Supplemental Appendix
Capturing Clicks: How the Chinese Government Uses Clickbait to Compete for Visibility[*]

Yingdan Lu[†] Jennifer Pan[‡]

May 1, 2020

Abstract

The proliferation of social media and digital technologies has made it necessary for governments to expand their focus beyond propaganda content in order to disseminate propaganda effectively. We identify a strategy of using clickbait to increase the visibility of political propaganda. We show that such a strategy is used across China by combining ethnography with a computational analysis of a novel dataset of the titles of 197,303 propaganda posts made by 213 Chinese city-level governments on WeChat. We find that Chinese propagandists face intense pressures to demonstrate their effectiveness on social media because their work is heavily quantified—measured, analyzed, and ranked—with metrics such as views and likes. Propagandists use both clickbait and non-propaganda content (e.g., lifestyle tips) to capture clicks, but rely more heavily on clickbait because it does not decrease space available for political propaganda. Government propagandists use clickbait at a rate commensurate with commercial and celebrity social media accounts. The use of clickbait is associated with more views and likes, as well as greater reach of government propaganda outlets and messages. These results reveal how the advertising-based business model and affordances of social media influence political propaganda and how government strategies to control information are moving beyond censorship, propaganda, and disinformation.

Keywords: propaganda, clickbait, social media, quantification, China

[*] Our thanks to Mengmeng Guo, Yueyi Li, Zeyu Li, Yuqian Lin, Yitian Liu, Sujing Lyu, Lu Tian, Chenyi Zhang, Jing Zhang, Bowen Zhao and others for superb research assistance; Erin Baggott, Angèle Christin, Charles Crabtree, Kingwa Fu, Holger Kern, Jon Krosnick, Carlo Horz, Dan Mattingly, Tamar Mitts, Rich Nielsen, Molly Roberts, Wei Wei, and others for helpful comments and suggestions; and to the NSF (Award #1831481) and the Stanford Graduate Fellowship for research support. Replication materials can be found at https://doi.org/10.7910/DVN/TALJOT.

[†] PhD Candidate, Department of Communication, Stanford University, Email: yingdan@stanford.edu

[‡] Assistant Professor, Department of Communication, Stanford University. Email: jp1@stanford.edu.

Contents

A. Collecting Data from WeChat Official Accounts

Figure A1 shows the year of creation of the 213 city-level government WeChat Official Accounts in our dataset. Accounts expanded rapidly in 2013 after the State Council issued "Opinions of the General Office of the State Council on Further Strengthening Government Information Disclosure in Response to Social Concerns and Enhancing Government Credibility" (国务院办公厅关于进一步加强政府信息公开回应社会关切提升政府公信力的意见), which required all local governments to establish social media accounts.[1] Since 2015, the increase in government accounts has slowed.

