中国人民大学
RENMIN UNIVERSITY OF CHINA

社交媒体如何影响青年婚育意愿？

—— 基于婚育观调查数据与微博文本数据的讨论

中国人民大学人口学系
博士生 闫誉腾
yanyuteng@ruc.edu.cn

2023/3/2

摘要

文献背景：社交媒体对观念分化、极化的研究已经是学术热点，但已有对社交媒体如何影响婚育观念的实证研究仍然很少。

数据与方法：本文以微博媒介为代表，基于2021年大学生婚育观调查数据与2022年6、7月微博婚育文本数据，使用工具变量法、百度飞桨（PaddlePaddle）ERNIE-Tiny 深度学习模型，对信息机制进行了实证探索。

研究发现：第一，工具变量分析认为，更高的微博使用频率与更低的青年群体婚育意愿强相关；第二，机制分析认为，更高的微博使用频率，可能通过负面新闻的暴露程度降低婚育意愿，具体而言，微博对婚育意愿的降低可能存在两种途径，一是焦点性负面事件的爆炸性冲击、二是日常负面话题的情感漩涡现象。

报告大纲

一． 文献回顾与研究假设

二． 数据与方法

三． 基于调查数据的模型结果

四． 基于文本数据的机制分析

五． 结论与不足

一. 文献回顾与研究假设

1. 互联网与观念极化的相关理论

- **社会心理学理路：**社会比较理论（Social Comparison Theory）、劝服辩论理论或有力论据理论（Persuasive Argumentation Theory）、选择性接触理论（Select Exposure Theory）、社会背书理论（Social Endorsements）；
- **传播学理路：**沉默的螺旋（The Spiral of Silence）、意见领袖理论等。

- **已有研究的理论问题，**忽视了**议程设置理论**。提出假设一。

假设一：微博使用频率会影响婚育意愿，并通过负面信息的暴露机会降低婚育意愿。

一. 文献回顾与研究假设

2. 互联网与观念极化的相关实证机制

- **微观方面。**陈福平和许丹红（2017）的同质偏好机制。**（个体选择）**
- **宏观方面。**虞鑫和许弘智（2019）与葛岩等（2020）的意见领袖机制。**（结构影响）**

- **已有研究的机制问题，**忽视了信息本身的属性。提出假设二、假设三。**（微观—信息机制）**

假设二：突发的、焦点性负面事件，会对婚育意愿产生巨大负面冲击；

假设三：负面事件本身具有情感的漩涡性，短期内随着时间的增加，舆论的负面情绪有加剧趋势。

报告大纲

一． 文献回顾与研究假设

二． 数据与方法

三． 基于调查数据的模型结果

四． 基于文本数据的机制分析

五． 结论与不足

二. 数据与方法

1. 数据

- 2021年中国大学生婚育观调查（样本量9,694份）
- 2022年微博文本数据（6月关键词，样本量126,332条）
- 2022年微博文本数据（3月至7月话题，样本量12,439条）

二. 数据与方法

2. 关键方法

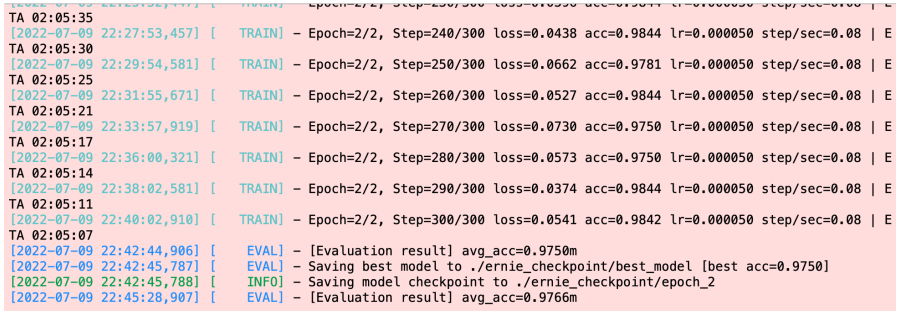
- 2021年中国大学生婚育观调查

  同群微博使用频率作为工具变量

- 2022年微博文本数据（关键词）

  百度飞桨（PaddlePaddle） ERNIE-Tiny 深度学习模型作为情绪识别

  测试集正确率达 97%



二. 数据与方法

Computer Science > Computation and Language

[Submitted on 4 Jun 2021]

ERNIE-Tiny : A Progressive Distillation Framework for Pretrained Transformer Compression

Weiyue Su, Xuyi Chen, Shikun Feng, Jiaxiang Liu, Weixin Liu, Yu Sun, Hao Tian, Hua Wu, Haifeng Wang

