REPUTATIONAL SELF-CENSORSHIP
Evidence from an Online Question-and-Answer Forum in China

Haohan Chen*

This Version: March 3, 2021 Latest Version

Abstract

Existing studies have primarily considered *state-focused self-censorship*, a precaution against state sanctions. Social pressure is an important but overlooked motivation for self-censorship under authoritarian rule. I define *reputational self-censorship*, the behavior of withholding genuine political views due to fear of reputational sanctions by fellow citizens, not the state. I test this new theoretical perspective with original online discourse data from Zhihu, an online question-and-answer forum in China. The data capture the nuance of political discourse in the world's largest authoritarian regime with high-quality semi-structured discussion. Its "anonymous answering" option, allowing users to hide identities from fellow users, but not from the state, provides a unique measure of reputational self-censorship. I combine document-term matrix, manual content analysis, Embedded Topic Modeling, and sentiment analysis to create measures of political expression in text. Variable selection methods are employed to find predictors of anonymous answering. Findings reveal reputational self-censorship both *extends* and *supplements* state-focused self-censorship. Extending state-focused self-censorship, reputational concerns cause self-censorship on discussions on regime support and mentions of specific time and places. Supplementing state-focused self-censorship, reputational concerns cause self-censorship on non-sensitive but divided political topics, sharing of personal experience, and engagement in debates.

*Postdoctoral Fellow, Center for Social Media and Politics, New York University, 60 5th Avenue, New York, NY 10011 (Email: haohan.chen@nyu.edu)

I thank Melanie Manion, Timur Kuran, Christopher Bail, Pablo Beramendi, Sunshine Hillygus, and Herbert Kitschelt for reading numerous drafts and sharing tremendously helpful comments. I thank the Duke "Mafia" for continuous support. Special thanks to Peng Peng, for the generous help with manuscript editing. Earlier versions of this paper were presented at MPSA, POLMETH, the 4th Quantitative Studies of China Conference (QCSS), the End-of-year Conference at Duke Political Science, as well as seminars at Vanderbilt, UIUC, UPenn, NYU, HKU, and HKUST. I am appreciative to the fruitful discussion with all the participants. Particularly, Yue Hou, Zhenhuan Lei, Wenfang Tang, Edmund Malesky, Xu Xu, Xiangqian Yi, Han Zhang, and Jiangnan Zhu shared constructive feedback.

1 Introduction

Self-censorship refers to the withholding of the expression of genuine political views. Observed under both authoritarian and democratic contexts, self-censorship is especially prevalent in authoritarian regimes, where it forms an important behavioral foundation for regime stability and survival (Kuran, 1995). Despite its theoretical importance, empirical studies of self-censorship are rare due to the difficulties involved in measuring unspoken genuine views.

This paper attempts to contribute to our understanding of self-censorship by employing a new theoretical perspective and a new empirical design. Theoretically, self-censorship has been primarily considered as a precaution against state sanctions in existing studies. I term such behavior as state-focused self-censorship. In this paper, I theorize a different type of self-censorship: reputational self-censorship. I define reputational self-censorship as the behavior of withholding genuine views as a precaution against reputational sanctions, a type of social pressure enforced by fellow citizens in everyday political talk.

I use unique online discourse data to test the new theoretical perspective. Most existing empirical studies on self-censorship use structured surveys or survey experiments to test self-censorship. I contend that such an approach is unable to capture the multi-dimensionality of political opinions and identify reputational sanctions of my theoretical interest. I present a novel research design using original political discourses on an online question-and-answer forum, Zhihu (知乎), in authoritarian China. Specifically, I take advantage of two unique features of this online community: First, the political discussion takes a semi-structured question-and-answer format, which captures the multi-dimensionality of political talk. Second, the website provides a unique "anonymity" option that allows users to hide their virtual identities from fellow users for posts of their choice. The feature effectively removes reputational—but not state—sanctions, providing an opportunity to identify reputational self-censorship in this virtual community.

I use a combination of text-as-data methods to measure opinions within these rich discourses. As the questions are short text prompts, I use a simple document-term matrix approach supplemented by manual content analysis. I detect the underlying semantic structures of answers of long and variable lengths with Embedded Topic Modeling (ETM), a new topic modeling method suitable for the large and messy corpus in my research. To my knowledge, this is the first application of this method in social sciences. I supplement it with a lexicon-based sentiment analysis to measure emotions in the answers. The generated features are used as predictors of the behavior of anonymous answering with four different types of models: Elastic Net Variable Selection, Poisson Regression, Logistic Regression with L1 Regularization (LASSO), and Logistic Regression.

The empirical inquiry discovers two patterns reputational self-censorship. First, reputational

self-censorship *extends* state-focused self-censorship. Consistent with previous works, people self-censor on discussions about regime support and mentions of specific entities of time and space. Second, reputational self-censorship *supplements* state-focused self-censorship. People self-censor discussions about non-sensitive but divided political topics, exposure of personal experience, and engagement in debates.

This study contributes to our understanding of authoritarian mass opinion and methods for opinion research under authoritarian rule. It attempts to draw attention to the social dynamics of citizen–citizen interactions as an important theoretical perspective overlooked in existing research on the topic. It also introduces novel methods for the measurement of political communication under authoritarian rule.

The remainder of this paper is structured as follows: Section 2 introduces reputational self-censorship with reference to the literature. Section 3 explains the data collection from Zhihu. Section 4 elaborates how measures of public opinions are generated from the text data. Section 5 discusses the empirical models and results. Section 6 concludes the paper and discusses the implications of this study.

2 Reputational Self-Censorship

Self-censorship refers to an individual citizens' behavior of withholding genuine views in political communications.[1] Self-censorship can be situated in the scholarship of authoritarian mass political behavior as a "weak" form of preference falsification (Kuran, 1995). Borrowing from Kuran's depiction of preference falsification, self-censorship means withholding "the truth," while preference falsification includes both withholding "the truth" and telling "lies."

Self-censorship is behavioral type of interest because of its implications for authoritarian survival. It has been considered a mass behavior that keeps unpopular dictators in office and makes revolutions unpredictable. As Kuran (1989, 1991a,b, 1995) theorized, unpopular dictators can survive, as citizens living under repression cannot form anti-regime collective action because they keep their dissent private. However, random shocks can push the society toward contagious mass revelation of dissent (known as a "cascade"), which can quickly mobilize collective action, overthrowing the dictator. This explains the unpredictability of revolutions. The theory was initially used to throw light on the sudden collapse of communism in Eastern Europe in the late 1980s and later, the 1998-2005 "Color Revolution" and the 2011 "Arab Spring" (Lynch, 2011; Hale, 2013).

[1] This study focuses on this behavior among individual citizens. Note that self-censorship differs from media self-censorship. The latter is not studied in this paper. See, for example, Germano and Meier (2013) and Stockmann (2013).

Empirical studies of self-censorship tend to be scarce due to the difficulties involved in measurement. However, recent methodological innovations have resulted in a small but growing stream of research on the topic. For example, experiments and quasi-experiments in authoritarian China consistently provide evidence of self-censorship in surveys and social media platforms, especially in connection with stressful political events, such as political purges (Jiang and Yang, 2016; Chang and Manion, 2020; Robinson and Tannenberg, 2019; Shen and Truex, 2020). An implicit association test among Egyptian citizens revealed a dissociation of explicit and implicit attitudes toward their authoritarian leaders, suggesting that some respondents misrepresented their political support (Truex and Tavana, 2019). However, not all studies testify to the existence of this behavior. For example, a list experiment in Russia suggests no evidence of preference falsification with regard to support for President Vladimir Putin (Frye et al., 2017).

Despite the varieties of contexts and methods, all these recent contributions focus on a particular type of self-censorship, namely, state-focused self-censorship. I define state-focused self-censorship as a type of self-censorship motivated by the fear of state sanctions. That is, citizens choose to withhold their genuine political views because they worry that the state will otherwise punish them. The forms of punishment range from official warnings to imprisonment and execution.

I contend that a different type of self-censorship, namely, reputational self-censorship, remains largely understudied. I define reputational self-censorship as a kind of self-censorship motivated by fear of social punishment by fellow citizens, which can harm one's social recognition. Citizens engaging in reputational self-censorship withhold their genuine views because they worry that fellow citizens in the audience, not the state, will otherwise punish them for what they say. These punishments can take the form of negative comments in a conversation or, more severely, harassment or intimidation.

Reputational self-censorship is theoretically rooted in the classic theory of preference falsification. According to Kuran (1995), citizens decide whether to review genuine political views by weighing two benefits (intrinsic utilities and expressive utilities) and a cost (reputational sanctions). An individual gains intrinsic utilities if revealing genuine views can induce a desirable policy outcome (e.g., a preferred policy adopted by the government). The second benefit, expressive utilities, arises from the psychological reward one earns for being honest. The cost of revealing genuine views, however, takes the form of reputational sanctions incurred by the disapproval of peers in social interactions, which has negative consequences for social recognition. Despite its intuitive plausibility, this behavioral model has inspired few follow-up discussions and has never been subject to systematic empirical tests. Given that self-censorship is a "weak" form of prefer-

ence falsification, I consider the theorized reputational self-censorship in this paper a revival and extension of the neglected behavioral model of the classic theory of preference falsification.

Reputational self-censorship is closely related to social desirability bias, a behavioral anomaly detected among participants of opinion studies. Social desirability bias refers to participants' bias toward giving socially desirable answers in surveys (Krumpal, 2013). Studies have found evidence of social desirability bias in association with numerous political issues in democratic countries. For example, voter turnout reports suffer from upward bias because absentees are embarrassed to admit it (Holbrook and Krosnick, 2010). Support for gay marriage, affirmative action, and affection towards certain religious groups may be overestimated (Blaydes and Gillum, 2013; Cilliers et al., 2015; Coffman et al., 2017; Kuklinski et al., 1997; Powell, 2013). In addition, citizens may hide their support for unpopular presidential candidates, leading to inaccurate polling results (Brownback and Novotny, 2018; Coppock, 2017). Arguably, reputational self-censorship and social desirability bias share the same set of motivations: social sanctions by the audience (although, for the latter, the sole audience is the interviewer). However, the two are different in that the former is studied as a type of mass behavior with institutional implications, while the latter is viewed as an anomaly in opinion research. That said, I expect this study to shed light on social desirability bias outside the democratic context, a matter that few existing studies have systematically discussed.[2]

I expect reputational self-censorship to be prominent in authoritarian countries where the state tolerates a variety of political discussions and citizens are capable of talking about a diverse set of political topics. China is a typical case meeting both criteria. Previous studies show that a variety of political discussions, including criticism against the regime, are evidently tolerated by the state as long as they do not pose an immediate threat of anti-regime collective action (Chen and Xu, 2017; King et al., 2013). At the same time, the variation in political opinions among Chinese citizens can be mapped onto a multi-dimensional ideological spectrum beyond the "pro- and anti-regime" cleavage (Pan and Xu, 2018). As a result, citizens in authoritarian China carry out diverse political discussions as part of their everyday social online and offline activities. They are aware that, for these conversations, the primary audience at stake is their fellow citizens, not the state. That is, sanctions are more likely to come as negative feedback from someone in the audience who disagrees with them, rather than arrest by the secret service that identifies the thoughts as being dangerous to the regime. Hence, China is an ideal case to test the theory of reputational self-censorship.

[2]To my knowledge, the only study on social desirability outside democratic countries is that of Zhou et al. (2020), who show that Chinese college students hide their support for the regime due to the social desirability of dissent.

3 Data

An empirical test of reputational self-censorship can be challenging in two ways. First, conventional methods can hardly gauge the richness of everyday political discourse among citizens in authoritarian regimes. Specifically, surveys with lists of structured questions miss a considerable amount of nuance in public opinions and operate in settings very different from those of real-life political discussions. In contrast, social media discourse, an emerging data source, gauges nuance more comprehensively. However, most data of this kind are generally too unstructured and apolitical for researchers to sample meaningful political discussions from. Second, and more critically, reputational self-censorship is unobservable because it comprises the unspoken part of political talk. Finding a convincing proxy for these unspoken words is difficult, especially for observational studies. To overcome the two challenges, I collected original political discourse data from Zhihu, a Chinese question-and-answer forum. As I will show in this section, the website's semi-structural discussion helps gauge the richness of political talk without being swamped in the messiness of social media data. Moreover, its unique "anonymous answering" option offers an opportunity to measure that part of the political talk that could have been self-censored due to reputational concerns.

3.1 Gauging Political Opinions

The data used for measuring opinions in authoritarian regimes fail to account for the richness of political discourse one way or another. Survey-based research examines self-censorship on a small set of pre-defined concepts such as regime support, approval of the government, and political trust (see, e.g., Jiang and Yang, 2016; Shen and Truex, 2020). The simplified questions do not account for the multi-dimensionality of citizens' political opinions, which are evidently much more complex than the "pro-regime" and "anti-regime" cleavage (Pan and Xu, 2018). Moreover, these questions concern sensitive political stances that people are unlikely to directly discuss in everyday political conversations, making them unfit for the empirical inquiry on reputational self-censorship, which is the focus of this study.

Online discourse data, on the other hand, offer the ability to measure a diverse set of topics. However, their drawback lies in their sparsity of political content and lack of meaningful exchanges resembling real-life conversations. First, in general, Chinese netizens primarily use social media for entertainment, while serious political discussion is scant (Leibold, 2011). For example, to measure public opinion with Weibo, the Chinese equivalent of Twitter, researchers usually need to use keywords to filter posts possibly related to a set of topics of interest (see, e.g., Chang and

Manion, 2020; Zhang and Pan, 2019; King et al., 2013). Such a strategy is sub-optimal since manual selection of an exhaustive list of keywords hinders my ability to capture the diversity of topics. Second, even if the social media posts contain political keywords, most contain far from meaningful exchanges of political opinions. The majority of posts on popular platforms such as Weibo or Twitter are typically cheerleading in favor of or opposition to opinion leaders than offering serious reasoning and deliberation. Hence, online discourse from most social media platforms is too scattered to approximate real-world political discussions in which participants exchange opinions with an audience about a loosely defined topic of interest.

Addressing the disadvantage of existing empirical approaches, I collect original data from Zhihu, a Chinese online question-and-answer forum. Zhihu is one of the few social media platforms where quality political discussions take place in authoritarian China. It is well-suited for this study given its popularity, unique question-and-answer structure, and active and influential community for political discussions.

Zhihu is one of the most popular social networking sites in China. Launched in December 2010, it has become one of the most popular and fastest-growing websites in China. In terms of traffic, Zhihu ranks 24th among Chinese websites and 105th in the world.[3] To provide readers with a point of reference, the only other two social networking sites whose traffic counts are larger than that of Zhihu are *Weibo* 微博(7th), a popular social networking site frequently appearing in previous studies on Chinese social media, and *Tianya* 天涯(22nd), a famous social networking site with a much longer history (launched in 1999). Zhihu is a populated community where many Chinese citizens enjoy their virtual social lives: the site's management announced that it had over 100 million registered users, 26 million active users per day, a visit time of 1 hour per person per day, and over 18 billion visits per month in 2017. [4]

Discussions on Zhihu follow a question-and-answer structure. When a user wants to propose an issue of interest for public discussion, they can submit a "question" containing a one-sentence title and (optionally) a brief elaboration. They may also tag the question with a set of keywords. To participate in an existing discussion, a user can submit an "answer" under a question of interest. Users can also choose to "sit in the audience" of a discussion stream by "following" a question to receive notifications about new answers, "upvote" an answer they approve of, "downvote" an answer they disapprove of, or "comment" on answers to express support or opposition. Questions and answers are generally open for the participation of all users of this virtual community. The summary statistics of positive or negative reactions to questions and answers (except for downvotes)

[3] Alexa. 2018. "zhihu.com." https://www.alexa.com/siteinfo/zhihu.com (accessed August 26, 2018).

[4] 腾讯科技. 2017. "知乎宣布注册用户数超1亿并开放机构号注册." tech.qq.com http://tech.qq.com/a/20170920/020694.htm (accessed on October 22, 2019).

are public information. This unique question-and-answer structure resembles the configuration of discussions in real life: Someone proposes a loosely defined theme, and those interested opt in for the discussion, expecting that their views will be evaluated by the other participants of the discussion. Pragmatically, keyword tagging makes it extremely convenient to identify political discussions in data collection: simply selecting all the questions with the keyword "politics" will suffice.

Zhihu hosts an active community for political discussions. Although most popular social networking sites in authoritarian China allow political discussions to some extent, Zhihu is especially known for its high-quality deliberation and tolerance of politically sensitive content.[5] In a few cases, discussion on this website shaped the public discourse of important political incidents in the past. For example, consider the *Lei Yang incident*, in which a young man died suspiciously during custody in Beijing. The victim's friends initially expressed concerns and shared information about the incident under a Zhihu question. This quickly resulted in large-scale media exposure and public attention outside Zhihu.[6] Although no official data about the number of people who participated in this political discussion are available, I can infer its stake by the number of members who contributed to political talk (i.e., by adding up all the forms of participation mentioned above). I estimate that 1.8 million people participated in this discussion, amounting to nearly 2% of the total number of Zhihu users as of 2017. This is a large proportion, considering the diversity of topics discussed on this generic website.

Given the above three features, Zhihu provides uniquely high-quality political discourse in authoritarian China, outperforming the two other social media platforms that are typically mentioned in the literature, Weibo and WeChat. As discussed in the previous section, Weibo is popular but offers sub-optimal measurement capability for my study due to the scarcity of political content and unstructured discussion. Moreover, its information feed centers around its users' networks, facilitating the tracking of news and posts from users' favorite opinion leaders and celebrities, but inconveniencing engagement in serious discussions on political topics of interest. WeChat, on the other hand, is primarily used as a tool to communicate with close connections. A recent study finds that people are unwilling to conduct political discussions on WeChat (Stockmann and Luo, 2017). In addition, although WeChat groups have been considered as venues for political discussion in recent years, the private nature of those groups drives up the cost of data collection to prohibitive

[5]The Economist. 2016. "Posers for the party: How an online forum catches censors unawares." economist.com. https://www.economist.com/news/china/21706331-how-online-forum-catches-censors-unawares-posers-party (accessed September 19, 2019).