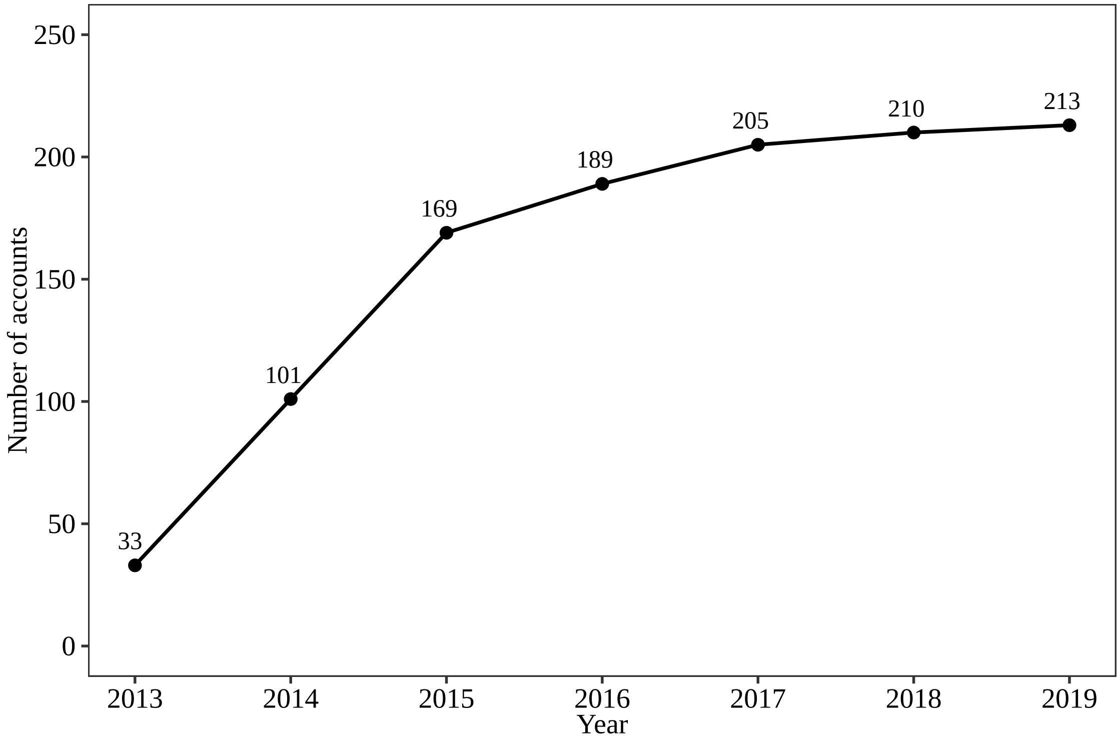

FIGURE A1. CUMULATIVE NUMBER OF CITY GOVERNMENT WECHAT OFFICIAL ACCOUNTS BY YEAR

Posts made by Official Accounts are publicly viewable upon login, but WeChat has implemented numerous anti-bot and anti-spider constraints that make automated scraping of Official Accounts unfeasible unless one has access to a large number of WeChat accounts that can be used for login credentials. The only large-scale collection of WeChat Official Account data is WeChatscope, a project developed and implemented by researchers at the Journalism and Media Studies Center of the University of Hong Kong.[2] The WeChatscope project is focused on tracking censorship on WeChat, so government Official Accounts, which are relatively unlikely to be censored, were not main targets of its data collection.

We find a solution to the challenge of collecting data from WeChat by developing automated scraping algorithms to collect the titles of posts from Sogou Weixin,[3] a platform

[1] See http://www.gov.cn/zhengce/content/2013-10/18/content_1219.htm (accessed Sept. 28, 2019).

[2] See https://wechatscope.jmsc.hku.hk/.

[3] See https://weixin.sogou.com/

developed by Sogou that is the default search engine for WeChat. Sogou Weixin shows the titles of WeChat posts. Titles are exactly what we require because that is the content users can see before making the decision of whether or not to click on a government post. By using Sogou Weixin, we also avoid the ethical risks of either buying WeChat login credentials or recruiting confederates willing to lend us their personal account credentials.[4] In our data collection process, we followed the terms of services of weixin.sogou.com, and adhered to requirements outlined in their robots.txt file.

We first collected the names and IDs of all WeChat Government Official Accounts by manually searching WeChat. Then, we built our crawler and ran the algorithms repeatedly to ensure we collected all posts in the time period of interest. Finally, we pre-processed the data to remove duplicates, generating a dataset of 197,303 titles.

B. Inter-coder Reliability

We computed inter-coder agreement on a random sample of 764 titles for each variable. Table A1 shows the percent agreement for all hand-coded variables. The first section are clickbait strategies that entailed hand-coding. The second section are hand-coded emotional and vision appeals. The last section are hand-coded topic controls that are included in the regression analysis of views and likes.

TABLE A1. INTER-CODER RELIABILITY FOR ALL VARIABLES

	Agreement (Percent)	Agreement (Number)	Disagreement (Number)	Total Cases (Number)
Listicle	94.2%	720	44	764
General noun	90.7%	693	71	764
Hyperbolic word	96.5%	737	27	764
Slang	96.5%	737	27	764
Joy	81.7%	624	140	764
Pride	80.6%	616	148	764
Anger	96.6%	738	26	764
Fear	88.5%	676	88	764
Warmth	99.0%	756	8	764
Vision	80.1%	612	152	764
Central ideology	96.7%	739	25	764
Leader activities	82.6%	631	133	764
Government administrative activities	80.2%	613	151	764
Guidance unrelated to politics	83.9%	641	123	764
Job openings	80.2%	613	151	764
Local claims to fame	83.8%	640	124	764

[4] WeChat login credentials, as required by Chinese law, must be linked to real identities and mobile phone numbers.