Pretrained language models (PLMs) such as BERT adopt a training paradigm which first pretrain the model in general data and then finetune the model on task-specific data, and have recently achieved great success. However, PLMs are notorious for their enormous parameters and hard to be deployed on real-life applications. Knowledge distillation has been prevailing to address this problem by transferring knowledge from a large teacher to a much smaller student over a set of data. We argue that the selection of thee three key components, namely teacher, training data, and learning objective, is crucial to the effectiveness of distillation. We, therefore, propose a four-stage progressive distillation framework ERNIE-Tiny to compress PLM, which varies the three components gradually from general level to task-specific level. Specifically, the first stage, General Distillation, performs distillation with guidance from pretrained teacher, gerenal data and latent distillation loss. Then, General-Enhanced Distillation changes teacher model from pretrained teacher to finetuned teacher. After that, Task-Adaptive Distillation shifts training data from general data to task-specific data. In the end, Task-Specific Distillation, adds two additional losses, namely Soft-Label and Hard-Label loss onto the last stage. Empirical results demonstrate the effectiveness of our framework and generalization gain brought by this http URL particular, experiments show that a 4-layer ERNIE-Tiny maintains over 98.0%performance of its 12-layer teacher BERT base on GLUE benchmark, surpassing state-of-the-art (SOTA) by 1.0% GLUE score with the same amount of parameters. Moreover, ERNIE-Tiny achieves a new compression SOTA on five Chinese NLP tasks, outperforming BERT base by 0.4% accuracy with 7.5x fewer parameters and9.4x faster inference speed.

二. 数据与方法

3. 变量

2021年中国大学生婚育观调查

表 1 核心变量基本情况

变量	样本量	均值/百分比(%)	标准差	最小值	最大值
期望婚龄 (岁)	9694	27.47	3.21	15	62
期望子女数 (个)	9694	1.35	0.85	0	10
互联网使用时长 (小时)	9694	4.60	2.45	1	19
微博使用强度(标化后)	9694	0.00	0.99	-0.78	9.48
同群微博使用强度(标化后)	624	0.067	0.78	-0.78	6.38
负面信息暴露强度	9694	2.39	2.29	1	5

资料来源：2021 年中国大学生婚育观调查

表 2 控制变量基本情况

变量	样本量	均值/百分比(%)	标准差	最小值	最大值
个人属性					
性别 (男)	9694	41.90	/	/	/
年龄 (岁)	9694	21.34	1.59	16	40
学校类型 (基准专科高校)					
双一流高校	9694	9.59	/	/	/
本科高校	9694	60.79	/	/	/
民族 (汉族)	9694	86.88	/	/	/
家庭属性					
Ln(家庭生活费)	9694	7.13	0.52	3.00	10.31
父母最高受教育程度 (基准初中及以下)					
大专及以上	9694	12.66	/	/	/
高中	9694	11.23	/	/	/
家庭关系 (基准差)					
好	9694	75.95	/	/	/
一般	9694	14.94	/	/	/
独生子女 (是)	9694	33.4	/	/	/
户口 (城市)	9694	41.22	/	/	/
地区 (基准东北)					
东部	9694	29.02	/	/	/
中部	9694	22.72	/	/	/
西部	9694	41.89	/	/	/

资料来源：2021 年中国大学生婚育观调查

3. 变量

2022年微博文本数据（关键词与话题）

表 3 微博爬取的婚育关键词

初始日期	截止日期	关键词名称	去重数量(条)	积极情绪占比(%)
2022/6/1	2022/6/30	婚姻	62123	69.30
2022/6/1	2022/6/30	生育/生孩子	64209	56.86

资料来源：爬取自 2022 年 6 月微博原创

表 4 微博爬取的婚育相关话题稳健性检验

初始日期	截止日期	话题名称	话题性质	去重数量(条)	积极情绪占比(%)
2022/3/31	2022/4/1	#这大概就是婚姻最好的样子#	积极	21	90.48
2022/4/10	2022/4/11	#这大概是最理想的婚姻状态#	积极	76	88.16
2022/7/5	2022/7/6	#见妻子被冲走旱鸭子丈夫扑进急流#	积极	547	87.75
2022/6/10	2022/6/17	#唐山一烧烤店内多名男子殴打女生#	消极	2809	34.43
2022/6/16	2022/6/17	#女子称怀孕 7 月被男友打致先兆流产#	消极	2239	33.50
2022/7/6	2022/7/7	#婚房烂尾 600 天 90 后女孩不敢告诉家人#	消极	2173	37.69
2022/3/6	2022/3/8	#建议鼓励在校硕士和博士生结婚生育#	中立	2141	58.94
2022/4/27	2022/4/28	#APP 使用习惯会影响大学生婚育观#	中立	406	47.04
2022/7/7	2022/7/8	#已婚男性死亡风险比未婚低 23%#	中立	286	58.39
2022/7/8	2022/7/9	#一个人过和婚姻不幸哪个更伤身#	中立	1741	48.13

