features, Marketized Preference and the Report of Inequality among Chinese Newspapers, 2003-2019

Yizhou Ye & Yi Zhang

1 Dec. 2020

Research aims:

- Figure out the overall and spatiotemporal pattern of the report of inequality and unfairness among Chinese newspaper, including topics and sentiment.
- Linking external data, including official statistics and general social survey features, to examine whether report preference is correlated with localized characteristics.
- Examine whether marketized media would have different preferences from party newspaper and whether marketization would moderate the correlation between localized inequality and reports.

Data

- Database: WiseNews (慧科電子剪報)
- Time: 2003-2019
- Source: all articles published in mainland newspaper
- Keywords: inequality / unfair (不平等/不公平)
- Original Article No.: 198,200
- Cleaned article No.: 172,784

Data preprocessing

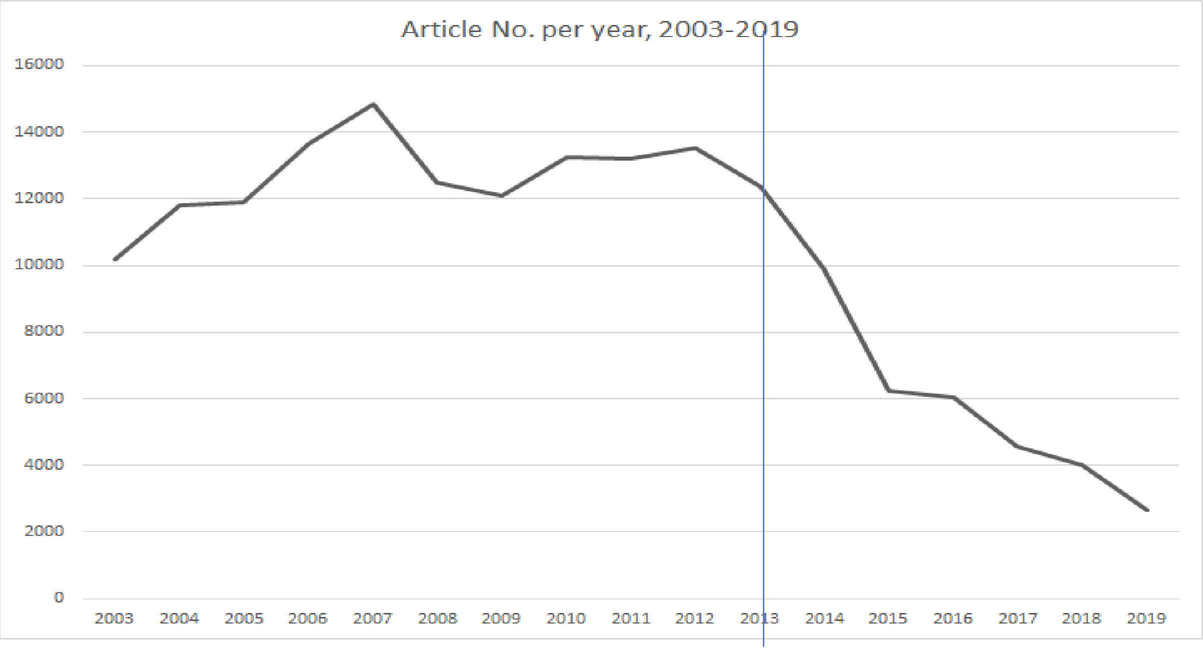
- -1,033 (removed if the length of content is less than 10 characters)
- -4,868 (removed if the content is not completed)
- -727 (removed if with duplicated content or duplicated date and title)
- -1633 (removed if media containing less than 51 articles during 17 years)
- -17,155 misguiding list = [‘上海證券報’,‘中國新聞社’,‘中國證券報’,‘中經網’,‘文摘周刊’,‘證券時報’,‘證券日報’,‘東方體育日報’,‘球報’,‘球迷報’,’體育周報','IT經理世界','通信世界周刊','贏周刊','新華社經濟資訊社','理財1周','互聯網周刊','新民周刊']
- 172,784 articles from 187 media are left
- Transform from traditional Chinese to simplified Chinese