[6]Wang Xiangwei. 2016. "A young life lost: it's time for justice to be served." scmp.com. http://www.scmp.com/news/china/policies-politics/article/1966336/young-life-lost-its-time-justice-be-served (accessed September 19, 2019).

levels.

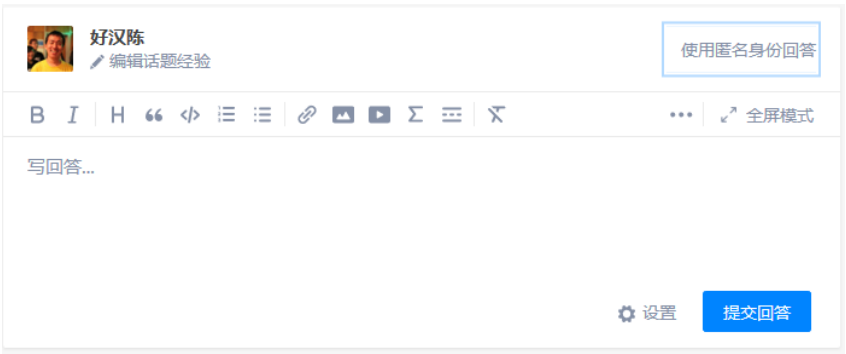
3.2 Observing the Reputationally Self-Censored Discourses

By definition, self-censorship is unobservable, because it is the unspoken part of a conversation. A unique feature of Zhihu, "anonymous answering," helps overcome this empirical challenge. In particular, as I will show in this section, the feature helps identify reputational self-censorship, the very behavior of theoretical interest to this study.

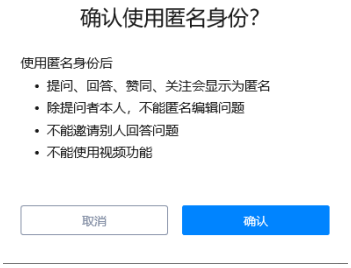
Like other social media platforms such as Weibo and Twitter, a Zhihu user can post answers under a unique virtual identity recognizable by fellow users. This platform is unique in that it allows users to detach their identities from specific posts with an "anonymous answering" option. As shown in Panel (a) of Figure 1, a regular Zhihu answer includes the author's virtual ID, profile picture, and a link to their profile page. In addition, a snapshot of the answer appears on the author's timeline, which is pushed onto the news feed of the author's followers and is visible to anyone viewing their profile page. However, when a user posts an answer "anonymously" ("匿名回答"), as shown in Panel (c) of Figure 1, the author's information is removed from their answer. In addition, the snapshot of the answer does not appear on the follower's news feed and is not visible to viewers of the author's profile page. The availability and effect of the anonymity option is well understood among Zhihu members according to observations of discourses in the community. Using the anonymity option needs active confirmation in a pop-up window, as shown in Panel (b) of Figure 1, making it reasonable to assume that an observed anonymous answer is most likely the result of a conscious choice.

Notably, using the "anonymous answering" function lifts reputational sanctions, but not state sanctions. "Anonymous" answers are anonymous to fellow users but not the website administration and the state. In an official user guide available to all users, the administration acknowledges that it is capable of linking anonymous answers with their authors using source data in the database, although the Zhihu administration reassures people about data security effected by a strict internal data access protocol. The Chinese state has access to the website data. The Chinese state maintains close partnerships with commercial Internet providers in its effort to control cyberspace (Miller, 2018). When needed, the authority can conveniently request information from social media platforms, including Zhihu, to trace the real identities of users of interest and enforce state sanctions accordingly (Qin et al., 2017).

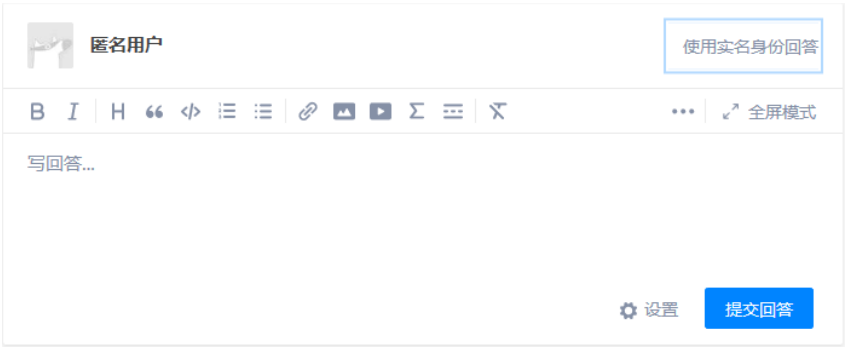
One possible objection is that the users of Zhihu may not understand the design of the anonymity option and are not actually using the anonymity option to publish uncensored opinions to avoid reputational sanctions. Qualitative evidence shows that this concern is not warranted. I analyze 155



(A) Regular answering



(B) Confirmation



(C) Anonymous answering

FIGURE 1: Zhihu User Interface

answers to a Zhihu question titled "Why do some people like to answer questions anonymously?". As the question suggests, users discussed their motivations for answering questions anonymously. The findings imply that Zhihu users' understanding of this unique feature is in line with the expectation of this research design.

The first finding to emerge from the above-mentioned discussion stream is that many users recognize their intention to express genuine views under the cover of anonymity. For example, one of the top-rated answers is simply a direct quote from Oscar Wilde: "Give a man a mask and he will tell you the truth." Another top-rated answer is "Because the story is true, the name must be fake" ("因为故事是真的，所以名字是假的."). Many other answers indicate that users use anonymity to reveal their true selves and the truth (e.g., "希望以知情人身份揭露真相", "匿名的我才是真实的我.").

Second, users understand that anonymity can help avoid reputational sanctions. The most endorsed answers listed three reasons for anonymity: protection against personal attacks, driving interests toward an answer itself instead of its author, and preventing an answer from appearing in the information feeds of followers. All three reasons are directly related to reputational sanctions.

In the other answers, the users' reasons were closely related to the need to avoid reputational sanctions by protecting their identities ("碍于身份"), personal image ("破坏形象"), and privacy ("涉及隐私"). As detailed in a number of answers, concerns about reputational sanctions were not limited to online sanctions; offline repercussions were also feared. Some users wish to hide from followers in the virtual community (e.g., in Chinese: "回答的问题不想让关注的盆友看到,"). Others want to hide from real-life connections such as colleagues and friends (e.g., "自从之前的回答被前同事、朋友发现后，就开始了匿名答题的习惯.").

Finally, the users are generally aware that anonymity cannot protect them from state sanctions. Among the total of 155 answers on users' reasons to answer anonymously, very few mentioned avoiding state sanctions as the motivation. Even though a few answers revealed misunderstandings, all of them were immediately corrected by fellow users in the comments section. For example, the following answer advocates providing anonymous responses to avoid state sanctions: "In a place without freedom of speech, anonymity is the only way to protect myself against harm" [7]. The comments section of this answer shows the following top-rated comment: "If the authority wants to make trouble, anonymity is useless. Anonymity is mainly for avoiding resentment and subsequent personal attack from Zhihu users." [8] Another use writes, "Anonymity is naturally good for answering highly sensitive questions that can lead the authority to check my water meter."[9] A fellow user corrects this misperception by commenting "Anonymity is useless if [the authority] wants to check your water meter." [10].

3.3 Data Limitation: Sampling, Censorship, and Unreality

Although the political discourses on Zhihu provide a unique opportunity to test reputational self-censorship, one may legitimately question three major limitations of the data, namely, non-random sampling, interference of state censorship, and the unreality of online political talk compared to the reality of such talk in real life. I contend that none severely limit the data's fit for answering the research question of interest. First, although non-random sampling limits the generalizability of this study, the unique demographic information available allows a clear profiling of the sample of interest. Second, although state censorship exists, it is evidently lenient on Zhihu. In addition, the focus on everyday political conversation as a social activity alleviates concerns about censorship.

[7] In Chinese: "在一个不可能有言论自由的地方，匿名就是保护自己不受伤害的唯一办法."

[8] In Chinese: "如果上头要找事匿名是没什么作用的匿名主要还是为了防止引起知乎用户的反感然后招致对人不对事的攻击吧."

[9] In Chinese: "有些容易被查水表的高危问题自然要匿名." (Note that "to check my water meter" is the literal translation of "查水表" in Chinese, a social media slang implying state sanctions.

[10] In Chinese: "要查水表你匿名也没用啊"

Third, the unreality of social media political talk is biased against my results, since the stake of reputational sanctions is lower online. This aspect confers conservativeness on my findings about the size of reputational self-censorship.

3.3.1 Non-random Sampling

The participants in political discussions on Zhihu do not constitute a representative sample of the Chinese population. According to summary statistics published by Zhihu's owners, their users are more likely than the average Chinese person to be urban, high-income, and college-educated. Specifically, as of 2017, 74% of Zhihu users were college-educated or higher, 20% were classified into the "high consumption group" (compared to 8.4% of all Internet users), 50% had a monthly expenditure of above 2,000 yuan (or about 308 US dollars). [11][12][13]

Moreover, other than the indirect evidence from news reports, I profile the sample with users' self-reported demographic information. This lack of demographic data poses a common limitation in social media studies. Researchers can easily collect a standard set of demographics, such as age, income, and party affiliation, from surveys. However, these elements are usually unknown when the data are sourced from the social media. That said, Zhihu provides demographic information of a much higher quality than most other social media platforms. Users have the option to list their gender, location, education, and occupation on their profile page. Zhihu does not have a differential privacy policy to restrict access to any information. Not surprisingly, a considerable proportion of the users opt out of self-reporting, as shown in Panel (a) of Figure 2. As many as 90% of the users report their gender, while the proportions of those reporting valid information about their location, education, and occupation are 29%, 23%, and 43%, respectively. These proportions of valid demographic data are much worse than the extent of information typically provided by surveys. Nonetheless, the quality is much higher than that available from most other social media sites, whose demographic information is more scattered or even nonexistent. Such information suffices for profiling the sample at the macro level.

The available demographic data show that male and college-educated residents residing in economically prosperous areas of China, who study or work in the information technology, higher education, law, or finance sectors, are over-represented in this sample. As shown in Panel (b) of

[11] According to the Statistics Year Book of 2018, as of 2017 only 6.4% of the Chinese population was college-educated. https://www.sohu.com/a/312444197$_9$9964340

[12] In 2017, the average monthly expenditure of a Chinese resident was 1,527 yuan (or about 236 US dollars) according to the National Bureau of Statistics http://www.stats.gov.cn/tjsj/zxfb/201801/t20180118$_1$574931.$html$.

[13] 南方都市报. 2017. "知乎晒"三高"用户数据：高学历、高收入、高消费." mp.oeee.com. https://m.mp.oeeee.com/a/BAAFRD00002017072745427.html (accessed September 19, 2019).

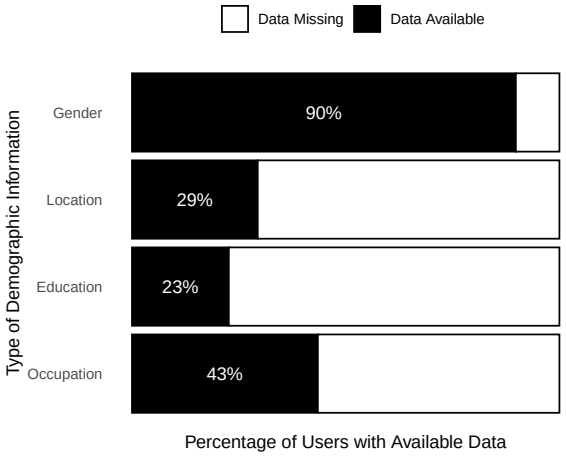
Figure 2, 78% of the users self-identify as college-educated (highly consistent with the estimate of 74% in the official report), and as many as 73% are male. Panel (c) of the same figure shows the geographic distribution of the users, with the top 10 locations labeled. Evidently, the users disproportionately self-identify as residing in Beijing, Guangdong, Shanghai, Jiangsu, or Zhejiang, four of the most economically prosperous provinces or municipalities in China. However, the users' locations are diverse enough to ensure non-zero samples in all provinces. Finally, the word cloud in Panel (d) of Figure 2 shows the most popular self-identified occupations. The size of each word is proportional to the frequency of its occurrence. Featured at the center of the figure, the terms "Internet," "high-tech," "computer software," "higher education," "law," and "finance" point to the most prevalent descriptions of the users' occupations.

The non-random sampling limits the generalizability of this study. The findings are expected to be generalizable only to the urban, high-income, and college-educated citizens of China. Whether they also apply to rural, low-income, and non-college-educated persons is an open empirical question that is beyond the scope of this study.[14] Despite this fact, this empirical study is the most appropriate of its kind for two reasons. First, given its capacity to accurately profile the sample, its results are based on the best possible available data quality among studies on social media discourses, whose samples may be biased in unknown ways. Second, the overrepresented population group is of interest to the academia. Young, urban, informed, and educated citizens are expected to drive institutional reforms and even democratization in authoritarian regimes. Hence, understanding their patterns of reputational self-censorship can help explain the regime stability of authoritarian China.

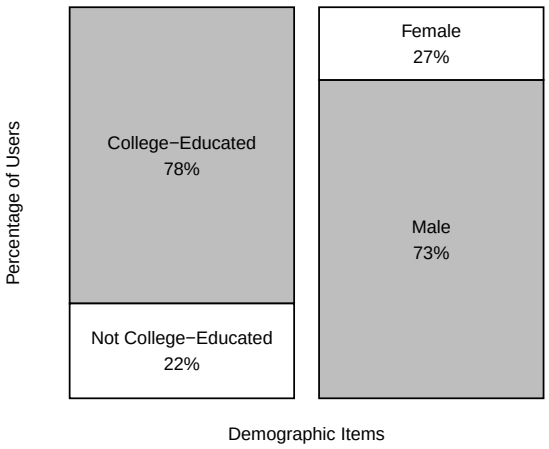
3.3.2 State Censorship

Interference by state censorship poses the second critical limitation of the data used in this work. The Chinese state has developed an immense capacity for online policing, censorship, and propaganda, which has effectively alleviated the potential harm caused by the new technology to the regime (Qin et al., 2017). It deploys such capacity strategically, for example by reducing speeds and success rates of access to foreign websites, removing information with collective action potential while tolerating some criticism, and distracting people from online criticism (King et al., 2013, 2014, 2017; Roberts, 2018). Furthermore, the state uses strategic information control to monitor local governance with minimal risks of anti-regime collective action (Chen and Xu, 2017; Lorentzen, 2014).

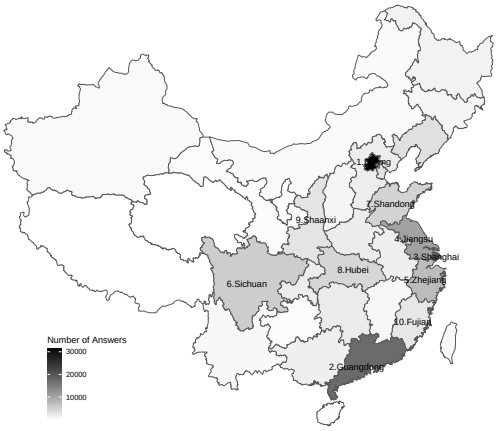
[14] Previous studies suggest that self-censorship may be less prominent among lower-income and not-college-educated groups (Jiang and Yang, 2016)



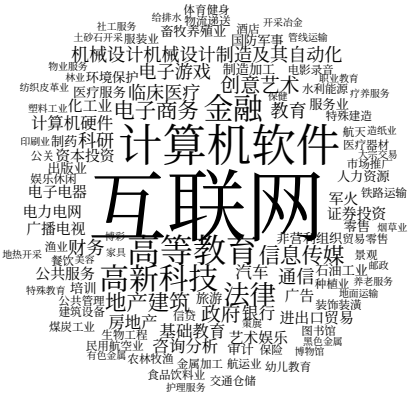
(A) Data Completeness



(B) Gender and College-education



(C) Locations



(D) Occupations

FIGURE 2: Demographics

Zhihu is not exempt from state censorship. However, two pieces of indirect evidence suggest that the level of censorship on Zhihu is low. Anecdotally, for the period of the political discourses examined in this study (from 2010 to 2016), the Chinese censors were more lenient on Zhihu than other popular social media platforms such as Weibo. Zhihu is famous among Chinese netizens for tolerating discussions on political issues that would have been deleted or prohibited elsewhere. Although no systematic study has been conducted to support this claim, the phenomenon was obvious enough to be discussed in a 2016 Economist article titled "Posers for the party: How an online forum catches Chinese censors unawares."[15] Citing multiple incidents where discussions on Zhihu about highly sensitive political incidents gained publicity and even caused policy changes, the article recognized the censors' leniency with Zhihu and attributed it to the fact that the website does not present the news or host videos, which the authorities consider more sensitive.

The second indirect evidence for the low level of censorship lies in the visibly censored answers in my dataset. Zhihu is no different from many other Chinese social media platforms in this regard; censorship is visible even if the censored posts are covered with only a line of warning. Zhihu is unique, however, as it allows authors to "revise and resubmit" (RR) their censored answers for a second review. The text of an answer censored for political sensitivity is covered by an RR request stating, "Revision suggested for the answer: Political content unfit for public discussion."[16] Answers of this kind show an empty text field in my dataset. Among the total answers collected, 0.5% of regular answers and 0.7% of anonymous answers have empty text fields, pointing to the proportion of answers for which a revision was recommended by the censor but ignored by the authors. The proportions are arguably low, and there is no significant difference between the number of regular and anonymous answers. This suggests that the overall magnitude of censorship is low and, more importantly, there is no evidence to indicate that censors treat regular and anonymous answers differently.