资料来源：爬取自 2022 年 3 月至 7 月若干微博话题

报告大纲

一． 文献回顾与研究假设

二． 数据与方法

三． 基于调查数据的模型结果

四． 基于文本数据的机制分析

五． 结论与不足

三、调查数据的模型结果

表 5 婚育意愿的基础回归表

	模型一 线性回归		模型二 泊松回归		模型三 线性回归		模型四 泊松回归	
	期望婚龄	期望婚龄	期望婚龄	期望婚龄	意愿子女数	意愿子女数	意愿子女数	意愿子女数
截距	22.294***	22.745***	3.127***	3.143***	2.459***	2.418***	1.422***	1.385***
	(0.796)	(0.795)	(0.045)	(0.045)	(0.017)	(0.068)	(0.215)	(0.216)
互联网使用时长	0.016	–0.041**	0.001	–0.001	–0.007*	–0.003	–0.006	–0.002
	(0.017)	(0.017)	(0.001)	(0.001)	(0.068)	(0.021)	(0.004)	(0.004)
微博使用强度		0.283***		0.010***		–0.020*		–0.019^
		(0.053)		(0.002)		(0.123)		(0.012)
期望婚龄					–0.053***	–0.053***	–0.054***	–0.054***
					(0.021)	(0.141)	(0.004)	(0.004)
意愿子女数	–0.782***	–0.772***	–0.029***	–0.028***				
	(0.068)	(0.067)	(0.002)	(0.002)				
控制变量	✓	✓	✓	✓	✓	✓	✓	✓
R2 / McFadden's R2	0.103	0.108	0.119	0.119	0.093	0.093	0.115	0.121
Adj. R2 / Nagelkerke's R2	0.101	0.106	0.118	0.118	0.127	0.128	0.133	0.139
样本量	9694	9694	9694	9694	9694	9694	9694	9694

注：^p<0.1, * p < .05, ** p < .01, *** p < .001; 括号内均为稳健标准误

三、调查数据的模型结果

表 6 婚育意愿的工具变量回归第二阶段表

	模型五 工具变量回归		模型六 中介效应检验	
	期望婚龄	意愿子女数	期望婚龄	意愿子女数
截距	22.867***	2.105***	22.641***	2.469***
	(0.826)	(0.264)	(0.783)	(0.220)
互联网使用时长	–0.056^	0.027**	–0.055^	0.024**
	(0.038)	(0.011)	(0.033)	(0.009)
微博使用强度	0.359*	–0.169***	0.348*	–0.149***
	(0.160)	(0.050)	(0.153)	(0.041)
负面信息暴露程度			0.027	–0.047***
			(0.019)	(0.005)
意愿子女数	–0.769***		–0.762***	
	(0.067)		(0.038)	
期望婚龄		–0.050***		–0.049***
		(0.005)		(0.003)
控制变量	✓	✓	✓	✓
R2	0.108	0.099	0.108	0.113
Adj. R2	0.106	0.097	0.106	0.112
样本量	9694	9694	9694	9694