C. Non-Government WeChat Official Accounts

To compare the rate of clickbait between government and non-government WeChat Official Accounts, we collected titles from three commercial and celebrity accounts in December, 2019. We selected the most popular WeChat Official Accounts that represent three different type of ownership structure and affiliation—an individual influencer account, a company account, and a commercial media account. Popularity was based on WeChat Official Account rankings from Qingbo Big Data Corporation.[5]

The individual influencer account is *Zhanhao* (占豪), run by a well-known investment expert and writer Hao Zhan and famous for its commentaries on international relations and financial investment. The company account is *Dingxiang Doctor (*丁香医生*)*, a healthcare-oriented account managed by Dingxiangyuan Company, which circulates science articles related to health, life and digital technology. The commercial media account is *Lifeweek (*三联生活周刊*)*, the official account of *Lifeweek* magazine, which publishes articles related to culture, entertainment, politics and technology. We collected all titles from these three accounts made between Feb 25, 2019 and May 25, 2019. In total, we collected 1,607 titles, and all titles were coded by the same human coders who were trained to code the government titles. Table A2 shows the basic characteristics of the non-government accounts.[6]

TABLE A2. STATISTICS OF NON-GOVERNMENT OFFICIAL ACCOUNTS AND POSTS

Account		Estimated Followers	Titles Collected	Avg. Reads/Article	Avg. Likes/Article
name	Type	(Millions)	(Number)	(Number)	(Number)
Zhanhao	Influencer	12.52	168	100,000	11,118
Dingxiang Doctor	Company	9.95	803	94,500	1,036
Lifeweek	Media	5.44	636	54,825	357

D. Identifying Clickbait

Our procedures for identifying clickbait titles are detailed below.

1. **Quotation marks**: we counted the number of exclamation marks, question marks, and ellipsis marks by writing a function in python to detect these marks. We included both Chinese-style and English-style marks (e.g. 。。。。。。 is the Chinese-style ellipsis mark, while ... is the English-style one).

2. **Hyperbolic words, slang**: to identify hyperbolic words and slang, human coders created an open-ended dictionary for each variable by reading all sampled titles. In total, 479 words were included in the dictionary for hyperbolic words, and 1,857 words were included for slang.[7] Then we applied the dictionary to our sample titles to find all titles containing words from the dictionary. Finally, human coders reviewed all identified titles to remove the false positives.

[5] See https://baijiahao.baidu.com/s?id=1635295121541068562wfr=spiderfor=pc (accessed in Dec 20, 2019).

[6] Follower data comes from http://data.xiguaji.com/ (accessed in January 18, 2019).

[7] For slang, we also combined the words derived from open-coding with existing lists of such words from the Sogou Cell Thesaurus: https://pinyin.sogou.com/dict/.

3. **Pronoun**: we used parts-of-speech tagging to tag pronouns. We started by applying the Stanford POS tagger (Toutanova et al. 2003; Tseng et al. 2005) on all segmented titles. We used JiebaR package (Qin and Wu 2019) for segmentation. We do not remove stop-words since most Chinese stop-word dictionaries contain pronouns. We classify all titles with pronoun (#PN) tags as containing pronouns. In total, 67 different pronouns were included.