Finally, state censorship has a limited effect, given the theoretical focus of this paper. As discussed in Section 2, the behavior of interest to reputational self-censorship is everyday political talk as part of people's social activities. I can reasonably expect that the type of talk that takes place in these social activities is not geared toward organizing anti-regime collective actions, which, according to King et al. (2013), are the main target of censorship.

[15]The Economist. 2016. "Posers for the party: How an online forum catches censors unawares." economist.com. https://www.economist.com/news/china/21706331-how-online-forum-catches-censors-unawares-posers-party (accessed September 19, 2019).

[16]In Chinese: "回答被建议修改：不宜公开讨论的政治内容."

3.3.3 Unreality

A final possible limitation to this empirical design concerns the differences between online and real-life political communication. While I acknowledge this limitation, it only biases against my argument, making the findings of this paper conservative estimates of the magnitude of reputational self-censorship.

Online discussions are accompanied with some cover of anonymity, and concerns about one's reputation being at stake in a virtual community are arguably lower than those in a real-life community. People participate in online discussions with virtual IDs. When needed, they can change or erase them. There is a fair chance, for most cases, that the audience will not link a user's virtual ID with their real-world identity. Applying this understanding to the users of Zhihu, I expect many of those who post regular answers with their traceable Zhihu IDs assume some level of "default anonymity" which generally lowers concerns about reputational sanctions compared to those in real-life political discussions. Thus, the differences between regular and anonymous answers are expected to be smaller than those between real-world self-censored and uncensored political talk. Hence, the data limitation biases against my results, and my findings are conservative estimates.

3.4 Descriptive Statistics

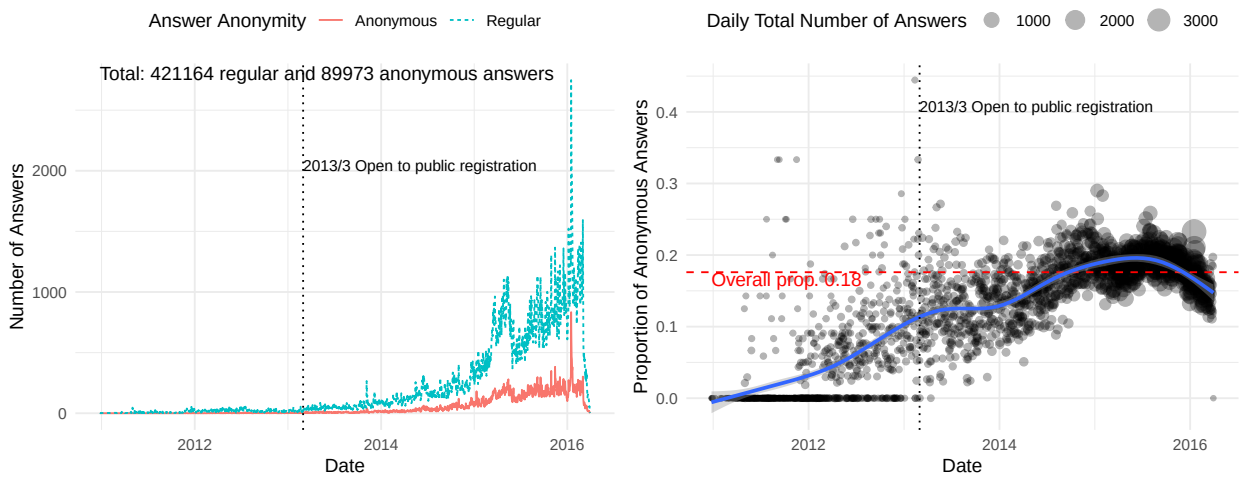
The web scraper developed for this study was run in March and April of 2016 to collect all questions and answers ever posted under the topic "politics" on Zhihu. The dataset contains a total of 101,532 questions and their 511,137 answers, contributed by at least 184,848 users.[17] The profiles of the users, including their self-reported genders, locations, educational qualifications, and occupations discussed in the previous section, are also part of the dataset.

The political discourses analyzed in this study were posted between December 25, 2010 and March 30, 2016, the median of the answers' timestamps being July 28, 2015. The earliest answer on a political topic was posted when the website had just been launched and was undergoing internal testing.[18] This suggests that politics is one of the oldest topics of discussion on the website, and it has continued to garner interest among Zhihu's users. The growth in the website's popularity was accompanied by a rapid increase in the number of political discussions from 2014 to early 2016, bringing the median timestamp to mid-2015.

About one 1 of 5 (18%) answers in the dataset is posted anonymously, with the volume and

[17] This is the count of unique authors of regular answers. Authors of anonymous answers are unidentifiable, although I expect most, if not all, of them are present in the pool of the identified users.

[18] 网易科技报道. 2011. 主题分享：知乎创始成员成远. https://tech.163.com/11/0516/23/747AV8JG00094LEH.html (accessed January 19, 2021).



(A) Number of Anonymous Answers (B) Proportion of Anonymous Answers

FIGURE 3: Descriptive Statistics

proportion of such answers being volatile, but nonetheless increasing over time. Figure 3 summarizes the specific patterns of anonymous posting. Panel (a) shows that the number of regular and anonymous answers increased considerably in 2014, 2015, and early 2016. Panel (b) shows that the proportion of anonymous answers increased steadily over time. Note the critical time point in March 2013 in both panels (annotated in the figure). During this time, Zhihu changed its eligibility for registration from invitation-only to open registration to the public. The introduction of more "strangers" to the community can impact users' risk evaluations of reputational sanction and change self-censorship behaviors. For instance, as Panel (b) shows, the said time point is associated with only a few spikes in the number of anonymous answers closely before and after this date, and it marks the plateauing of the extent of anonymity after a period of growth. The lack of a significant change in behavior around this important time point confirms that I need not treat the data prior to and after it separately.

To examine question-level variations, I narrow down the dataset to a subset of 14,228 questions with at least 20 followers and at least 1 answer. I conduct this exercise for two reasons. First, theoretically, the questions are considered to be political talk when reputational sanctions are at stake and an audience exists. I define a lower bound of the number of followers to filter out questions for which an extremely small audience makes reputational concerns negligible. Second, from an empirical standpoint, as in the case of all social media data, the distribution depicting the attention around Zhihu has an extremely long tail. Most questions received few or no answers, leaving their anonymity metrics undefined. Thus, I focus on a subset with definable metrics. The

proportion of anonymous answers varies across the questions. Panel (c) of Figure 3 shows the distribution of the proportions of anonymous answers in this subset. On average, a question has 15% anonymous answers (variance = 0.02). Finally, Panel (d) lists the questions that obtained the largest number of anonymous answers. The topics of these questions span across a set of politically sensitive and contentious issues in China, including nationalism, the one-child policy, Taiwan, the death penalty, and corruption.

Interestingly, the discourses examined in this study were posted during a period when both the Chinese regime and the website went through tremendous change. Regarding the regime, Xi Jinping took office in 2012, which marked the start of a series of political changes compared to the Hu–Wen government, including escalated information control and enhanced state propaganda. Regarding the website, the growing number of users is anecdotally associated with demographic changes, featured by the addition of non-college-educated and lower-income users. The empirical inquiry of this study focuses on the average magnitudes and variations of reputational self-censorship in this virtual community throughout this period; I leave assessments about the time dynamics to follow-up studies.

4 Feature Engineering with Text Data

I use a variety of text-as-data methods to extract features from the text of questions and answers. Feature engineering using the text of the questions is relatively straightforward. As the questions are short text prompts (14 tokens on average and 58 at the maximum), I use a document-term matrix approach and supplement it with manual content analysis. The texts of the answers are, on average, much longer than those of the questions and show larger variations (114 tokens on average and 45,115 at the maximum). Thus, I require dimensional reduction methods beyond document-term matrix to create interpretable features. I employ ETM, a state-of-the-art unsupervised text clustering algorithm, to summarize the text into topics (Dieng et al., 2020). This is the first application of this method in political science. I supplement it with a lexicon-based sentiment analysis, which measures the emotions contained in the answers.

4.1 Questions Asked

Using the 14,228 questions with at least 20 followers and at least 1 answer, I construct a document-term matrix of unigrams, bigrams, and trigrams using term frequency–inverse document frequency

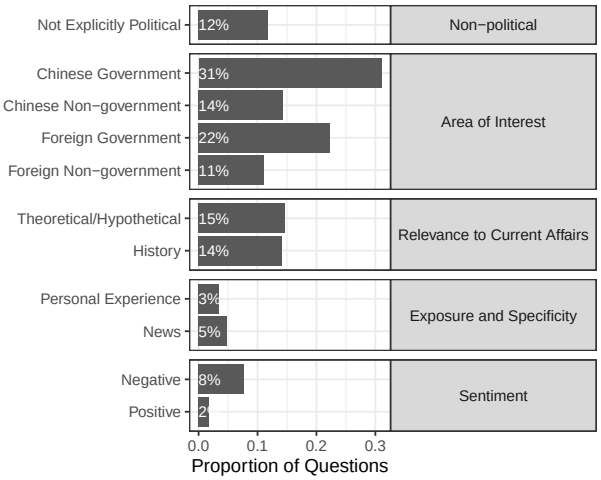
(TF–IDF) weighting.[19] The resulting matrix has 1,504 unique tokens. Panel (a) of Figure 4 shows a word cloud of the top 100 most frequent tokens (summations of the TF–IDF scores of the tokens). Notably, two of the most frequent words are "China" (中国) and "politics" (政治), indicating that the primary focus of the collected discourses is likely Chinese politics. In addition, discussions about other countries, for example the United States and Japan, are prevalent. This finding points to a group of discussions about foreign countries where the user's self-censorship behaviors differ from those associated with discussions of Chinese domestic politics. Two other most frequently used words are "evaluate" (评价) and "view" (看待), suggesting that a considerable number of questions explicitly required users to take positions. Diverse sets of specific political entities, domestic and foreign, contemporary and historical, are also present in the word cloud. The descriptive statistics show that analysis of the questions on Zhihu can address the challenge posed in this empirical work, namely capturing the richness of everyday political conversations among Chinese citizens.

The above data-driven approach for measuring questions is supplemented by a manual content analysis. I manually label 1,016 questions whose number of followers are within the top 1 percentile. After a careful examination of all the questions, I code them with a set of non-mutually exclusive labels in four categories of interest: area of interest, relevance to the real world, relevance to personal experiences, specificity, and sentiment. A "non-political" label is assigned to questions that do not appear to be explicitly political. Panel (b) of Figure 4 summarizes the frequencies of the labels.[20] First, 12% of the questions do not appear to be explicitly political. For example, this group includes informational questions about some entities related to politics, but the users are unlikely to form political opinions about them (e.g., "各地博物馆的镇馆之宝有哪些?"). They also include questions about business or social issues that that may lead to a discussion about politics and policies (e.g., "中国内地最值得敬佩的信息技术公司是哪一家?"). In terms of areas of interest, 31% of the questions pertain explicitly to the Chinese regime, while 22% concern foreign regimes. Moreover, a substantial number of questions on social issues do not explicitly refer to a regime (14% concern Chinese society, and 11%, foreign societies). Not all questions are related to current affairs: 15% are theoretical or hypothetical inquiries (e.g., "社会主义国家的审美特点是什么?") and 14% concern history (e.g., "为什么朱元璋用重刑解决不了贪腐问题?"). A small proportion of questions (3%) invites users to share their personal experiences (e.g., "不想让儿子入少先队，我有错吗?"). With regard to specificity, 5% invite discussions on specific news items, and 33% of the questions mention specific names and places. Finally, some questions refer

[19]The TF–IDF weighting downweighs common terms appearing in many documents.

[20]Note that the summation of the percentages exceeds 1 because the labels are not mutually exclusive. For example, a question can be labeled as being about the Chinese regime as well as relevant to personal experiences.

(A) Most Frequent Tokens by TF-IDF



(B) Distributions of Manual Labels

FIGURE 4: Text Features of Questions

to a sentiment: 8% are asked with a negative sentiment (e.g., "江苏省委组织部发文大学生村官不再续聘，是否为卸磨杀驴?", "香港的现状如何？较之以前变差了吗？为什么?"), while 2% are asked with a positive sentiment (e.g., "你从什么时候开始意识到中国正在逐渐变强?").

4.2 Answers Given

I generate interpretable features with ETM to discover the hidden semantic structures in the answers. Supplementarily, I conduct a sentiment analysis to measure the emotions in the answers. ETM, developed by Dieng et al. (2020), is a state-of-the-art method combining the strengths of two important natural language processing algorithms: Word Embedding and Topic Modeling. I choose the algorithm because of its empirically proved capacity to generate interpretable topics even when the size of the vocabulary is large, a typical feature of the social media discourses of interest to this work. The sentiment analysis is conducted to compensate for a disadvantage of topic modeling: it is unable to explicitly capture emotions, an important feature of social media political discourses. I use a state-of-the-art lexicon developed by the NLP Laboratory of the Dalian University of Technology, China, to construct fine-grained measures of 12 types of positive emotions and 8 types of negative emotions (徐琳宏et al., 2008).

4.2.1 ETM

ETM combines the strengths of two important natural language processing algorithms, Word Embedding and Topic Modeling. Word Embedding, also known as distributed semantics, learns numeric representations of text in a low-dimensional space (typically a few hundreds of dimensions) (Mikolov et al., 2013). The algorithm is based on the linguistic finding that words with similar meanings or semantic functions tend to appear alongside one another or co-appear with the same sets of other words. Based on this intuition, the Word Embedding Algorithm trains a neural network model to learn the relationships among words that appear in local windows of specified lengths in documents. The algorithm has two variants: skip-gram and Continuous Bag of Words. The former learns words' numeric vector representations to optimize the model's capacity to predict words surrounding a focal word, while the latter predicts focal words with surrounding words. The outputs of word embeddings are numeric vectors that can represent the semantic relations among all words in the vocabulary. The trained model can be tested by qualitatively examining whether words whose vectors with high cosine similarities have related meanings.

Using all the text data in the Zhihu political discourses, I construct a corpus as training data. [21] Word embeddings (skip-gram with a local window of 5) are trained on the data, resulting in 255,391 unique tokens of 300-dimensional word vectors. Qualitative examination indicates that the learned vectors capture the semantics of the words very well. Table 1 shows four political entities of interest and their semantically related words or phrases detected by the algorithm. The algorithm is capable of cluster variants of the same political entity type. For example, the word vector of "政府" (government) is close to those of "中央政府" (the central government) and "中国政府" (the Chinese government), and even "执政党" (the ruling party) and "执政者" (the ruler). In addition, it identifies the slang terms for these entities and their nicknames. For example, "ZF," the Pinyin of the Chinese word "government," which netizens usually use to avoid censorship, is identified as a term similar to "government." Similarly, "GCD" and "TG," nicknames for the Communist Party, are identified as similar terms by the algorithm. Finally, the algorithm can capture memes. For example, a set of memes, known as "mo ha" (膜蛤), based on the speeches of Jiang Zemin, a retired CCP General Secretary, is popular among Chinese netizens, including Zhihu users (Fang, 2020). Users of the meme refer to Jiang as "the elderly" (长者), in connection to a quote from his famous heated exchange with journalists from Hong Kong during a press conference. Using this term as the query word, the algorithm successfully associates it with other terms in the corpus of the meme, such as "蛤蛤," "人生经验," "闷声发大财," and "谈笑风生."

[21] I pre-process the data with sentence tokenization, as word embeddings are suitable for sentences, not long articles such as answers on Zhihu.

Token	Similar Tokens
政府(Government)	ZF — zf — 中央政府— 中国政府 美国政府— 执政党— 联邦政府 当地政府— 执政者— 日本政府
共产党(CCP)	共党— 国民党— gcd — 中共— tg 中国共产党— GCD — TG 共产党人— 我党
腐败(corruption)	贪腐— 贪污腐败 贪污— 腐败问题 清廉— 廉洁— 贪官 贪— 官僚主义— 官员
长者(the elder)	蛤蛤— 一位长者— 人生经验— 蛤 某位长者— 长辈 闷声发大财— 谈笑风生

TABLE 1: Measurement Using Word Embedding Models

The second step of ETM is training a generative model that draws documents (i.e., answers) from a set of topics, each of which is a distribution over the vocabulary.[22] Latent Dirichlet Allocation, the traditional topic modeling method, uses a document-term matrix approach to generate features from documents, represents both topics the vocabulary with one-hot encoding Blei et al. (2003). Thus, the model has no information about the semantics of words that would have been revealed by its local structure. This setup makes it vulnerable to a large vocabulary and unexpected stop words, both being the exact features of the social media political discourse types I study here.[23] The ETM makes up for this drawback by representing both words and topics with dense vectors in the same numeric space. The model is specified in Algorithm 1, following Dieng et al. (2020):

I train a set of embedded topic models, varying the number of topics from 50 to 400.[24] The

[22]I train word embeddings and then fit them as untrainable parameters into the embedded topic model. A different variant of the model trains word embeddings and the generative model simultaneously. I do not consider this variant because it is less interpretable and, more importantly, performs worse according to the evaluation of Dieng et al. (2020).

[23]Text pre-processing methods, such as removing rare terms and stop words, help to prepare formal text for effective topic modeling. However, it is less effective for social media discourses with large vocabularies due to the informality of the language used on social media (e.g., slangs, memes, and emojis). For these types of data, radical removal of rare terms may cause loss of important sub-communities. In addition, stop word removal with commonly used dictionaries can result in a high false negative rate, leading to the appearance of stop words in many topics.

[24]My implementation is an adapted version of the replication code of Dieng et al. (2020). The model is prototyped with the deep learning framework PyTorch on Python 3. It is trained and evaluated using GPU on Google Colab.