注：^p<0.1, * p < .05, ** p < .01, *** p < .001；括号内均为稳健标准误

三、调查数据的模型结果

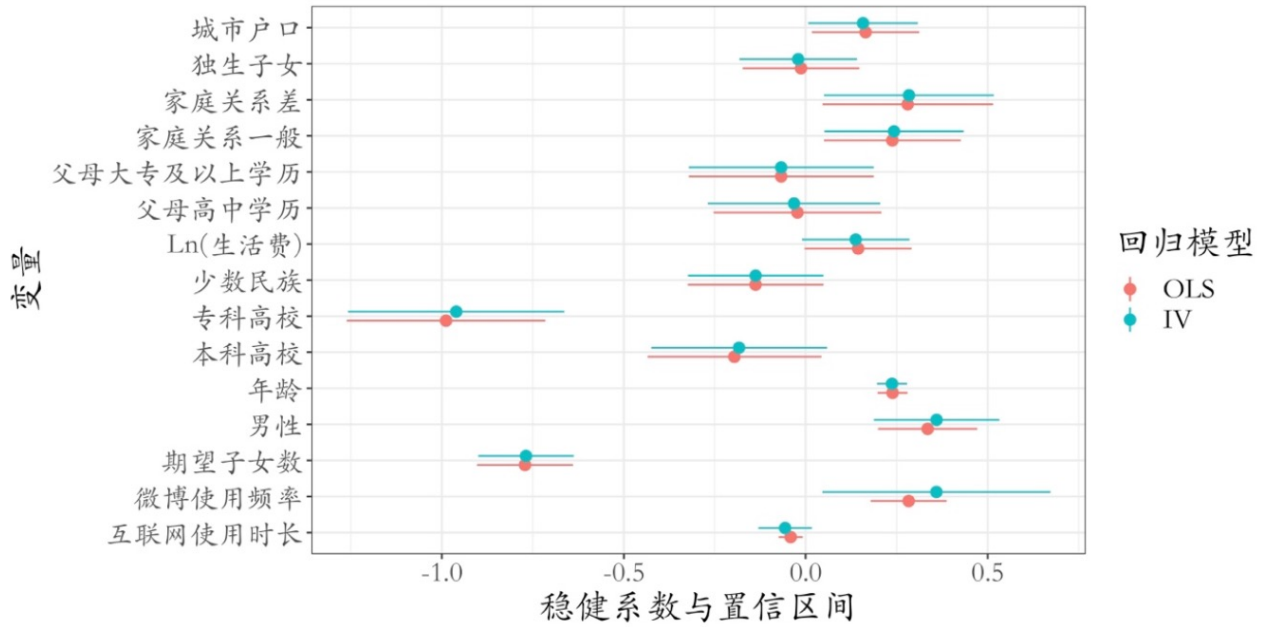


图 1 微博使用强度与期望婚龄工具变量检验

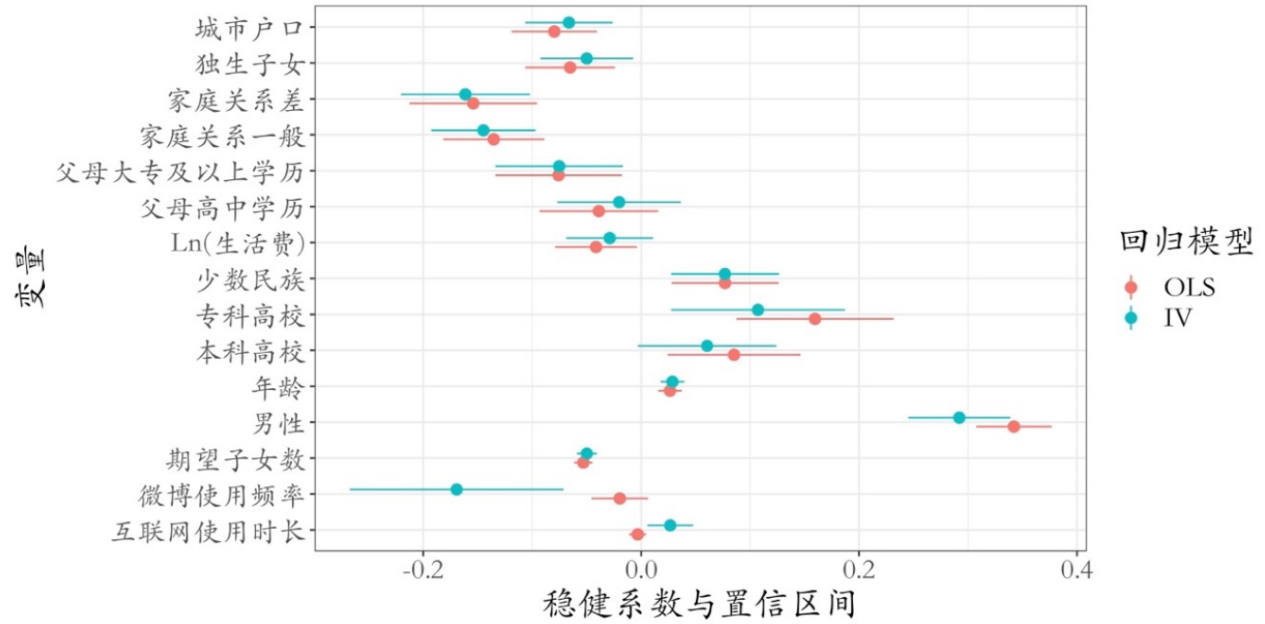


图 2 微博使用强度与意愿子女数工具变量检验

报告大纲

一． 文献回顾与研究假设