4. **Fixed-phrase patterns**: We conducted n-gram analysis by using the NLTK package for Python (Bird et al. 2009) on all scraped titles, and extracted 21,6234 bi-grams, 15,4359 tri-grams, and 91,478 four-grams. After excluding rare n-grams (those that occur less than 50 times in the entire dataset), human coders coded a random sample of n-grams to determine whether they contained fixed-pattern phrases relevant for clickbait. This random sample included 1,362 bi-grams, 327 tri-grams and 102 four-grams from stratified sampling on gram type. Two human coders coded these sampled n-grams to determine whether they are common fixed-pattern clickbaity phrases or not, achieving intercoder reliability of 92.7%. We used this human-coded subset to determine a n-gram frequency threshold—the total number of times the n-gram occurs divided by the total number of titles—that would maximize recall. Our thresholds ranged from 0.0002 to 0.0065, and we found we could achieve recall of 0.9 at the 0.0003 threshold. We then applied this frequency threshold to all the extracted n-grams. Finally, one human coder went through all of the output of this step and manually removed n-grams specific to propaganda (e.g., ideological slogans) and n-grams found in other clickbait dictionaries we had constructed (e.g., slang). After that, 17 n-grams remained, and we then combined these with phrases identified in prior studies (Wei and Wan 2017). In total, our dictionary contained 54 fixed-pattern words. Then, we applied this dictionary to our sampled titles to find all titles containing words from the dictionary.

E. Topic Modeling

Figure A2 plots the hold-out likelihood, semantic coherence, residuals, and lower bound of our topic model with topics ranging from 10 to 80. This shows that setting the topic number equal to 30 balances between maximizing held-out likelihood, semantic coherence, and minimizing residuals. Of the 30 topics, As Figure A3 shows, we were able to label 29 topics after reading the top ten titles associated with the topic and the most frequently occurring words.