Algorithm 1 Embedded Topic Modeling

1: Draw topic proportions $\theta_d \sim \mathcal{LN}(0, I)$
2: For each word n in the document
    1: Draw topic assignment $z_{dn} \sim$Cat(θ_d).
    2: Draw word $w_{dn} \sim$ softmax($\rho^T \alpha_{z_{dn}}$).

Note: A draw θ_d from the logistic normal distribution $\mathcal{LN}(.)$ is obtained as: $\sigma_d \sim N(0, I); \theta_d =$ softmax(σ_d).

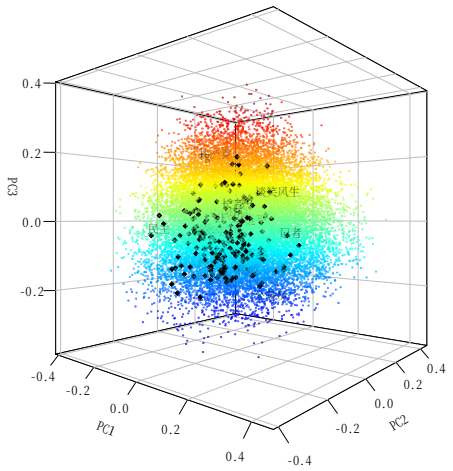
text is pre-processed by removing standard stop words and extremely frequent and infrequent tokens.[25] A trained model with 200 topics is selected based on both quantitative and qualitative evaluations.[26] Figure 5 demonstrates the output of ETM. Panel (a) shows the first three principal components of the word vectors. Panel (b) shows the most prevalent topics in the corpus measured by the summations of θ's. The most popular ones include a diverse set of political topics related to current events in contemporary China (e.g., Taiwan, war, Sino–Japan relations, and the Chinese civil service), topics concerning theoretical or historical inquiries (e.g., freedom and rights, ideology, the KMT, European history, the law and the courts, and democracy), topics about personal experiences, (e.g., student life and family members), topics about specificity (e.g., names of Chinese locations) as well as topics that are not explicitly political (e.g., consumption, travel, and entertainment). Panels (c) and (d) demonstrate the procedure of assigning a concept to a machine-identified topic with an example. Panel (c) shows the first three principal components of the vector representations of the topic and the top 30 words close to it. Based on these words, I consider whether this topic is about the Chinese government. Panel (d) shows the first 100 characters of 4 answers the algorithm identifies as the most associated with the topic. Learning these four articles to discuss different governmental agencies further confirms the concept assigned to this topic. I use the same procedure to interpret all the identified topics of interest discussed in this paper.

4.2.2 Sentiment Analysis

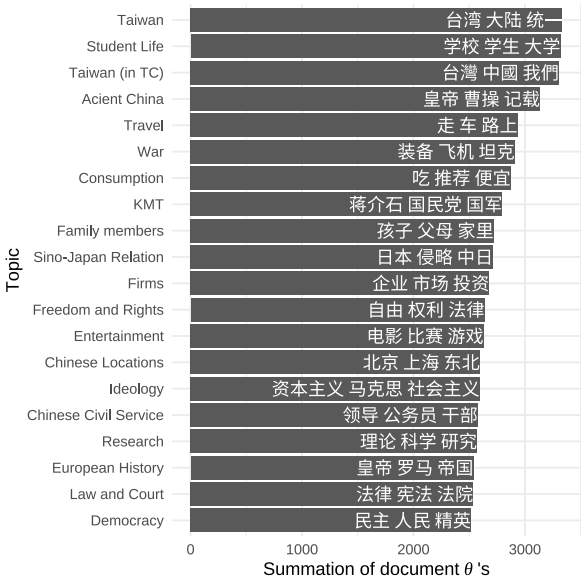
Finally, I use sentiment analysis to gauge the emotions contained in the answers. One limitation of the ETM is that it does not provide systematic measures of sentiment or emotions expressed in

[25] Words or phrases appearing in less than 100 and more than 70% of the answers are removed.

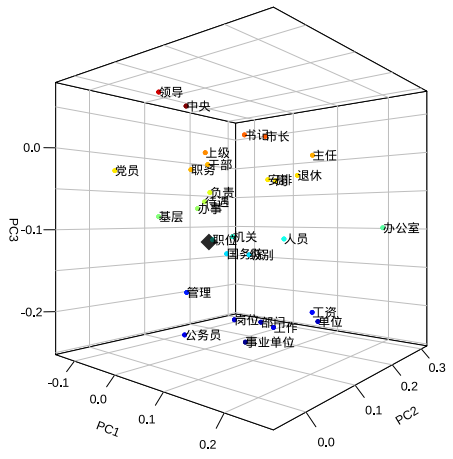
[26] In topic modeling, choosing the number of topics is an art rather than a science. Optimization of quantitative evaluation metrics, such as topic coherence, topic diversity, and document completion task perplexity, does not guarantee interpretable topics Chang et al. (2009). The choice is a result of both quantitative evaluation metrics and manual examination of the text. For trained models with reasonable results in quantitative evaluations, I read samples of topics and their representative documents and words under different model specifications. The objective is to balance topic redundancy, separability, and the inclusion of less frequent topics.



(A) First 3 PC of Embedded Words



(B) Most Frequent Topics



(C) Example Words of Topic *Corruption*

ID	θ	text
690489	0.94	各级国资委领导任命及上下级关系国务院国资委是属于国务院直属的特设机构,代表国家出资人的身份,监管范围是中央所属企业（不含金融类企业）。国资委是国务院直属特设机构，不是国务院组成部门。直属机构的行政长官
670135	0.94	个人戏测，请勿当真。降息决策过程可能有两条线，一条是：货政司根据经济形式和数据研究后起草调整利率请示→人民银行党委会讨论通过→司长签字→主管副行长签字→行长签字报国务院→主管副秘书长签字→副总理签字→
322765	0.94	中纪委是副国级的，所以中纪委副书记是正部级干部。在诸如发改委、公安部、中组部、政法委等重要的党政机关，领导高配是很普遍的现象，所以中纪委有副国级的副书记非常正常。赵洪祝是中纪委第一副书记，更是中央书记
698461	0.90	常务副总理：张高丽　　负责：国务院常务工作；负责发展和改革、财政税务、国土资源、环境保护、住房和城乡建设、三峡工程、南水北调工程工作；协助负责：经济和社会发展、审计、能源工作。　　分管

(D) Example Answers of Topic *Corruption*

FIGURE 5: Embedded Topic Modeling

the answers.[27] As a supplement, I use a lexicon developed by the NLP Laboratory of the Dalian University of Technology. The lexicon stands out among a few existing options for the fine-grained measures in the types of emotions as well as their intensities. Instead of classifying emotions into simple "positive" and "negative" categories in line with most other Chinese sentiment analyses, I split the positive sentiments into 8 categories and the negative ones into 12 categories. In addition, each sentiment word is assigned a fine-grained intensity score ranging from 1 to 9. Figure 6 shows the prevalence of the 30 types of sentiments in the answers identified.[28] The average intensity is low; in other words, many answers contain no sentiment. "Blame" is the most prevalent negative sentiment, and the average intensity of the anonymous answers exceeds that of the regular answers. "Praise" is the most prevalent positive sentiment, and the average intensity of the regular answers exceeds that of the anonymous answers.

5 Empirical Models and Results

To examine what kinds of political discussions are more likely to be subject to self-censorship, I fit the four sets of text features discussed in the previous section as predictors' anonymous answers with four types of statistical learning models. The first empirical inquiry of interest concerns what kinds of political questions create reputational sanctions inducing self-censorship. I fit Elastic Net Models to predict the proportion of anonymous answers to questions using the document-term matrix. As a supplemental analysis, I fit a Poisson Regression with an offset to explain the number of anonymous answers with labels of the manually coded questions. The second empirical inquiry concerns what kinds of expressions in answers are associated with anonymous postings. I fit the normalized scores of the answers in topics (θ) to predict statuses of anonymity with LASSO and select and interpret the best predictors. As a supplemental analysis, I fit a Logistic Regression using the sentiment scores of the answers as independent variables.

Note that these models are chosen for interpretability, not predictive performance. The goal of my empirical inquiries is to make sense of the text features associated with the outcome variable, not to build an artificial intelligence with the best capacity to predict the behavior of anonymous posting. Admittedly, more flexible machine learning models, such as ensemble models and deep neural network models, can produce better predictive performance than the methods employed in this section. However, they are not selected due to their inability to systematically identify

[27] A few topics are predominantly words of a positive or negative nature. For example, a "name-calling" topic (referring to users' nationalistic name-calling) is identified. Nevertheless, it is not a systematic identification.

[28] An answer's score on a sentiment type is the summation of sentiment intensities normalized by the length of its text.

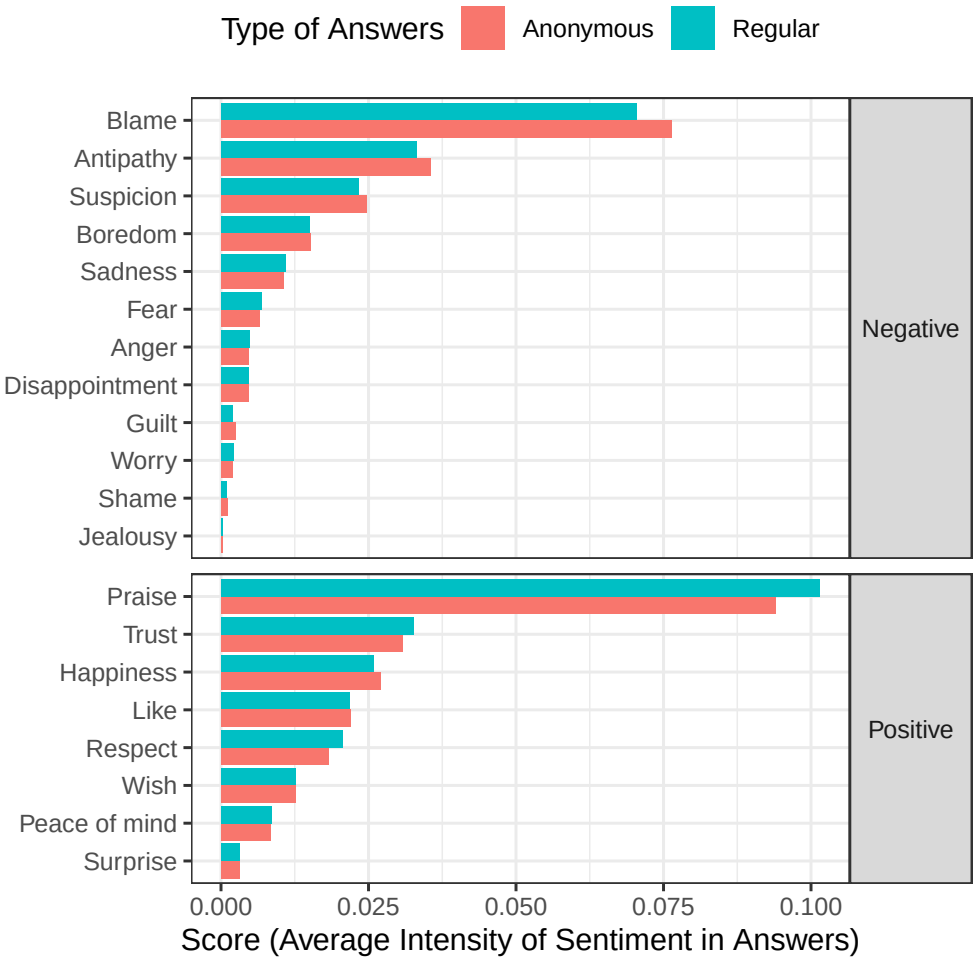


FIGURE 6: Sentiment Analysis

what text features are positively or negatively associated with the outcome variable.[29] That said, although some predictive performance is sacrificed for interpretability, the empirical models are still reliable because the features are carefully engineered, as discussed in the previous section, and the models are carefully tuned with the parameter grid search.

5.1 Questions That Induced Anonymous Answers

I fit Elastic Net Models to select variables to identify the words and phrases that are the best predictors of the proportion of anonymous answers to a question. Using the document-term matrix, a part of the text features (which are unigrams, bigrams, and trigrams in my case) are expected to be highly correlated, because words belonging to the same concept can co-appear in documents frequently. The goal of this empirical inquiry is to interpret which concepts, not words, are the best predictors of the outcome variables. Hence, a model capable of selecting important highly correlated predictors in a group, instead of choosing only one of them, is required. The Elastic Net Model fits this need, given its combination of L1 and L2 regularizations (Zou and Hastie, 2005). Formally, the estimates of the Elastic Net Model are specified below:

$$\hat{\beta} \equiv \underset{\beta}{\operatorname{argmin}}(\|y - \mathbf{X}\beta\|^2 + \gamma_2 \|\beta\|^2 + \gamma_1 \|\beta\|_1) \quad \text{where } \|\beta\|_1 = \sum_{j=1}^{p} |\beta_j|$$

where X denotes the document-term matrix of questions, y refers to the proportion of anonymous answers to the questions, and β denotes the coefficients.

The best Elastic Net Model selected 86 out of 1,504 tokens as features positively associated with the proportions of anonymous answers.[30] Figure 7 summarizes the results. Panel (a) is a word cloud of all 86 selected tokens, the size of the text being the size of the individual coefficients. Panels (b) and (c) interpret the results. Upon reading the words and questions containing them, I categorize them into 10 categories and then calculate the sums of the coefficients of the variables by category. As shown in Panel (c), words about the Chinese regime are the most strongly associated with the proportion of anonymity. The second strongest category of predictors concerns student and campus life. This is likely a result of the demographic composition of the forum: the

[29] With Random Forest Models, variable importance sheds light on the most important predictors of the outcome. However, it cannot reveal the directions of the association. Similarly, for Deep Neural Network Models, such as Convolutional Neural Networks and Transformers, visualization tools are available to provide piecemeal interpretations of that part of the text activating neurons. Nonetheless, they do not allow a systematic interpretation either.

[30] I tune the Elastic Net Model with a grid search along the size of the L2 penalty. Out-of-sample Mean Squared Error (MSE) is used as the evaluation metric. The best model achieves an MSE of 0.026.

users are young, predominantly college-educated, and likely to talk about their ongoing or recent lives as students. The third prominent category, "debate," consists of words that ask people to take a clear position on an issue or join a debate. The fourth prominent category, "personal experience," contains words related to the sharing of stories about family, romance, or encounters in daily life. Moreover, discussions about general social issues that are not necessarily political ("society"), religion, and some foreign entities are selected as strong predictors, although generally at much smaller magnitudes. The category "theory," which indicates questions about specific types of political theories, induces anonymity. Finally, questions mentioning a specific time and space are found to have received more anonymous answers.

The results are supplemented by an analysis with the manually labeled questions. I fit a Poisson Regression explaining the counts of anonymous answers to questions with manually assigned labels informed by the above inductive analysis. To account for the popularity of the different questions, the total counts of answers are added as an offset. Formally,

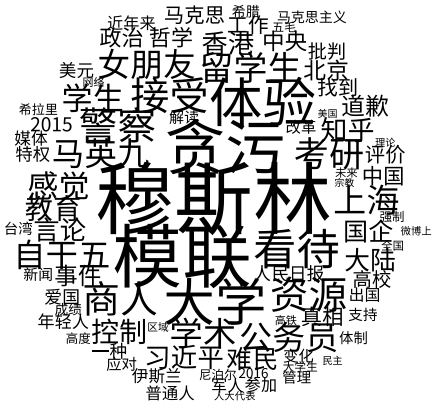
$$\log(E(\text{N of Anonymous Answers}|\mathbf{X})) = \log(\text{Total N Answers}) + \mathbf{X}\beta$$

The result is highly consistent with those from machine-generated text features, while providing more nuance. Figure 8 shows the estimated coefficients. Questions about the Chinese regime and social issues in China are positively associated with more anonymous answers. In contrast, discussions about foreign regimes or social issues are negatively associated with anonymous answers (although less statistically significantly so). Both theoretical and historical inquiries dissuade the use of the anonymous option.[31] Requests for sharing personal experience provide a strong incentive for anonymous answering. Questions about specific political entities, including those referring to a certain news item, time, or place, received more anonymous answers. Importantly, the results of this work revealed an important phenomenon that the machine-coded approach is unable to detect: questions asked with a negative tone pushed users toward anonymity, and this finding is statistically significant. However, questions asked with a positive tone effected the opposite (although this finding is not statistically significant).

5.2 Answers That Are More Likely To Be Anonymous

To understand the kinds of political expression associated with anonymous answering, I fit LASSOs to find topics with the strongest predictive power for the anonymity statuses of the answers.

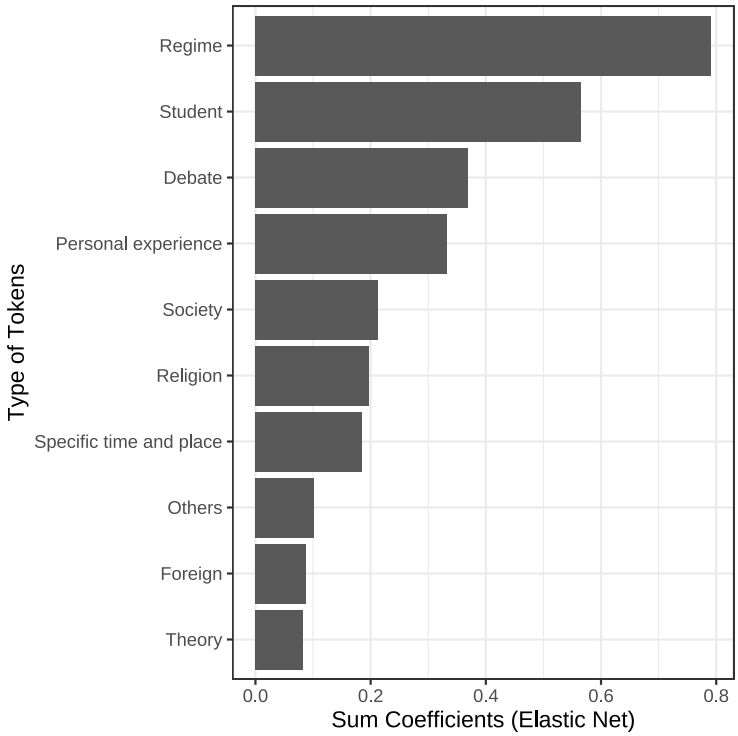
[31] The bag-of-the-words approach finds a few specific topics of theoretical discussion, such as Marxism and democracy, which are self-censorship-inducing. This result shows that the broader topic of theoretical discussion does not induce self-censorship, which is in line with the intuition.