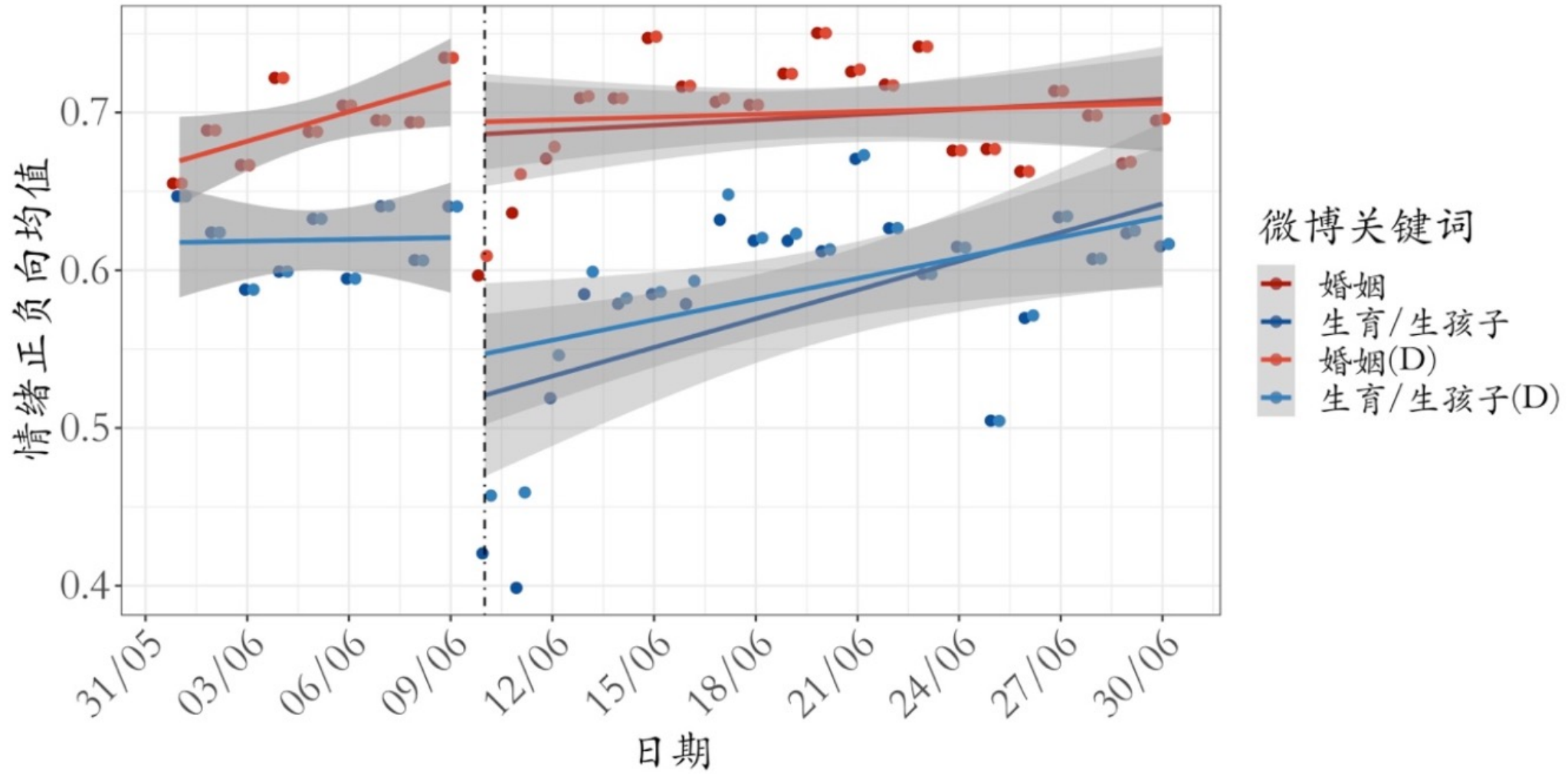
二． 数据与方法

三． 基于调查数据的模型结果

四． 基于文本数据的机制分析

五． 结论与不足

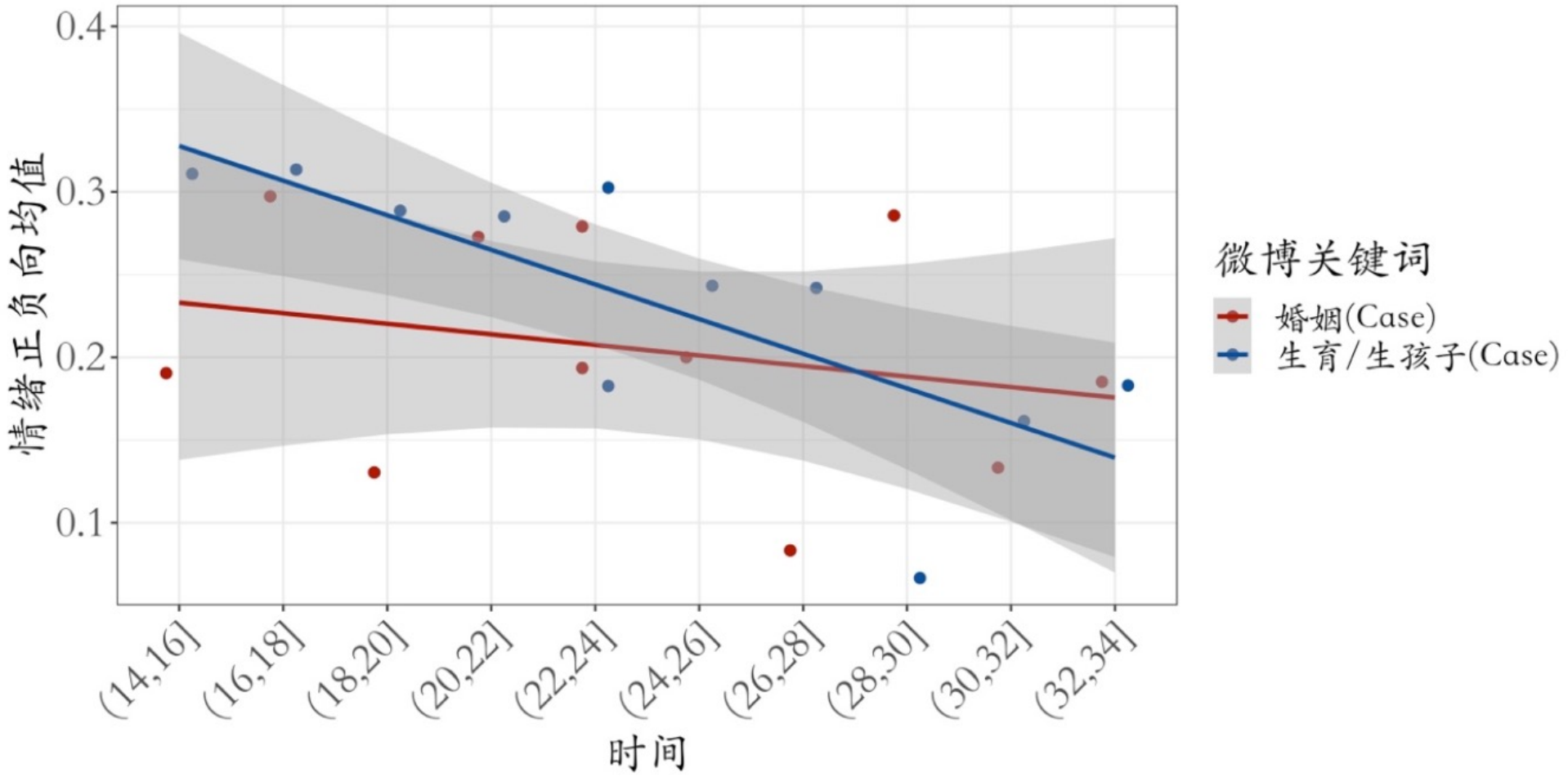
四. 文本数据的机制分析



数据来源：2022 年 6 月微博文本数据（126332 条）

图 3 2022 年 6 月微博婚育舆情（焦点负面事件冲击前后）

四. 文本数据的机制分析



数据来源：2022 年 6 月微博文本数据（截取 2022 年 6 月 10 日 14 时至次日 10 时，2643 条包含某焦点事件关键词的样本）

图 4 2022 年 6 月某焦点负面事件 20 小时婚育舆情

3. 变量

2022年微博文本数据（关键词与话题）

表 3 微博爬取的婚育关键词

初始日期	截止日期	关键词名称	去重数量(条)	积极情绪占比(%)
2022/6/1	2022/6/30	婚姻	62123	69.30
2022/6/1	2022/6/30	生育/生孩子	64209	56.86