Data preprocessing (Cont'd)

南方都市报	羊城晚报(全国版)	新京报	新快报	广州日报
10091	5488	5109	4618	4380
东方早报	21世纪经济报道	南方日报(全国版)	新民晚报	北京青年报
3952	3884	3840	3597	2984

47,943 / 172,784

≈ 0.27



Data preprocessing (Cont'd)

- Segmentation using Python package pkuseg
- Stopword dictionary generated from “中文常用停用词表”

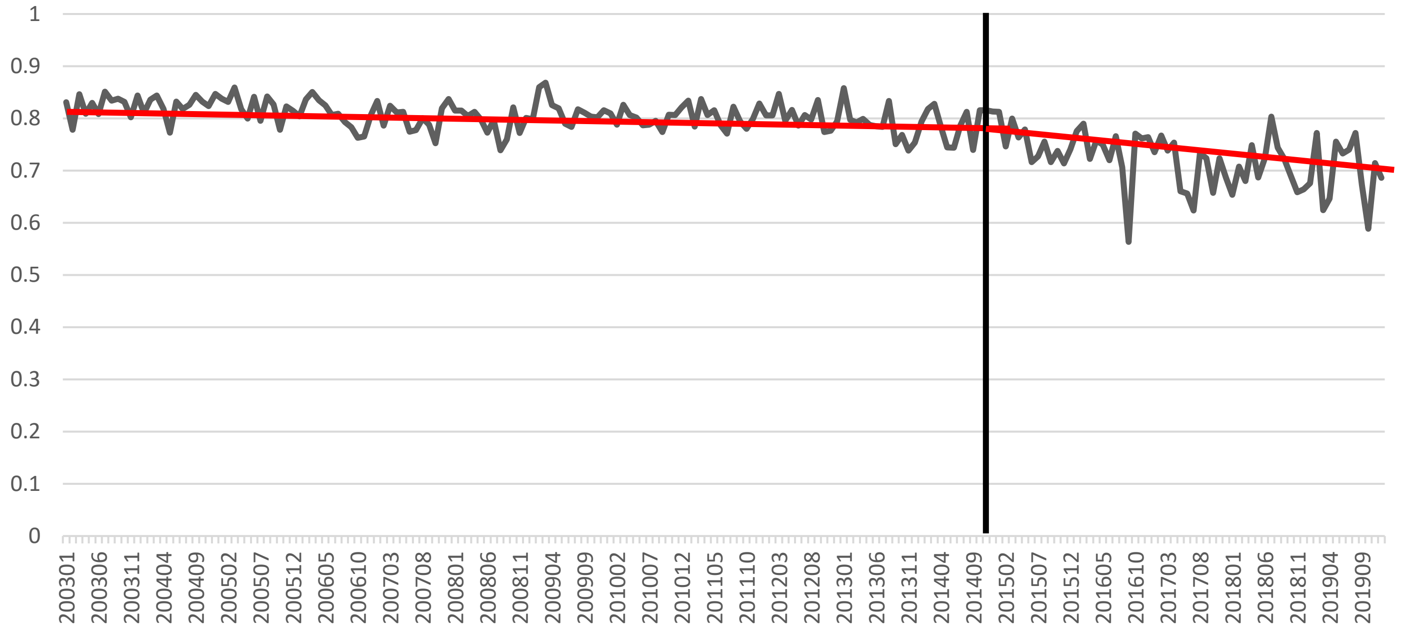
- Add province, city, time and marketized status based on media name
- Add corresponding spatiotemporal features from externally official statistics & CGSS