(A) Tokens Selected by Elastic Net

Regime	贪污 警察 公务员 马英九 控制 习近平 言论 香港 大陆 国企 中央 中国 爱国 人民日报 特权 媒体 军人 改革 新闻 台湾 强制 全国 人大代表
Student	模联 大学 考研 留学生 学术 学生 高校 年轻人 成绩 大学生
Debate	看待 接受 自干五 道歉 真相 评价 批判 解读 支持 五毛
Personal experience	体验 女朋友 感觉 工作 找到 普通人 参加 出国
Society	资源 商人 教育 管理 高铁 网络
Religion	穆斯林 伊斯兰 宗教
Specific time and place	上海 事件 北京 2015 近年来 2016 未来 区域
Others	知乎 一种 应对 变化 高度 微博上
Foreign	难民 美元 希腊 希拉里 尼泊尔 美国
Theory	政治_哲学 马克思 体制 马克思主义 民主 理论

(B) Selected Tokens by Category



(C) Sums of Coefficients by Category

FIGURE 7: Question-level Results with Document-Term Matrix

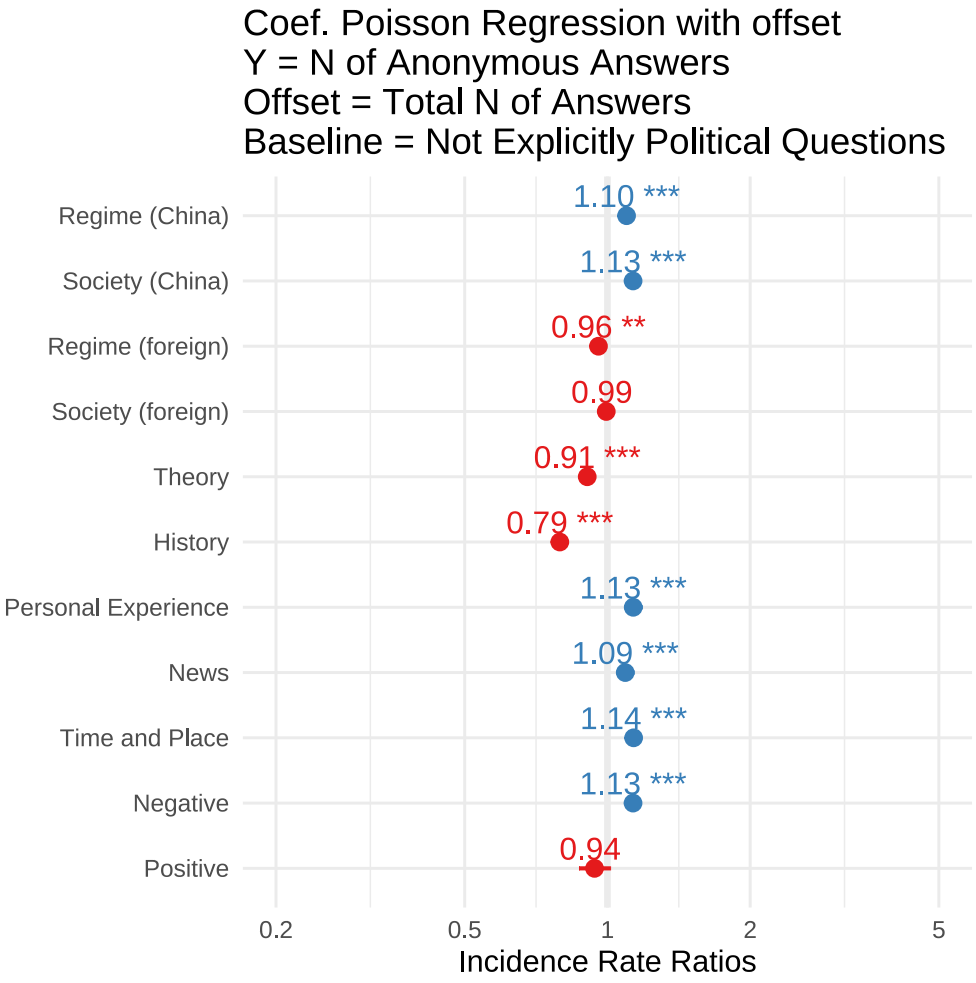


FIGURE 8: Question-level Results with Manual Labels

The L1 regularization is chosen to tackle topic redundancy. As discussed briefly in Section 4, this involves balancing topic redundancy, separability, and the inclusion of less frequent topics. Thus, to include "niche" topics, it is necessary that redundancy of more prominent topics be tolerated to some extent. As a result, the chosen 200 topic text features show a considerable number of redundant topics, that is, some topics map onto the same concept. Thus, I require a modeling strategy with a goal different from that of the model using the document-term matrix of the questions. For a group of highly correlated topics, only the strongest predictor is to be selected. LASSOs are ideal for this goal.

The best model selects 28 among the 200 topics as positive predictors of the outcome, of which 24 have interpretable meanings. Upon manual examination, I classify these selected topics into seven categories. Figure 9 summarizes the results. From the top to the bottom, the topics are grouped by categories, and the categories are sorted by sums of coefficients, indicating the strengths of the categories as predictors of anonymity, in descending order. Topics on personal experience have the highest predictive power for anonymity. Topics in this group include discussions about family members, sharing of events in the users' personal lives, romantic relationships, and friendships. The second strongest group of predictors involves political discussions about the Chinese regime. Topics in this category include the Chinese government (i.e., names of governmental organizations and titles of its officials), corruption, and the Communist Party. Taiwan is also a political topic strongly associated with anonymity. In addition, the users opt for anonymity when discussing the media, censorship, and propaganda. Among the discussions about foreign countries in the forum, the topic on Sino–Japan relations stands out as the only positive predictor of anonymity. Engaging in serious debate is the third largest predictor of anonymity. This category includes a topic that resulted in the use of name-calling words and a topic on functional words. The latter was started to allow users to assume clear positions in the discussions. Topics on social issues are also strong predictors of anonymity and included discussions on women and minorities, the doctor–patient relationship, inequality, and social service in general. The next best predictors of anonymity comprised special topics on personal experience and student lives on campus. Religion, a special discussion type, also shows considerable predictive power for anonymity. Finally, mentioning specific entities, including a specific news item, place, and time, increases the odds of anonymous posting. Note that the "news" topic included in this category refers to a particular news item about an incident connected to the backlash against a Taiwanese artist for allegedly supporting Taiwan's independence. Hence, its effect is likely a mixture of a politically sensitive topic about the regime and its specificity as a news item.

Using sentiment scores as features, I investigate how emotions in political expressions are

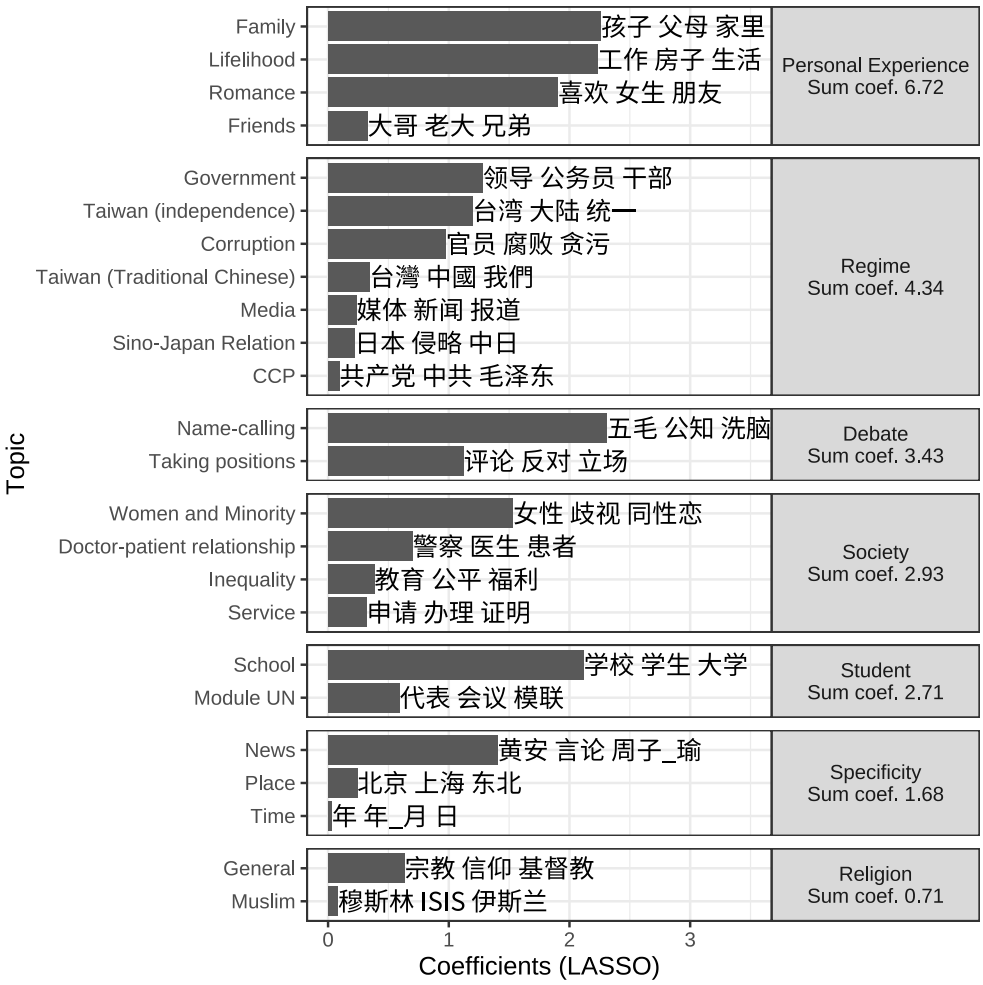
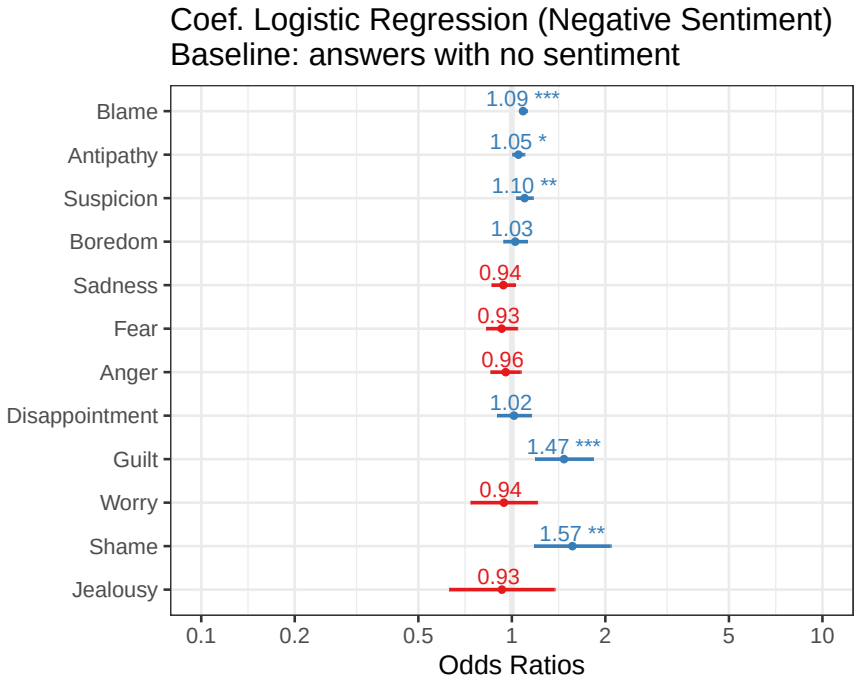
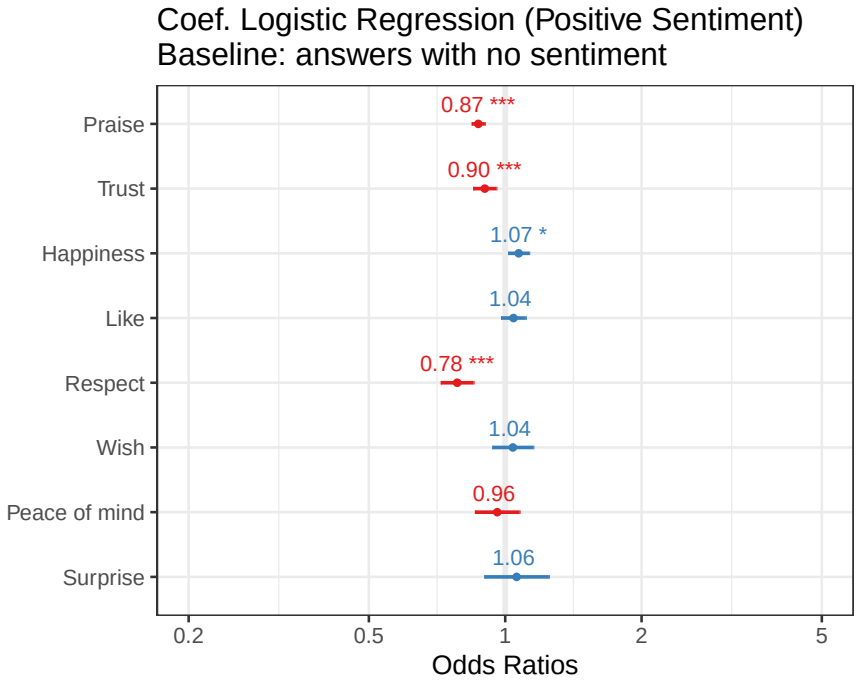


FIGURE 9: Answer-level Results: Embedded Topics Selected by Logit Model as Predictors of Anonymity

associated with anonymous postings. Figure 10 shows the odds ratios of 12 negative and 8 positive sentiments as predictors of anonymous answering. As shown in Panel (a), five types of negative sentiments have statistically significant positive associations with anonymity: blame, antipathy, suspicion, guilt, and shame. In contrast, three types of positive sentiments are negatively associated with anonymity at a statistically significant level: praise, trust, and respect. This result means that when people show more trust and respect and praise others more, they are not as compelled to use the anonymity option to hide their virtual identities.



(A) Negative Sentiment



(B) Positive Sentiment

FIGURE 10: Answer-level Results with Sentiment Scores

5.3 Discussion: Reputational Self-Censorship Extends and Supplements State-Focused Self-Censorship

The empirical inquiries into the political discussions on Zhihu, cross-validating one another, can be summarized as five findings of reputational self-censorship. The first two suggest reputational self-censorship *extends* state-focused self-censorship. The remaining three suggest reputational self-censorship *supplements* state-focused self-censorship.

Two findings suggest that reputational self-censorship *extends* state-focused self-censorship. That is, it affects political communication in the same way as state-focused self-censorship. First, people reputationally self-censor in discussions about regime support. Specifically, a considerable amount of self-censorship is evident in questions and answers about governmental organizations and officials, the Communist Party, corruption, and media control. This shows fear of social sanctions can deter citizen from engaging in discussions about critical issues on regime support. This finding is intuitive as it resonates with the theme of almost all previous research on (state-focused) censorship (see, for example, Jiang and Yang, 2016; Robinson and Tannenberg, 2019; Shen and Truex, 2020). But it brings new knowledge by providing rare empirical evidence that fear of social sanctions, on top of fear of the state, is an important motivation for this most expected pattern of self-censorship.

The second evidence that reputational self-censorship *extends* state-focused self-censorship connects to a recent study. My results show that citizens reputationally self-censor discussions on specific political entities. The evidence pertaining to both questions and answers shows that Zhihu users self-censor discussions pertaining to specific times, places, and events. This resonates with Chang and Manion (2020), who recently showed that people self-censor when referring to focal times and places (namely, the focal-point self-censorship). This finding suggests people are also engaged in focal-point self-censorship for reputational concerns.

Three additional findings suggest that reputational self-censorship *supplements* state-focused self-censorship. That is, it has unique ways to affect political communication. First, the results identify a rich set of topics beyond regime support for which reputational self-censorship is prevalent. In this study, the participants of explicit political discussions self-censor on two topics beyond those about regime support: Taiwan and Sino–Japan relations. While the discussions on these two topics are arguably not politically sensitive enough to justify major worries about state sanctions, they are contentious enough among citizens to raise concerns about reputational sanctions, leading to the observed self-censorship. Similarly, a set of social issues closely related to politics is highly self-censored, namely women and minorities, the doctor–patient relationship, and inequality. These topics are even less sensitive for the state, but the contentious views among users

arguably lead to high levels of self-censorship. Finally, religion is another discussion topic for which a high level of self-censorship is observed. In qualitative terms, discussions on religions and religious practices are prevalent on Zhihu, and users visibly express their concerns about the possibility of fellow participants, particularly those with religious beliefs, imposing reputational sanctions.