FIGURE A2. TOPIC SELECTION

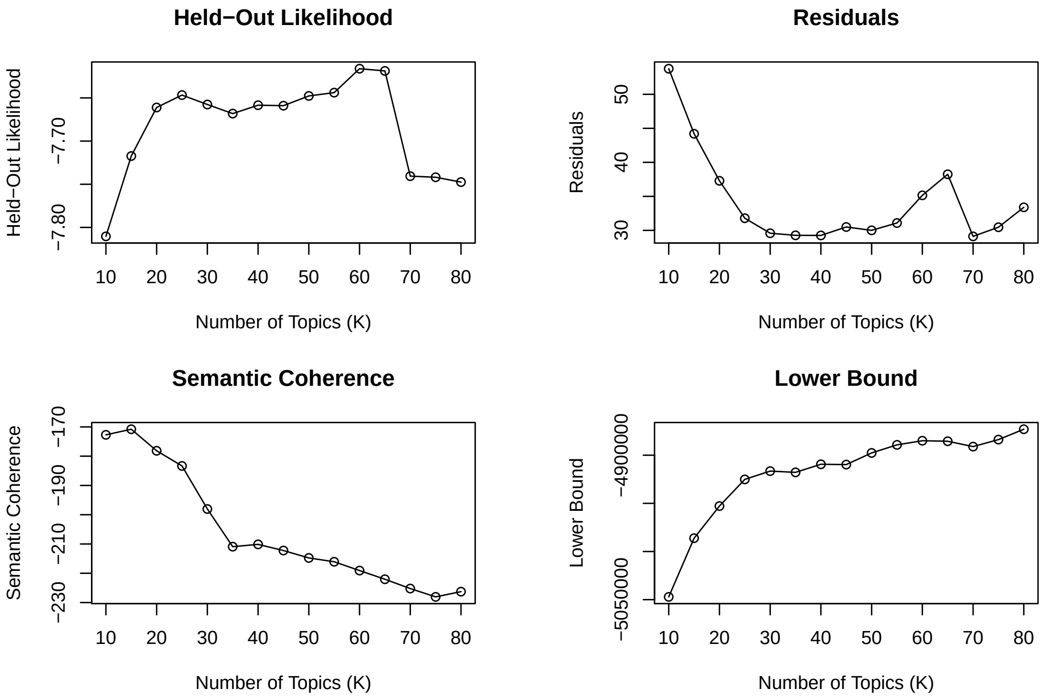

TABLE A3. TOPICS AND LABELS

ID	Topic Label	Words (Chinese)	Words (English translation)
1	Social security, welfare policies	元, 项目, 万, 亿, 金, 保, 房	Yuan, project, ten thousand, hundred million, gold, insure, house
2	National media coverage of locality	关注, 武汉, 聚焦, 央, 视, 新闻, 人民	attention, Wuhan, focus, central, vision, news, citizen
3	Model citizen stories	多, 双, 尔, 鄂, 斯, 心, 暖	more, pair, Er, E, Si, Yichun, heart, warmth
4	Extreme weather	今日, 南, 火, 钦州, 开, 常州, 运	today, south, fire, Qinzhou, open, Changzhou, move
5	Subway, highway, rail construction	路, 大, 走, 年, 红, 桥, 高速	road, big, go, year, red, bridge, high-speed
6	Crime and punishment	车, 曝光, 动, 主, 电, 交警, 注意	car, exposure, move, main, electricity, traffic police, attention
7	Motivational messages	新, 城市, 时代, 生活, 文明, 正, 规	new, city, era, life, civilization, positive, rule
8	Customs and festivals	微, 信, 市民, 宿, 份, 送, 单	micro, letter, citizen, live, share, send, single
9	Notable events	这, 事, 今年, 件, 定, 干, 关	this, thing, this year, item, fix, do, close
10	Local government meetings	会, 会议, 市, 召开, 市委, 工作, 学习	meet, meeting, city, convene, municipal committee, work, study
11	Local cultural events	节, 今天, 活动, 文化, 湖州, 树, 场	festival, today, activity, culture, Huzhou, tree, space
12	Weather forecast	天气, 风, 天, 雨, 未来, 迎, 大	weather, wind, day, rain, future, welcome, big
13	College entrance examination	报, 服务, 举, 考, 进, 高考, 清远	register, service, raise, exam, enter, entrance exam, Qingyuan
14	Propaganda slogans for local development	发展, 高, 经济, 质量, 产业, 要, 商	development, high, economy, quality, industry, require, business
15	Local history, culture	阳, 阜, 听, 故事, 德, 揭, 江	sun, Fu, listen, story, virtue, uncover, lake
16	Advice on daily life	办, 证, 卡, 分钟, 太, 跑, 办理	office, certificate, card, minute, too, run, transact
17	Local wins in nat'l/int'l competition	赛, 第一, 集, 征, 亚, 大, 世界	competition, first place, collect, victory, Asia, big, world
18	Local government activities	市, 发布, 重, 最新, 招聘, 磅, 问题	city, release, repeat, latest, recruitment, pound, question
19	cannot label	图, 一, 生, 读, 成都, 健康, 养	image, one, birth, read, Chengdu, health, raise
20	Local recognition by upper-level government	点, 省, 全国, 国家, 名单, 全, 州	point, province, nationwide, nation, name list, all, state
21	Six-city policy	城, 建, 创, 港, 宣, 文, 之	city, build, create, harbor, propagate, language, go
22	Local claims to fame	网络, 机场, 航, 试, 班, 飞, 首	Internet, airport, sail, try, class, fly, first
23	Tourism and travel	旅游, 美, 区, 花, 景, 山, 园	travel, beautiful, district, flower, scenery, mountain, garden
24	Local news	快, 来, 家, 有, 看看, 你, 变	fast, come, home, have, look, you, change
25	Public transportation, travel advisories	交, 起, 公, 日, 条, 交通, 注意	cross, from, public, day, stripe, traffic, attention
26	Local officials' activities	工作, 县, 调, 贫, 研, 书记, 市	work, county, tune, poverty, research, party secretary, city
27	Local implementation of central propaganda	改革, 开放, 党, 周年, 新, 中国, 系列	reform, open, party, anniversary, new, China, series
28	Local implementation of nat'l anti-gang campaign	项, 扫黑, 除恶, 专, 行动, 督, 开展	item, sweep-crime, eradicate-evil, specific, action, supervision, launch
29	Advice on healthy living and safety	人, 吃, 知道, 注意, 别, 提醒, 佛山	people, eat, know, attention, don't, warn, Foshan
30	Local construction projects	市, 州, 区, 两, 发, 中心, 个	city, state, district, two, send, center, entity

F. Analysis of Views and Likes

Figure A3 shows the histogram of views (upper) and likes (lower). There is a clear over-dispersion in in views, which is why we use a negative binomial model. For likes, there is over-disperson and 19.5% of likes are zeros. Given this structure, we use a zero-inflated negative binomial model. Table A4 shows the full regression results where the outcome is views. Table A5 shows the full regression results where the outcome is likes.