Analytical strategies

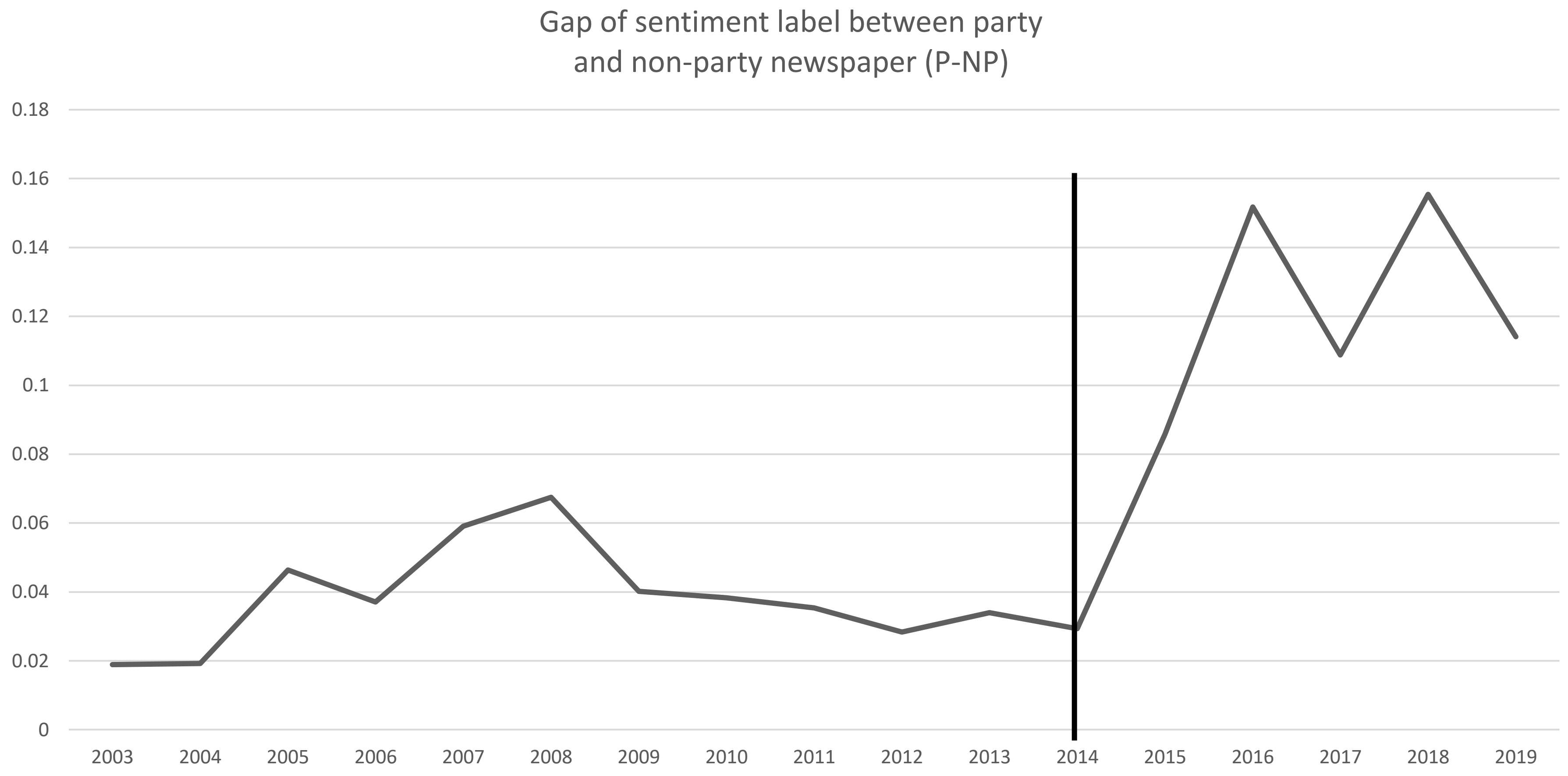
- Computing sentiment score for each article using package Baidu' PaddleHub BiLSTM model.
- Generating topics for each article using package gensim
- For descriptive part,
- By month, draw the changing pattern of overall sentiment score and trend of some topics.
- By year and province, draw the sentiment and dominant topic color card.