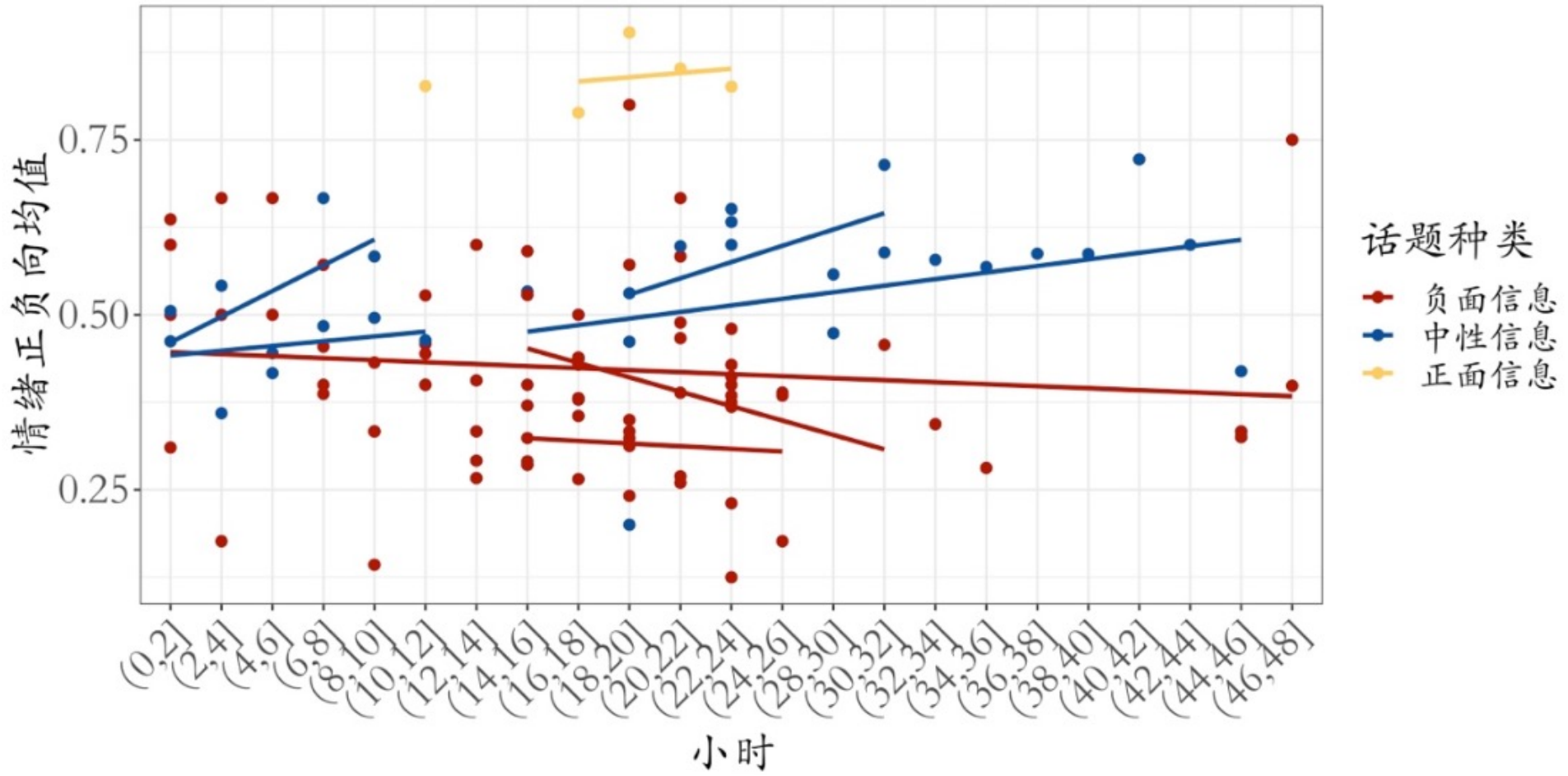
资料来源：爬取自 2022 年 6 月微博原创

表 4 微博爬取的婚育相关话题稳健性检验

初始日期	截止日期	话题名称	话题性质	去重数量(条)	积极情绪占比(%)
2022/3/31	2022/4/1	#这大概就是婚姻最好的样子#	积极	21	90.48
2022/4/10	2022/4/11	#这大概是最理想的婚姻状态#	积极	76	88.16
2022/7/5	2022/7/6	#见妻子被冲走旱鸭子丈夫扑进急流#	积极	547	87.75
2022/6/10	2022/6/17	#唐山一烧烤店内多名男子殴打女生#	消极	2809	34.43
2022/6/16	2022/6/17	#女子称怀孕 7 月被男友打致先兆流产#	消极	2239	33.50
2022/7/6	2022/7/7	#婚房烂尾 600 天 90 后女孩不敢告诉家人#	消极	2173	37.69
2022/3/6	2022/3/8	#建议鼓励在校硕士和博士生结婚生育#	中立	2141	58.94
2022/4/27	2022/4/28	#APP 使用习惯会影响大学生婚育观#	中立	406	47.04
2022/7/7	2022/7/8	#已婚男性死亡风险比未婚低 23%#	中立	286	58.39
2022/7/8	2022/7/9	#一个人过和婚姻不幸哪个更伤身#	中立	1741	48.13