Second, the results suggest that citizens self-censor the sharing of personal experiences. Notably, words and topics constituting personal experiences are one of the strongest predictors of anonymity, as indicated by both the question- and answer-level analyses. Two plausible reasons may be attributed to this behavior. First, sharing personal information can lead to personal attacks in the virtual community. While users typically use personal experiences to support arguments in their answers, fellow users may easily deviate by judging the author based on the shared story, leading to reputational sanctions online. Second, as a more severe consequence, sharing personal experiences risks the exposure of one's real personal information. When fellow users are able to associate a user's real-world identity with their virtual identity, the threat of reputational sanctions drastically increases. Topics and words related to students' campus lives constitute special cases of personal experiences as well as strong predictors of self-censorship. As noted in Section 3, users of the forum are typically college-educated and young, making the sharing of their experiences as students a prominent selection by the model.

Third, the results suggest that people reputationally self-censor to avoid engagement in debates and conflicts. Both question- and answer-level analyses show that people self-censor their explicit adoption of a stance in arguments. The fact that questions with negative sentiments, a name-calling topic, and the negative sentiment analysis of the answers each predict anonymity, on the other hand, suggest that people self-censor their involvement in intense, and sometimes unfriendly, debates. More generally, this implies that the questions or answers alone (i.e., the substantive content) do not matter; the manner in which the questions are asked and answered is also important. A question that requests responders to explicitly take a stance may create social pressure and raise concerns about reputational sanctions. In turn, when users consider making their arguments in an assertive manner, they may take precautions against reputational sanctions.

The three newly discovered patterns of reputational self-censorship add to our understanding of political communication both under authoritarian context and from a broader comparative perspective. First, specific to the authoritarian context, they reveal a hidden part of self-censorship that the scholarship has overlooked. While previous research focused on a narrowly defined set of issues on regime support, the results show the prevalence of self-censorship in a broader set of political and social topics. In addition, while previous research lacked the empirical leverage to study how

the dynamics of social interactions could induce self-censorship, my results show self-censorship vary by the level of exposure and intensity in political conversations in authoritarian China. Such discoveries enhance our understanding of self-censorship as the backbone of authoritarian stability: We learn that self-censorship suppresses the exchange of opinions on a broader set of topics, sharing of personal experience that could have drawn empathy, and engagement in serious political debate. All three can potentially disorganize anti-regime collective actions, which deepens out understanding on how political communication in the social media era may help authoritarian rulers.

Second, beyond the authoritarian context, the findings speak to classic and frontier empirical research on political communication around the world. The first finding resonates with findings in the classic literature on social desirability tested in democratic countries that citizens hold back their genuine views on socially contentious issues. The second and third findings further connect with an emerging literature on political polarization in democracy in the social media era. Hence, on top of their unique implications for authoritarian politics as discussed above, these findings also make authoritarian China an interesting case for comparative studies on political communication, which can lead to future theoretical and empirical innovations.

6 Conclusion

This paper makes theoretical as well as empirical contributions to our understanding of self-censorship. Self-censorship, the behavior of withholding genuine political views from others, has long been considered a prevalent political behavior with important implications for authoritarian stability. Existing works primarily treat the behavior as an outcome of state sanctions: citizens self-censor to avoid being punished by the state for expressing their opinions. This theoretical perspective has undoubtedly captured an important part of the motivations behind the behavior, but it has also overlooked a different element, namely the characteristics of political talk. A considerable proportion of the political talk in authoritarian regimes, such as China, where public discussions including criticism against the state are tolerated, takes place as a part of citizens' everyday social activities. When people talk politics in their everyday lives, their intended audience should be primarily fellow citizens, not the state. In this manner, social pressure should be an important, if not dominant, motivation for self-censorship in everyday political talk under authoritarian rule. Reviving the understudied behavioral foundations of preference falsification specified in Kuran's (1995) seminal work, I theorize that self-censorship under social pressure is an outcome of the fear of reputational sanctions, namely social punishment enforced by the audience that dislikes

a person's expressed political view. I argue that people self-censor expressions of political opinion when the cost, namely reputational sanctions, outweigh the benefits, namely the intrinsic and expressive utilities.

This new theoretical perspective is tested using original online discourse data in authoritarian China. The use of discourse data from Zhihu, a popular Chinese question-and-answer forum, helps overcome two empirical challenges. First, its question-and-answer format allows the gauging of multi-dimensional political opinions without being overwhelmed by the messiness of the data. Second, its unique "anonymity" option facilitates observation of the discourses that would otherwise have been subject to reputational self-censorship.

This empirical inquiry finds five distinct patterns of self-censorship as the outcomes of the fear of reputational sanctions, two of which resonate with previous findings on state-focused self-censorship, while the remainder are unique to reputational self-censorship. First, consistent with conventional wisdom, people evidently self-censor on politically sensitive topics related to regime support, implying that reputational sanctions are an extension of state sanctions. Second, self-censorship about discussions of focal times and places is observed, consistent with the result of a recent study that used other data and focused on a different context. The remaining three findings are unique patterns of reputational self-censorship. First, I detect a rich set of topics on political and social issues associated with reputational self-censorship, namely Taiwan, Sino–Japan relations, women, minorities, the doctor–patient relationship, inequality, and religion. Second, reputational self-censorship is prevalent with regard to the sharing of personal experiences. Specifically, citizens self-censor when questions or answers refer to social relations or experiences connected to their daily lives. Third, citizens reputationally self-censor to avoid engaging in debates. Specifically, people are worried about inviting reputational sanctions caused by taking a clear stance in an argument and/or expressing views that can drag them into intense exchanges, possibly leading to expression of negative emotions.

This empirical inquiry employs a combination of text-as-data methods. For the short text in the questions, I use a document-term matrix approach to automatically code the text and supplement it with manual content analysis. For the long text in the answers, I generate text-based measures with ETM to account for the large vocabulary size in the social media data. To my knowledge, this is the first application of these methods in social sciences studies. I also code emotions in answers with a lexicon-based sentiment analysis, which provides fine-grain measures in 12 negative and 8 positive sentiments.

This study has implications for future research on authoritarian mass political behavior and opinion research methods under authoritarian rule. It is a common assumption among students of

authoritarian politics that an authoritarian state is responsible for most of the political phenomena. That is, everything happening among "the ruled" is theorized as resulting from "authoritarian control," following the definition of Svolik (2012). As a result, state-of-the-art studies on authoritarian mass behavior predominantly focus on the dynamics of state–citizen interactions. While I do not contest the importance of this theoretical perspective, my study draws attention to a different angle: mass behavior as an outcome of the dynamics of citizen–citizen interactions under authoritarian rule. Social forces take over spaces left open to the public by the state. It is critical to understand the social force behind authoritarian mass behavior, because it is an important micro-foundation of authoritarian politics and has implications for authoritarian control. For example, to extend this study, questions may be asked about how reputational sanctions in political talk can help or harm the authoritarian state or how the state can strategically manipulate these sanctions.

This study also has implications for opinion research methods in connection to authoritarian rule. At present, scholars still tend to rely primarily on self-reported data in the field. Hence, understanding when and how participants choose not to share their genuine views is critical to avoid measurement errors. The results of this study indicate that both the manner in which the questions are asked and the answers are given affects the level of self-censorship under an authoritarian context. These specific findings indicate areas for which future scholars may take due precautions when designing their studies. Moreover, this work suggests a new form of data processing in opinion research worth further investigation. Using rich semi-structural discourse data from a question-and-answer forum, this study uncovers patterns of interest in a manner that neither completely structured surveys nor completely unstructured social media discourses can achieve. Future works on authoritarian mass opinion may consider data generated in similar forms with innovative open-ended survey or laboratory experiment designs based on semi-structured political discussions.

References

Blaydes, L. and R. M. Gillum (2013). Religiosity-of-interviewer effects: Assessing the impact of veiled enumerators on survey response in Egypt. *Politics and Religion 6*(3), 459–482.

Blei, D. M., A. Y. Ng, and M. I. Jordan (2003). Latent Dirichlet Allocation. *Journal of Machine Learning Research 3*, 993–1022.

Brownback, A. and A. Novotny (2018). Social desirability bias and polling errors in the 2016 presidential election. *Journal of Behavioral and Experimental Economics 74*(March), 38–56.

Chang, C. and M. Manion (2020). Political Self-Censorship in Authoritarian States: The Spatial-Temporal Dimension of Trouble. *Comparative Political Studies forthcoming*.

Chang, J., S. Gerrish, C. Wang, and D. M. Blei (2009). Reading Tea Leaves: How Humans Interpret Topic Models. *Advances in Neural Information Processing Systems 22*, 288—-296.

Chen, J. and Y. Xu (2017). Why do authoritarian regimes allow citizens to voice opinions publicly? *The Journal of Politics 79*(3), 792–803.

Cilliers, J., O. Dube, and B. Siddiqi (2015). The white-man effect: How foreigner presence affects behavior in experiments. *Journal of Economic Behavior and Organization 118*, 397–414.

Coffman, K. B., L. C. Coffman, and K. M. M. Ericson (2017). The size of the LGBT population and the magnitude of antigay sentiment are substantially underestimated. *Management Science 63*(10), 3168–3186.

Coppock, A. (2017). Did Shy Trump Supporters Bias the 2016 Polls? Evidence from a Nationally-representative List Experiment. *Statistics, Politics and Policy 8*(1), 29–40.

Dieng, A. B., F. J. Ruiz, and D. M. Blei (2020). Topic modeling in embedding spaces. *Transactions of the Association for Computational Linguistics 8*, 439–453.

Fang, K. (2020). Turning a communist party leader into an internet meme: the political and apolitical aspects of china's toad worship culture. *Information, Communication & Society 23*(1), 38–58.

Frye, T., S. Gehlbach, K. L. Marquardt, and O. J. Reuter (2017). Is Putin's popularity real? *Post-Soviet Affairs 33*(1), 1–15.

Germano, F. and M. Meier (2013). Concentration and self-censorship in commercial media. *Journal of Public Economics 97*, 117–130.

Hale, H. E. (2013). Regime Change Cascades: What We Have Learned from the 1848 Revolutions to the 2011 Arab Uprisings. *Annual Review of Political Science 16*(1), 331–353.

Holbrook, A. L. and J. A. Krosnick (2010). Social desirability bias in voter turnout reports: Tests using the item count technique. *Public Opinion Quarterly 74*(1), 37–67.

Jiang, J. and D. L. Yang (2016). Lying or believing? Measuring preference falsification from a political purge in China. *Comparative Political Studies 49*(5), 600–634.

King, G., J. Pan, and M. Roberts (2013). How censorship in China allows government criticism but silences collective expression. *American Political Science Review 107*(917), 326–343.

King, G., J. Pan, and M. E. Roberts (2014). Reverse-engineering censorship in China: Randomized experimentation and participant observation. *Science 345*(6199), 1251722–1251722.

King, G., J. Pan, and M. E. Roberts (2017). How the Chinese government fabricates social media posts for strategic distraction, not engaged argument. *American Political Science Review 111*(3), 484–501.

Krumpal, I. (2013). Determinants of social desirability bias in sensitive surveys: A literature review. *Quality and Quantity 47*(4), 2025–2047.

Kuklinski, J. H., P. M. Sniderman, K. Knight, T. Piazza, P. E. Tetlock, G. R. Lawrence, and B. Mellers (1997). Racial Prejudice and Attitudes Toward Affirmative Action. *American Journal of Political Science 41*(2), 402.

Kuran, T. (1989). Sparks and prarie fires: A theory of unanticipated revolution. *Public Choice 61*(1), 41–74.

Kuran, T. (1991a). Now out of Never: The Element of Surprise in the East European Revolution of 1989. *World Politics 44*(01), 7–48.

Kuran, T. (1991b). The East European Revolution of 1989: Is it surprising that we were surprised? *American Economic Review 81*(2).

Kuran, T. (1995). *Private Truth, Public Lies: The Social Consequences of Preference Falsification*. Cambridge and London: Harvard University Press.

Leibold, J. (2011). Blogging alone: China, the internet, and the democratic illusion? *Journal of Asian Studies 70*(4), 1023–1041.

Lorentzen, P. (2014). China's strategic censorship. *American Journal of Political Science 58*(2), 402–414.

Lynch, M. (2011). After egypt: The limits and promise of online challenges to the authoritarian Arab state. *Perspectives on Politics 9*(2), 301–310.

Mikolov, T., K. Chen, G. Corrado, and J. Dean (2013). Efficient Estimation of Word Representations in Vector Space. pp. 1–12.

Miller, B. (2018). The limits of commercialized censorship in China. *Working paper*.

Pan, J. and Y. Xu (2018). China's ideological spectrum. *The Journal of Politics 80*(1), 254–273.

Powell, R. J. (2013). Social Desirability Bias in Polling on Same-Sex Marriage Ballot Measures. *American Politics Research 41*(6), 1052–1070.

Qin, B., D. Strömberg, and Y. Wu (2017). Why does China allow freer social media? protests versus surveillance and propaganda. *Journal of Economic Perspectives 31*(1), 117–140.

Roberts, M. E. (2018). *Censored : Distraction and Diversion inside China's Great Firewall*. Princeton University Press.

Robinson, D. and M. Tannenberg (2019). Self-censorship of regime support in authoritarian states: Evidence from list experiments in China. *Research and Politics 6*(3).

Shen, X. and R. Truex (2020). In Search of Self-Censorship. *British Journal of Political Science*, 1–13.

Stockmann, D. (2013). *Media commercialization and authoritarian rule in China*. Cambridge University Press.

Stockmann, D. and T. Luo (2017). Which social media facilitate online public opinion in China? *Problems of Post-Communism 64*(3-4), 189 – 202.

Svolik, M. W. (2012). *The Politics of Authoritarian Rule*. Cambridge: Cambridge University Press.

Truex, R. and D. L. Tavana (2019). Implicit attitudes toward an authoritarian regime. *Journal of Politics 81*(3), 1014–1027.

Zhang, H. and J. Pan (2019). *CASM: A Deep-Learning Approach for Identifying Collective Action Events with Text and Image Data from Social Media*, Volume 49.

Zhou, Y. J., W. Tang, and X. Lei (2020). Social Desirability of Dissent: an IAT Experiment with Chinese University Students. *Journal of Chinese Political Science 25*(1), 113–138.

Zou, H. and T. Hastie (2005). Regularization and variable selection via the elastic net. *Journal of the Royal Statistical Society. Series B: Statistical Methodology 67*(5), 301–320.

徐琳宏, 林鸿飞, 潘宇, 任惠, and 陈建美(2008). 情感词汇本体的构造. 情报学报 *27*(2), 180–185.

Appendices

A Embedded Topic Modeling Evaluation **2**

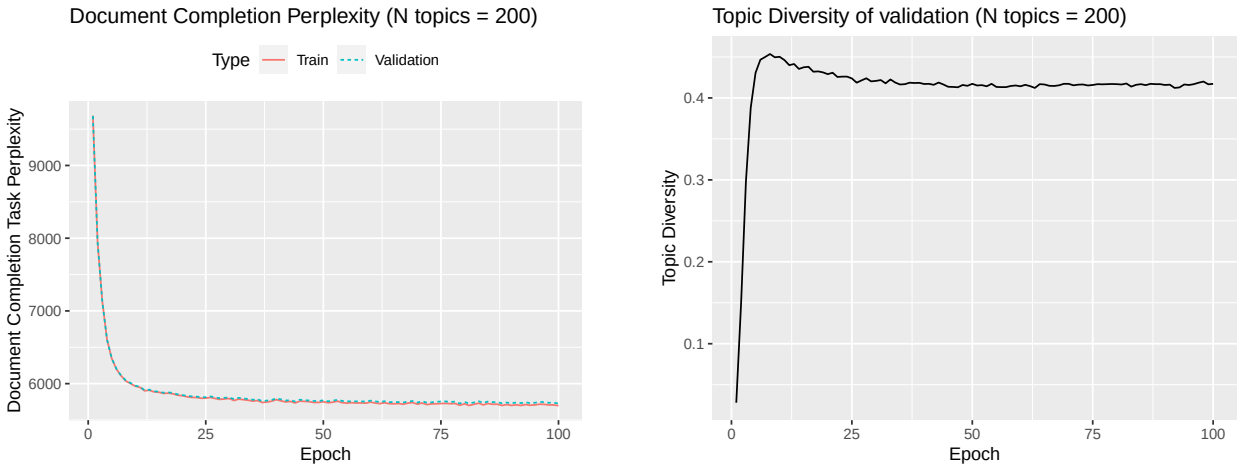
B How Question Types Explain Anonymity **3**

C Topics Selected as Predictors of Anonymity **4**

D How Sentiment of Answers Explains Anonymity **16**

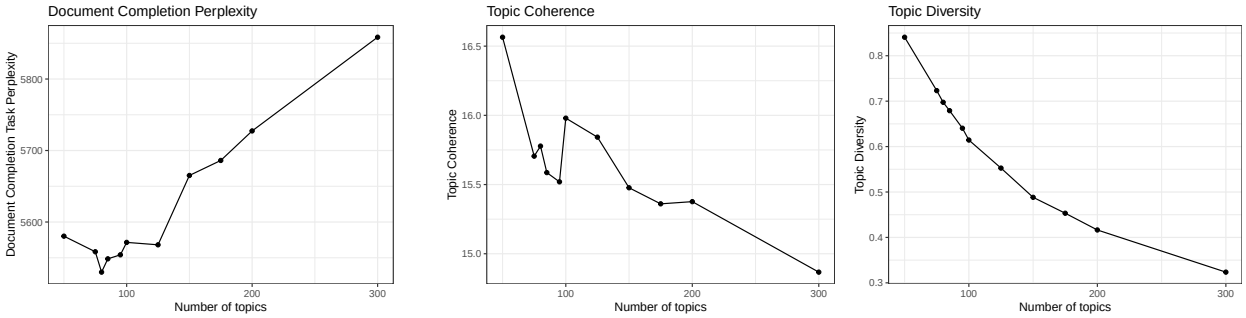
A Embedded Topic Modeling Evaluation

FIGURE A.1: Training of the Selected 200-Topic Model



(A) Document Completion Perplexity (B) Topic Diversity

FIGURE A.3: Evaluation Metrics of Models with 50-300 Topics



(A) Perplexity (B) Topic Coherence (C) Topic Diversity

B How Question Types Explain Anonymity

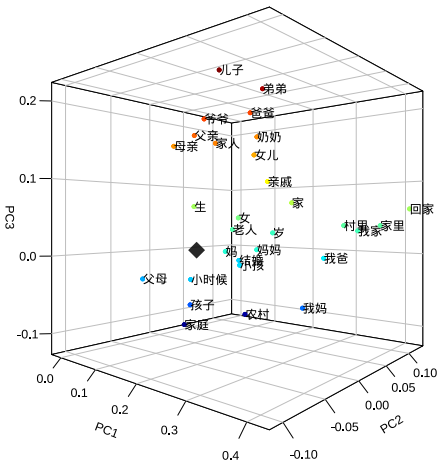
TABLE B.1: POISSON REGRESSION

	DV: N Anonymous Answer
Regime (China)	0.093***
	(0.011)
Society (China)	0.124***
	(0.014)
Regime (foreign)	−0.044***
	(0.016)
Society (foreign)	−0.006
	(0.019)
Theory	−0.098***
	(0.019)
History	−0.231***
	(0.019)
Personal Experience‘	0.125***
	(0.021)
News	0.087***
	(0.021)
Time and Place‘	0.127***
	(0.012)
Negative	0.124***
	(0.017)
Positive	−0.063
	(0.039)
Constant	−1.819***
	(0.013)
Observations	1,015
Log Likelihood	−3,857.084
Akaike Inf. Crit.	7,738.169