Preliminary result

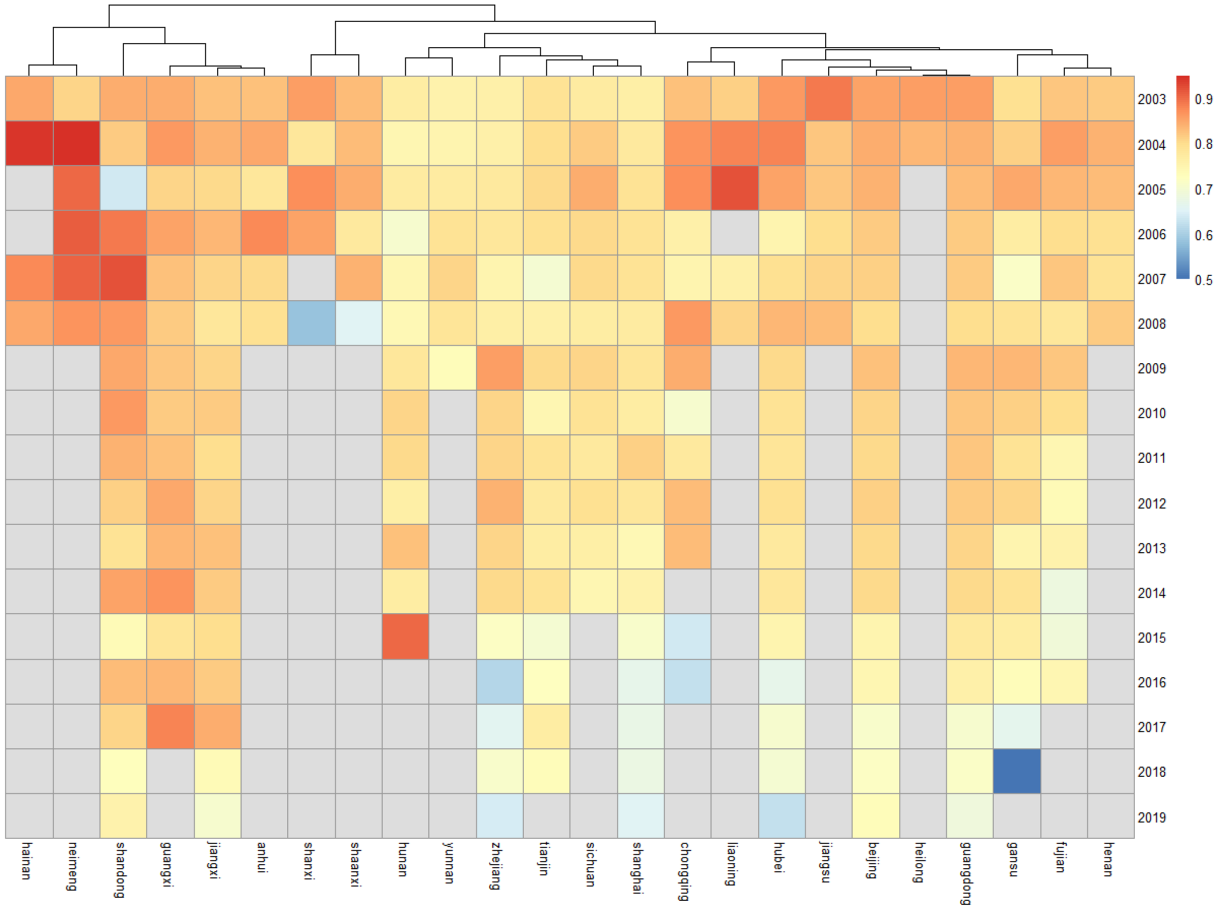
Sentiment label trend, 2003-2019 (1=negative, 0=positive)



Preliminary result (Cont'd)



Preliminary result (Cont'd)



Heat map of aggregated sentiment label score, 2003-2019, by province

Analytical strategies (Cont'd)

- For explanatory part,
- Aggregate article features by year and province as outcome variable.
- Aggregate features from externally official statistics & CGSS. independent variables.
- To further analyze the effect of marketization, add marketized status into the dimension of outcome variable and examine the interaction effect.

Thanks!