FIGURE A3. DISTRIBUTION OF VIEWS AND LIKES OF SAMPLE TITLES

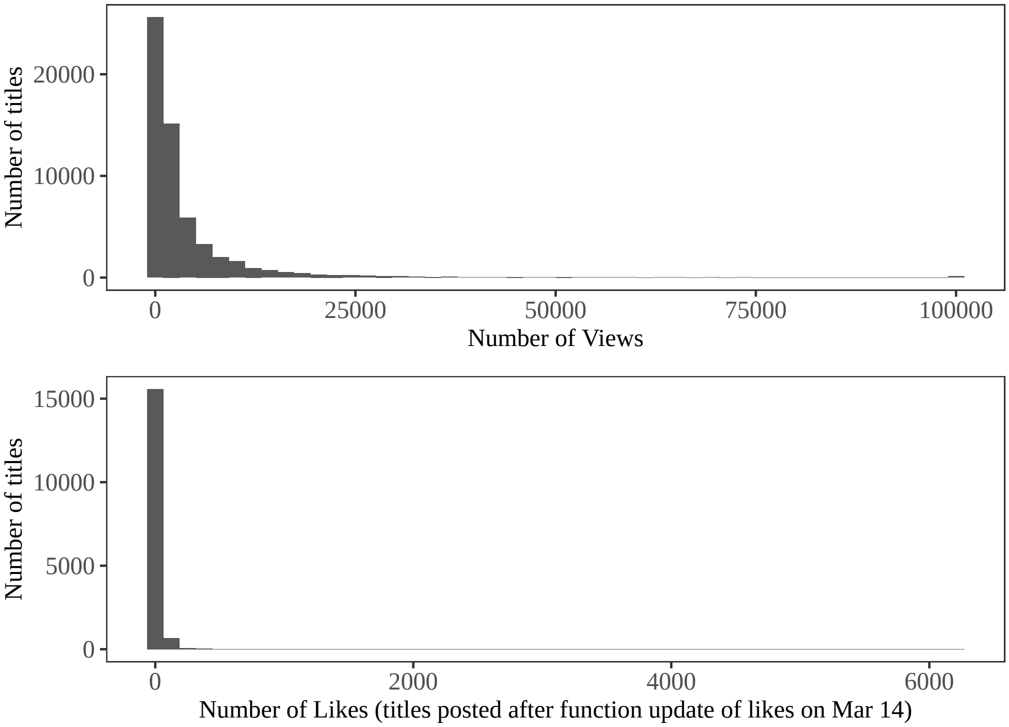

TABLE A4. PREDICTORS OF VIEWS

		(1)	(2)	(3)
Clickbait strategies	Hyperbolic words	0.244***	0.243***	0.259***
		(0.018)	(0.018)	(0.018)
	Exclamation marks	0.229***	0.238***	0.245***
		(0.010)	(0.010)	(0.010)
	Ellipsis marks	0.039***	0.041***	0.046***
		(0.014)	(0.014)	(0.014)
	Fixed phrases patterns	0.012	0.012	0.016
		(0.012)	(0.012)	(0.012)
	Listicles	0.008	0.010	0.016
		(0.018)	(0.018)	(0.018)
	Question marks	−0.006	−0.009	−0.005
		(0.015)	(0.015)	(0.015)
	Pronouns	−0.074***	−0.073***	−0.061***
		(0.010)	(0.010)	(0.010)
	Slang	−0.080***	−0.079***	−0.063***
		(0.013)	(0.014)	(0.014)
	General nouns	−0.083***	−0.082***	−0.086***
		(0.020)	(0.020)	(0.020)
Appeals	Pride		0.055**	0.086***
			(0.026)	(0.026)
	Joy		−0.077***	−0.070***
			(0.016)	(0.016)
	Anger		−0.082	−0.081
			(0.065)	(0.065)
	Fear		−0.186***	−0.180***
			(0.044)	(0.044)
	Warmth		−0.220***	−0.213***
			(0.048)	(0.048)
	Vision		0.046**	0.041**
			(0.018)	(0.018)
Topic controls	Central ideology			−0.332***
				(0.025)
	Job openings			0.245***
				(0.036)
	Government administrative activities			0.198***
				(0.020)
	Leader activities			0.055***
				(0.020)
	Local claims to fame			−0.056***
				(0.012)
	Guidance unrelated to politics			−0.028**
				(0.012)
Constant		6.539***	6.538***	6.519***
		(0.063)	(0.063)	(0.063)
City FE		YES	YES	YES
Observations		58,411	58,411	58,411