Note: aaaaaa. Heteroskedastic standard errors clustered at prefecture level are reported in parentheses. * $p < 0.05$, ** $p < 0.01$

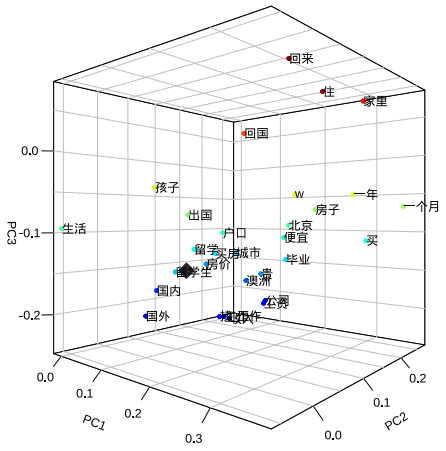
C Topics Selected as Predictors of Anonymity

FIGURE C.1: Personal Experience: Family



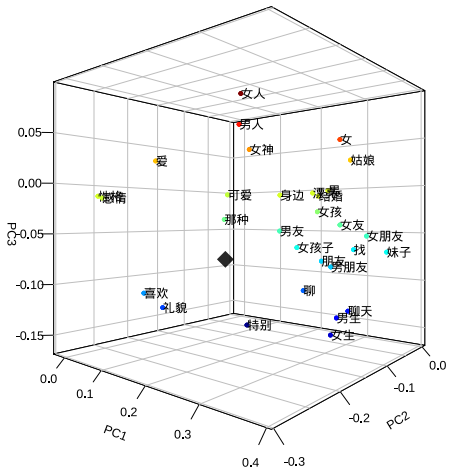
ID	θ	text
42754	0.91	科科，还记得去丈母娘家里吃饭的典故吗。过年，爸爸打电话给大儿子：“你在哪里吃饭。”大儿子：“丈母娘家里。”给二儿子打电话：“你在哪里吃饭。”二儿子：“丈母娘家里。”爸爸：“你们怎么这么没出息，都在丈母
310869	0.79	没看过这篇报道，单纯就“男人比女人更加孝顺”这句话谈谈我的看法。没有大数据，但是我看过的大部分家庭都是女方嫁到男方家，和男方一大家子人生活在一起。就拿我妈妈来说，和爸爸在大学认识，嫁到我们家。虽然娘家
13083	0.77	妈：儿子，等你考上大学，妈妈就享福了。妈：儿子，等你娶了媳妇，妈妈就享福了。妈：儿子，等把你孩子带大了，妈妈就享福了。妈妈总说要享福，结果吃了一辈子的苦。
338371	0.76	父母结婚那年我爸22岁，我妈20岁，为什么结婚呢，因为我那时已经4个月大了。在我的童年记忆里，他们不是在吵架就是在准备吵架，原因是我爸怀疑我妈有了外遇。吵着吵着就打，打完了继续吵，终于在我11岁那年我

FIGURE C.3: Personal Experience: Livelihood



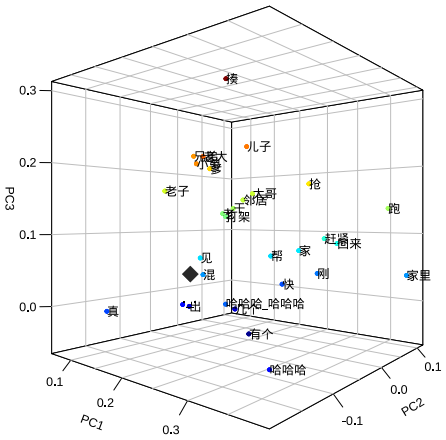
ID	θ	text
223636	0.84	一名公务员，收入不会太高，也不会太低，应该处于当地的中等水平。如果当地是这样的收入，我很想知道54万的房子是多大，地段如何。他家里的贷款，是因为他购入了和收入不符的房子。家里存款少，是因为妻子无工作。
68670	0.80	穷的要死的体验。我现在每月实习工资税后2500多刀，单身狗租个房间500刀，吃饭400刀，娱乐200刀，养车200刀，这就1300刀了。真要是找了妹子，2000刀不够花。一个月攒不下多少钱。话说我们工
310815	0.78	上海25w人民币，你活得紧巴巴的。多伦多八万，还是税前，也不是很宽裕。但是，以我一家三口的生活情况来讲。如果都是不买房，多伦多的八万会比上海的25万舒服。首先你在加拿大没啥人情往来，在国内这块开销不小
310781	0.76	我觉得是，底层人也可以容易过的幸福。加拿大人工贵，做清洁工服务生都可以有20刀一小时，申请低保政府会有补贴，生孩子每个月发六百刀。假如一个单身狗每月收入三千刀（算是比较低的工资了），那么他租房子可能会

FIGURE C.5: Personal Experience: Romance



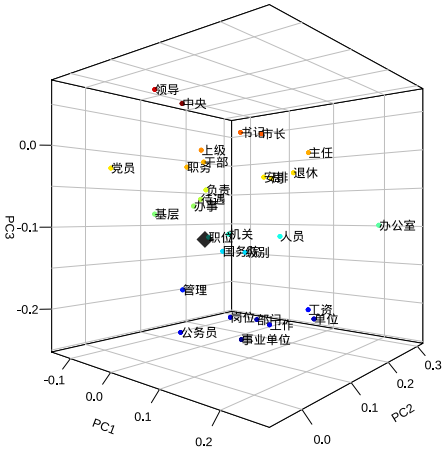
ID	θ	text
255272	0.69	长辈：说什么你听就行了，不要问为什么。奶奶：女孩的裤子不能放在男的上衣上面，就算是洗衣篮里堆着的时候，脏。来月经清明不要去扫墓，脏。妈妈：巴啦巴啦打死你，巴啦巴啦打死你。我爸：不要和男同学来往多。不知
255747	0.65	笑死宝宝了。可能1米57的妹子也开了一个贴。是等一个高富帅出现呢还是将就嫁给你一个屌丝呢。本人一米57 长得很普通 自己其实很想找一个高一点 帅一点 有点钱的男生 但是迫于年纪有点大了 家里催的急
255761	0.63	既然都这么喜欢女神了，那还是追女神吧，别祸害人家姑娘了，你既然嫌弃人家胖，又嫌弃人家矮，种种比不上你喜欢的女神，还纠结什么，所谓“得不到的就是最好的”，你就是跟追你的这个姑娘在一起了，心里也得想着没有
309158	0.62	给题主科普一下165以上的姑娘对男朋友的身高认识 本人女169 就想告诉你别嫌弃姑娘这个姑娘那个的 以你现在这种状态有人喜欢真是烧高香了还有啊一般165左右的姑娘对男朋友的身高幻想在175以上(但也

FIGURE C.7: Personal Experience: Friends



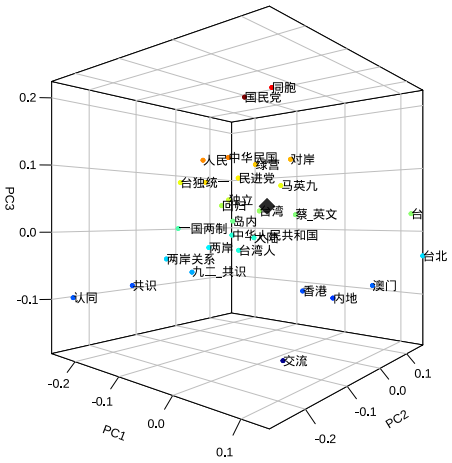
ID	θ	text
309182	0.84	@木月 ——一码归一码。我家有个祖传的金塑但巨丑，所以我每天没事就骂它这么丑还非得摆在客厅的正中间，没事说不准还会拿它垫个桌子什么的。可是忽然有一天有个选房亲戚（关系很不好，多年没来往了）过来二话不说
691802	0.81	我就说一点…你家里兄弟两个，你是老大当家的，隔壁老王家兄弟来你家闹事，你作为老大出门拼他丫的，不仅自己打伤了，还动员了一批丫鬟佣人打，然后你弟弟安安稳稳笼络人心，抢你的当家地位，完了老王打跑了还跟哥哥
146728	0.78	没有出师表，难道要有这东西吗？你亮叔我跟你讲几句：你爸当年出来混，半道上就给挂了；现在地盘又分成三块了，益州好像咱也罩不住了，这世道眼瞅着要杯具了。但是你爸留下的保镖还很忠心啊，出去砸场的那些二杆子也
177451	0.78	首先先说明一下文件编号问题，文件的编号按照发布或者处理的时间顺序而制定的。中央一号文件也就是中央每年发布文件的第一个。本来这些序号没有什么特殊的意义，就像每个人的电话号码一样，仅仅是一个代号。但是，文

FIGURE C.9: Regime: Government



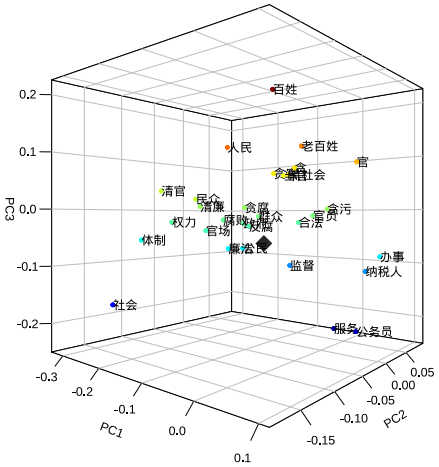
ID	θ	text
690489	0.94	各级国资委领导任命及上下级关系国务院国资委是属于国务院直属的特设机构,代表国家出资人的身份,监管范围是中央所属企业（不含金融类企业）。国资委是国务院直属特设机构，不是国务院组成部门。直属机构的行政长官
670135	0.94	个人戏测，请勿当真。降息决策过程可能有两条线，一条是：货政司根据经济形式和数据研究后起草调整利率请示→人民银行党委会讨论通过→司长签字→主管副行长签字→行长签字报国务院→主管副秘书长签字→副总理签字→
322765	0.94	中纪委是副国级的，所以中纪委副书记是正部级干部。在诸如发改委、公安部、中组部、政法委等重要的党政机关，领导高配是很普遍的现象，所以中纪委有副国级的副书记非常正常。赵洪祝是中纪委第一副书记，更是中央书记
698461	0.90	常务副总理：张高丽　　负责：国务院常务工作；负责发展和改革、财政税务、国土资源、环境保护、住房和城乡建设、三峡工程、南水北调工程工作；　　协助负责：经济和社会发展、审计、能源工作。　　分管

FIGURE C.11: Regime: Taiwan (independence)



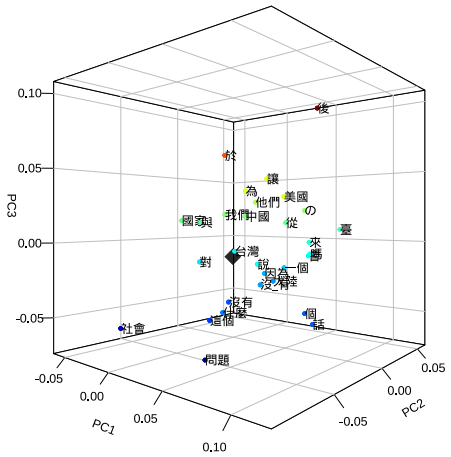
ID	θ	text
33953	0.68	相关法律文件太冗杂，就简单说说现状和定义吧。1945年，台湾便已从日本回归了中华民国。1949年，国共内战国民党领导的国民政府失败，退守台湾。共产党在北京建立中华人民共和国。从此两岸分治。所以，现在的
40209	0.62	我是大陆人，作为一介草民，其实我对台湾问题不太关注，因为台湾不回归，我是草民，将来台湾回归了，我还是个草民，收入待遇不会好很多，我的生活状态基本不变。这些年看腻了台湾所谓的民主，还有台湾许多人对大陆的
37307	0.61	西元1992年，大陆和台湾方面达成92共识，其中最主要的内容是一中各表，什么是一中各表呢，就是一个中国两岸各自表述，大陆方面，一个中国是中华人民共和国，台湾为未解放区，台湾方面，一个中国是中华民国，注
570566	0.59	谢邀台独分子说了，大陆不让他们台独，伤了他们的心，影响了两岸关系。意思就是，大陆人允许台湾人台独，台独分子就高兴了，两岸关系就好了？什么逻辑。两岸关系根本不是台独分子可以定义的，周子瑜事件影响的事实上

FIGURE C.13: Regime: Corruption



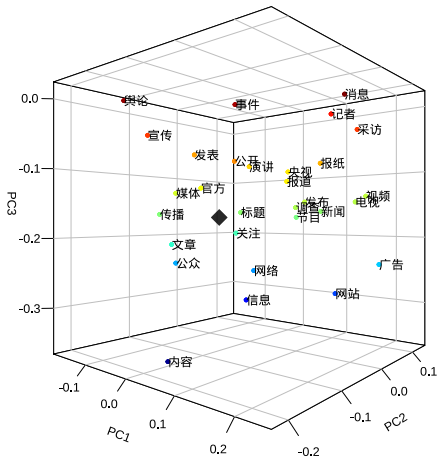
ID	θ	text
555464	0.64	中国经济高速增长是因为起点较低。而中国的腐败一般就是拿钱办事，相较于政府的低效率，贪官办事效率很高呀！而且贪腐只涉及了官员和商人，并没有盘剥底层劳动者。所以我国贪腐才没造成重大影响。
707210	0.64	如果一个清官不能让百姓都过上好日子，说明这个清官没有本事！如果一个贪官上任，但能让百姓都过上好日子，老百姓都很富生活安居乐业，说明这个官有本事！他贪不贪也没人会说什么。如果这个官不贪又有本事，那么这个
46213	0.58	我觉得评价一个官员是成功还是失败还要看政绩啊，即使他是一个清官如果在位期间不作为，或者乱作为我觉得这就是失败。现在社会升官发财观念深入人心，无官不贪逐渐被人认可，人民只是希望官员能够少贪一点，多做点真
68453	0.57	穷官与好官之间不是一个＝符号。身为一方执政者，官员需要的是能写出来的政绩，而百姓需要的是切实的实惠。不怕当官的贪污受贿，就怕不仅只会贪污受贿还没一点能力。尽从百姓身上贪

FIGURE C.15: Regime: Taiwan (Traditional Chinese)



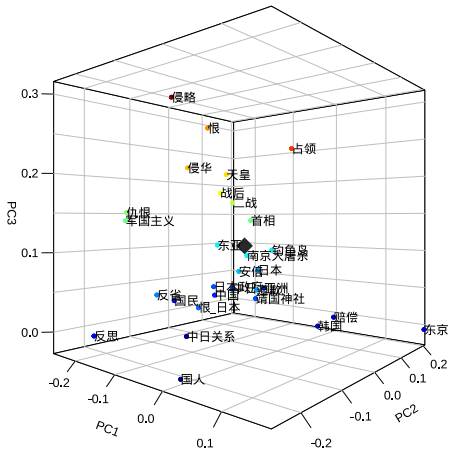
ID	θ	text
648652	0.94	我還蠻喜歡閃靈的～從皇軍到尼古拉斯到暮沉武德殿，特別是民謠版的暮沉武德殿，簡直不能更棒了，聽起來都超爽DER~咳咳，花癡結束了。現在要談的，是立法委員林昶佐先生，而不是閃靈主場飛踢(Freddy)兄。
712968	0.93	這件事我真的不吐不快。作為一個從未去過大陸的台灣人，因為興趣愛好混跡大陸網站5年，非常能了解地區不同，領導不同，民情不同，價值觀的差異高至天南地北，更何況還跨越一個海峽。這五年來我一直包容大陸對台灣的
546100	0.93	我看大陸跟臺灣的討論，並不是一天一夕，也不是從 ptt 開始的，早在臺灣郵網和新聞組的時代，我已經在看了．當年的香港，還差不多是清一色「反對臺灣獨立」，「中國必須統一而且強大大家才有幸福
40050	0.93	以下為朱立倫致辭全文： 馬總統、吳副總統、連前主席、吳前主席、毛院長、王院長、兩位副主席，各位中常委、中評委、黨代表、以及各位先進同志和媒體朋友們，大家好！ 在今天會議的一開始，我必須先向大家報告，我