资料来源：爬取自 2022 年 3 月至 7 月若干微博话题

四. 文本数据的机制分析



数据来源：2022 年 3 月至 7 月微博若干话题文本数据（12439 条）

图 5 2022 年 3 月至 7 月若干婚育相关话题事件舆情

四. 文本数据的机制分析

表 7 影响婚育意愿的信息传播机制

	模型六 Probit 回归 事件冲击模型		模型七 Probit 回归 舆论发酵模型		模型八 Probit 回归 舆论发酵模型稳健性检验	
	婚姻样本	生育样本	婚姻样本	生育样本	婚育相关	婚育相关
截距	0.597***	0.363***	–0.359	–0.001	–0.478***	–0.357***
	(0.011)	(0.012)	(0.379)	(0.126)	(0.028)	(0.042)
事件前后 (10 日及以后)	–0.037**	–0.172***				
	(0.012)	(0.012)				
事件时长			–0.017	–0.031***	0.007***	–0.000
			(0.015)	(0.005)	(0.001)	(0.002)
中性信息					0.478***	0.299***
					(0.024)	(0.051)
正面信息					1.547***	1.657***
					(0.066)	(0.276)
事件时长*中性信息						0.010***
						(0.003)
事件时长*正面信息						–0.006
						(0.015)
手机品牌 (iPhone 或华为)	–0.192***	–0.118***	–0.036	0.105^	–0.053*	–0.057*
	(0.011)	(0.010)	(0.160)	(0.056)	(0.023)	(0.023)
McFadden's R2	0.004	0.004	0.004	0.014	0.058	0.059
Nagelkerke's R2	0.007	0.007	0.007	0.023	0.102	0.104
样本量	62123	64209	310	2333	12439	12439

注：^p<0.1, * p < .05, ** p < .01, *** p < .001，表格中系数非边际效应

报告大纲

一． 文献回顾与研究假设

二． 数据与方法

三． 基于调查数据的模型结果

四． 基于文本数据的机制分析

五． 结论与不足

五. 结论与不足

1. 结论

- **支持假设。**

数据显示了强相关，基于理论，我们认为更高的微博使用频率会降低婚育意愿，且微博使用频率通过增加负面信息的暴露机会降低婚育意愿。

具体的信息机制方面。第一，突发的、焦点性负面事件，会对个体婚育意愿产生巨大负面冲击；第二，负面事件本身具有“情绪的漩涡”性，短期内会随时间的增加，加剧个体负面婚育情绪的概率。本文对议程设置理论在婚育意愿方面的有效性提供了经验支持。

五. 结论与不足

2. 不足

- **工具变量法的因果问题。**

  因果识别三要素，时间先后、相关性、排除其他可解释变量 。本文仅作强相关解释。

- **文本数据的机制问题。**

  差群还是同群？初始时中立态度的人群不再发声，还是同群的婚育意愿降低？

- **微博语料库的稳健性问题。**

  反讽语气的识别，“听我说，谢谢你，因为有你，温暖了四季。”

- **宏微观环境的交互问题。**

  媒介作为现实社会的镜子，根本问题仍是社会变化，不能将问题归咎为社交媒体。

感谢周思瑶、王力、梅自颖、郑叶昕等同学提供的建议

五. 结论与不足

3. 贡献

- **婚育意愿的转变侧写。**

虽然数据本身有一定问题，社交媒体上的痕迹可能也是言不由衷的，这仍然体现了中国人家庭观念的部分转变特征，即中国青年群体的婚育观念发生了转变，婚姻仍是绝大多数人的必需品，但其价值基础已经潜在变化了，生育不再是急切的传宗接代。

诚如统计学家们谈及的：“所有的模型都是错的，但还是有点用的。”本文的一手数据，显示了以微博这一社交媒介所代表的舆情场与婚育情绪、意愿的紧密联系。

谢谢各位老师、同学！

中国人民大学人口学系
博士生 闫誉腾
yanyuteng@ruc.edu.cn