Note: *p<0.1; **p<0.05; ***p<0.01

Table A5. Predictors of Likes

		(1)	(2)	(3)
Clickbait strategies	Slang	0.17^{***}	0.16^{***}	0.12^{***}
		(0.02)	(0.02)	(0.02)
	Hyperbolic words	0.13^{***}	0.12^{***}	0.06^{*}
		(0.02)	(0.02)	(0.02)
	Pronouns	0.05^{***}	0.05^{***}	0.03^{*}
		(0.01)	(0.01)	(0.01)
	Question marks	-0.04	-0.03	-0.03
		(0.02)	(0.02)	(0.02)
	Ellipsis marks	-0.05^{*}	-0.05^{**}	-0.05^{*}
		(0.02)	(0.02)	(0.02)
	Exclamation marks	-0.08^{***}	-0.10^{***}	-0.08^{***}
		(0.01)	(0.01)	(0.01)
	General nouns	-0.08^{*}	-0.08^{*}	-0.07^{*}
		(0.03)	(0.03)	(0.03)
	Fixed phrase patterns	-0.10^{***}	-0.10^{***}	-0.07^{***}
		(0.02)	(0.02)	(0.02)
	Listicles	-0.13^{***}	-0.14^{***}	-0.15^{***}
		(0.03)	(0.03)	(0.03)
Appeals	Warmth		0.70^{***}	0.67^{***}
			(0.08)	(0.07)
	Anger		0.10	0.10
			(0.08)	(0.08)
	Joy		0.09^{***}	0.04
			(0.02)	(0.02)
	Pride		0.09	-0.02
			(0.08)	(0.08)
	Fear		-0.12	-0.08
			(0.07)	(0.07)
	Vision		-0.08^{**}	-0.07^{*}
			(0.03)	(0.03)
Topic controls	Central ideology			0.26^{***}
				(0.04)
	Job openings			-0.80^{***}
				(0.05)
	Government administrative activities			-0.27^{***}
				(0.03)
	Leader activities			0.05
				(0.04)
	Local claims to fame			0.23^{***}
				(0.02)
	Guidance unrelated to politics			-0.16^{***}
				(0.02)
Logged Views		0.85^{***}	0.85^{***}	0.85^{***}
		(0.01)	(0.01)	(0.01)
Constant		-4.24^{***}	-4.24^{***}	-4.26^{***}
		(0.06)	(0.06)	(0.06)
City FE		YES	YES	YES
Observations		16,384	16,384	16,384

Note: *p<0.1; **p<0.05; ***p<0.01

References

Bird, S., E. Klein, and E. Loper (2009). *Natural language processing with Python: analyzing text with the natural language toolkit.* " O'Reilly Media, Inc.".

Qin, W. and Y. Wu (2019). *jiebaR: Chinese Text Segmentation*. R package version 0.10.99.

Toutanova, K., D. Klein, C. D. Manning, and Y. Singer (2003). Feature-rich part-of-speech tagging with a cyclic dependency network. In *Proceedings of the 2003 conference of the North American chapter of the association for computational linguistics on human language technology-volume 1*, pp. 173–180. Association for Computational Linguistics.

Tseng, H., D. Jurafsky, and C. Manning (2005). Morphological features help pos tagging of unknown words across language varieties. In *Proceedings of the fourth SIGHAN workshop on Chinese language processing*.

Wei, W. and X. Wan (2017). Learning to identify ambiguous and misleading news headlines. *IJCAI International Joint Conference on Artificial Intelligence*, 4172–4178.