FIGURE C.17: Regime: Media



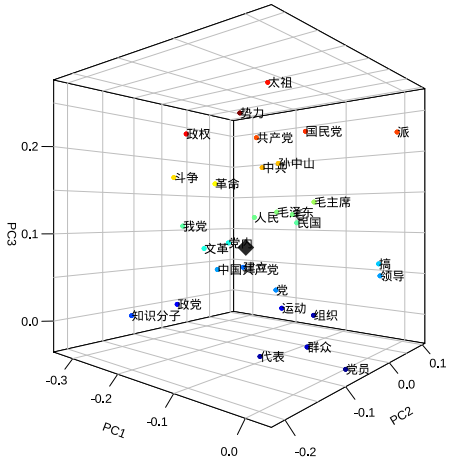
ID	θ	text
618392	0.65	网络编辑每日的KPI就是保证一定的PV，在新闻素材有限的情况下，必须极大地挖掘新闻素材的吸引力，把文章中最刺激的一言半语当标题来吸引用户，产生耸人听闻、误导舆论的效果就不足为怪。另外，正规渠道（比如门
121161	0.56	现在很多媒体都是转载的，真实可靠的新闻来源主要还是新华网和人民网以及部分其它媒体。互联网时代，信息更新很快所以很多新闻都来不及查证，一般都是只要符合网站宗旨的都发布，但又担心发布后新闻里的真实性、言论
50520	0.55	中国的各大小媒体，无论是直播的节目，还是转播、编辑后的节目，电视台都是知道内容的，具有可控性，也就是说经过了各地方宣传部门的审核才能播，即使现场直播的文艺表演也是经过n次彩排的，而马英九的讲话媒体完全
59454	0.55	有的是由于讲话内容偏长、对话内容不连贯等，不适合作为新闻全部播发，所以按上级指示进行了摘要报道；有些是由于讲话内容并未准备全部公开发表，所以按照上级指示摘要报道了其中可以公开报道的部分内容；当然，同一

FIGURE C.19: Regime: Sino-Japan Relation



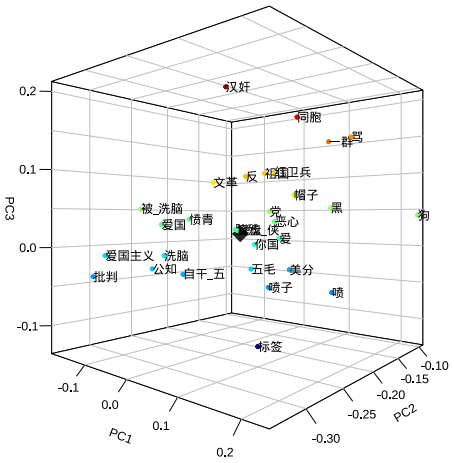
ID	θ	text
612698	0.59	额，不明白为什么要以父子关系来形容发动侵略战争的日本人和现在的日本人。如果说发动战争的是当时的日本人，所以有罪的是当时的日本人，但是对于我们而言的是，这是日本对中国的侵略战争，是国与国的关系，是日本发
631472	0.57	法德和解,是因为德国赔款、道歉付出的太多太多。中日之间与法德正相反，中国为了与日本和好，一、没有让日本赔款，二、没有让日本道歉。使得日本领导人杨凤武威　每年几次参拜 有杀害中国人民的甲级战犯的靖国神
638873	0.55	日本，日本国，日本人，日本人民，这几个概念在这个问题上是什么敏感的，日本反省了吗，答案是模糊的，因为日本等于天皇制，天皇还在，那个只是终战诏书，而不是投降诏书。日本国反省了吗？当然反省了，日本国宪法就
686152	0.54	题主问的是：日本有没有为侵略中国或他国道歉过？注意，是有没有道歉，不是有没有诚意，也不是首相去不去参拜靖国神社，更不是日本人有多恶心多虚伪多残暴。谁能代表日本国民进行道歉？是日本国的国家元首，日本天皇

FIGURE C.21: Regime: The Communist Party



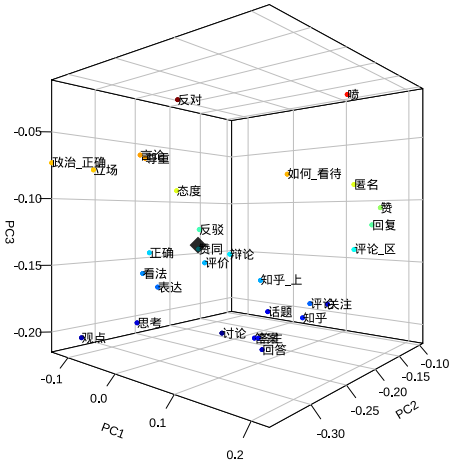
ID	θ	text
244574	0.86	1.披着毛左皮的邓右；2.陈主席在理论建设、组织建设和攀科技树上开的金手指太大；3.革命那么多年，一次大规模肃反都没有，没有左倾、右倾、反右、反反右的反复党内斗争，这点是最大的败笔。
614001	0.73	因为执政党是中国共产党。《中华人民共和国宪法》第一章 总纲 第一条摘录：中华人民共和国是工人阶级领导的、以工农联盟为基础的人民民主专政的社会主义国家。《中国共产党章程》总纲摘录：中国共产党是中国工人阶
181795	0.70	中国共产党要始终代表中国先进生产力的发展要求中国共产党要始终代表中国先进文化的前进方向中国共产党要始终代表中国最广大人民的根本利益
244901	0.68	因为分裂。60年代初，中苏矛盾尖锐化，日本左翼逐渐分裂为中立派、亲华派和亲苏派。1966 年，日共与中共断绝关系，诬中国为“帝国主义”。这引起日共产内部严重分裂，大批党员脱党，成立许多亲中共党派。19

FIGURE C.23: Debate: Name-calling



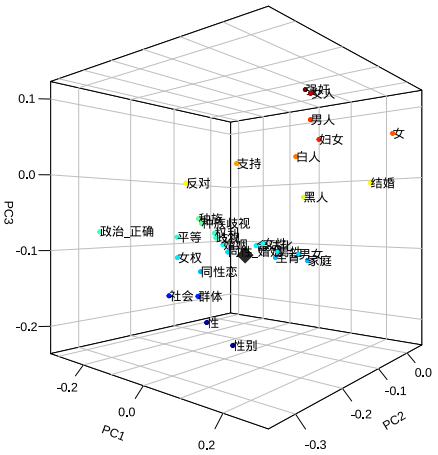
ID	θ	text
237042	0.76	精神外国人不太恰当，真正的精神外国人的特征应该是处处替外国人找想，而不是天天黑中国，这才叫精外，例如，武装保卫苏联。至于逢G必反的不算精外，中G不等于中国，不爱D不代表不爱国。此前有小粉红还污蔑本人替
629155	0.73	为什么要黑那些向往自由民主的人呢？我们谁喜欢被束缚着被独裁着？我们每个人都是向往自由民主的。我们黑的是那些打着"自由民主"旗号的傻逼，因为他们的"自由民主"只不过是另一种独裁而已。"爱国主义"是流氓最
339572	0.61	跑个题吧：在中国网络上，按照美分、五毛分划阵营，谁在指责谁不爱国？五毛指责美分不爱国？no，恰恰相反，是美分在指责五毛不爱国。极端的情况下，五毛会骂美分为“汉奸、卖国贼”，这是指责对方“卖国”，而不是
680931	0.59	扣帽子的概念大家自行百度。总之就是把人贴上让人反感的标签然后引起大家的反感。但是1。自干五通常是自封的，五毛才是通常用来扣帽子的词，自干五还以此为荣，所以不是扣帽子2。自封的自干五不一定是真自干五。两

FIGURE C.25: Debate: Taking positions



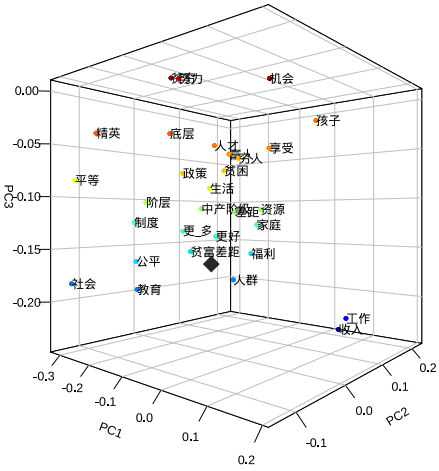
ID	θ	text
563798	0.55	目前王先生说的很对了，之所以开个答案，也就是为了和@孟庆斌撕逼；评论完就拉黑，鸵鸟战术用的不错，他说的很对，我确实经常在生活中感到困惑不解，没办法，蠢货太多，我又不能强行拉低自己的智商站在他的“高度”
159116	0.55	那般人，很明显不会做最基本的利弊权衡分析，不知自己所需、所求，不明自己所往所终，分不清敌友亲疏……==========那位高票的匿名答主的评论区已经关闭
557056	0.53	首先，我是彝族。然后，关于如何看待贴吧里这些言论，不知道楼主又想得到什么样的回答？我把所有答案都翻了一遍，觉得也确实都不够理智。有个答主说本来应该自称倮倮之类的，这些是楼主想看到的么？每个贴吧每个帖子
613848	0.53	我本来以为不友善是说脏话或者辱骂或者人身攻击，结果我在某位答主的答案下评论，那位答主莫名其妙的开始辱骂我人身攻击，我提醒她不要进行人身攻击，他变本加厉，我只好举报。结果管理员告诉我她的答案被删除了，但

FIGURE C.27: Society: Women and Minority



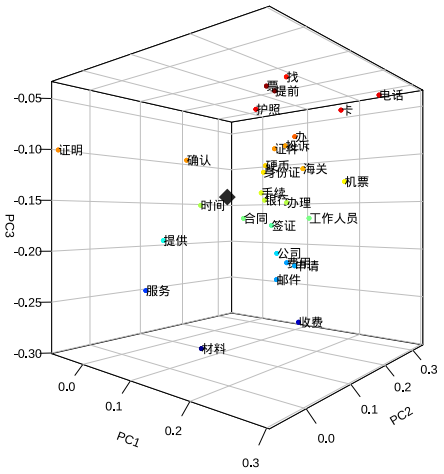
ID	θ	text
104075	0.67	不同的女权会有不同的看法。有女权认为是歧视；有女权认为隐含了歧视；有女权认为即便不是歧视也导致了歧视；有女权认为是过度保护，还算不上歧视；有女权认为不是歧视，是保护；有女权认为不是歧视，是补偿；有女权
60708	0.64	新中国的婚姻法，规定了女性有离婚的权利呢。女性也越来越知道权益被侵犯可以离婚。这是一种进步。95%的离婚是由女性提出，也说明，在这些婚姻中女性处于很不利的位置，处于被侵犯利益的位置，想必还有很多女性觉
182997	0.63	女权主义是男女权利、义务的平等，只强调权利闭口不提义务，这是哪门子女权主义，举个例子，女权们一方面反对物化女权，反对中国父权传统，但另一方面，在结婚的时候，就开始强调男方必须有房有车，婚前房子必须加女
183021	0.61	不支持同性恋，不歧视同性恋，接受民事结合，反对同性婚姻。我就不明白同性恋为什么要政府和社会把他们结合的方式纳入到“婚姻”的含义当中。叫民事结合不好么？一个黑人过去受歧视，现在不受歧视了，就一定要说“我

FIGURE C.29: Society: Inequality



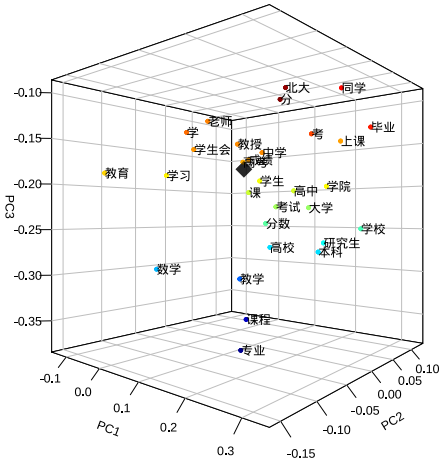
ID	θ	text
165765	0.60	先说一个现实：这个社会很不公平；再说一个找骂的观点：当前的社会，远好过完全公平和平均的社会。介于原题干中，公平等同于平均，以下也按照公平等于平均来论述，谢绝咬文嚼字。这篇答案将围绕以下几个问题，展开论
701337	0.58	人与人之间最终的差距，在于起点（家庭背景等）是否公平以及过程（竞争规则等）是否公平。显然，在起点、过程两个方面都存在着极大的不公平。因此，一个相对正义的社会旨在追求过程公平，以尽量弥补起点不公平所带来
19360	0.57	首先，社会经济的蛋糕如果不做大，能容纳的上层人有限。。自然不可能所有人都在上层。人和人之间自然会分出三六九等。。而且因为各种主观的（起点不同加优势者设置障碍）客观的（天赋差异）因素，这种差异往往不会消
333325	0.55	先谈结论，提供社会福利对于社会主义（中国特色）还是资本主义来说，在社会效果、经济效果、现实意义、未来发展上都是很划算的。前提是这个福利压力不会把自身的社会经济压垮。首先，资本社会的行为是有经济意味的。

FIGURE C.31: Society: Service



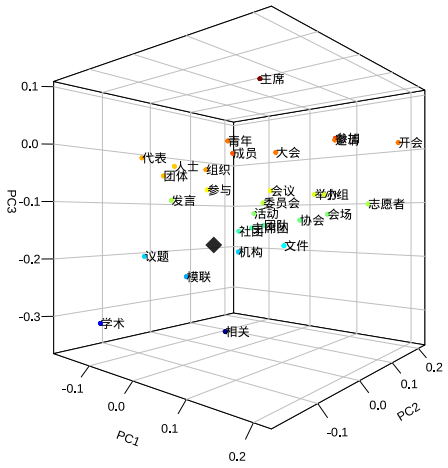
ID	θ	text
258660	0.92	没太明白是哪种签证，Q2多次签证每次停留180天吗？如果预计停留超过180天可以提前去当地出入境管理部门或者派出所询问办理延期停留签证，一次出境后就没用了；如果不办延期直接出境边防查到了逾期10天内会
27862	0.90	谢邀。上面的姑娘说了一些，我再补充一下吧。（刚开店其实不用办那么多复杂的证件，你可以等着他们上门催的时候再来办也不迟，当然前提是你要遇到好说话的工作人员）第一是：消防　这是最难过的一关　　商铺的消防
75046	0.90	1、营业执照①申请个体工商户名称预先登记应提交的文件、证件； 办理资料： （1）申请人的身份证明或由申请人委托的有关证明； （2）个体工商户名称预先登记申请书； （3）法规、规章和政策规定应提交的其他
568267	0.86	工商税控变更发行操作流程是怎么样的呢，下面大家就和轩辕邑会计（400---080--1589）小编一起来看一下吧。　　第一步：签到、取号、领取资料　　第二步：报税　　第三步：审核资料及纸质空白发票缴销

FIGURE C.33: Student: School



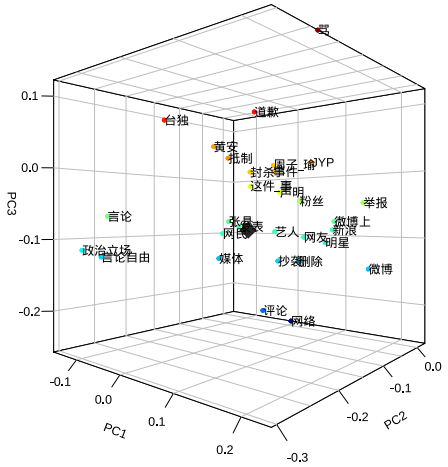
ID	θ	text
258285	0.76	初一，这是一个打基础的时候，很重要，初一学好了，初中没问题了，大家快报补习班。初二，这是一个过渡的时候，很重要，初二学好了，初三就稳了，大家快报补习班。初三，这是一个冲刺的时候，很重要，以前多少同学，
545511	0.73	首先，请楼主简单了解一下AP课程，AP全称为“高级课程班(AdvancedPlacement)”，是由美国大学委员会(CEEB)在美国高中设立的一个教育项目。旨在为有能力的高中生提供机会，允许他们在高
293706	0.73	AP是Advanced Placement的缩写，中文一般翻译为美国大学先修课程、美国大学预修课程。指由美国大学理事会（The College Board）提供的在高中授课的大学课程。美国高中生可以选
626442	0.72	AP课程（大学先修课程）是给一些学有余力的高中生，在10-11年级可以开始选修的，且难度较高的高中课程。有的美国学校是10/11年级，可以通过考试提前选修，有的学校是GPA高就可以选修，有的学校是任何

FIGURE C.35: Student: Module United Nations



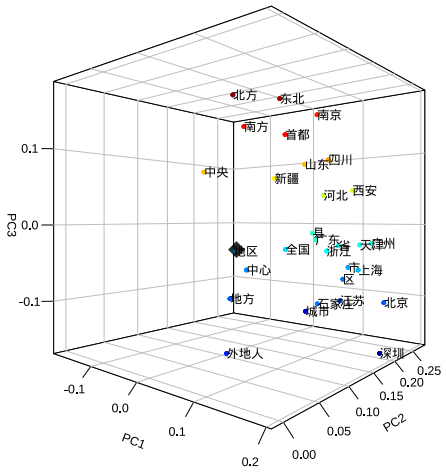
ID	θ	text
98148	0.84	这个问题倒是挺好的，授人以鱼不如授人以渔，与其列举近期的议题，我来说说一般情况下我是如何找议题的好了。step1：确定你希望的是一个常委，还是一个特委（历史委、内阁委、联动委……）step2：以下以联
103601	0.81	三个方面，题材，规则，玩法题材，委员会不再局限于联合国下设机构，议题也不再局限于现行时空。规则，基于通用规则进行创新，或另起炉灶使用新的规则。玩法，出现了一个委员会的学术进程成为另一个委员会的学术背景
40824	0.75	我是握梦模联的创始人、筹备委员会主席吴泽文，我看到咱们的回答中有人对我们的会场选择方式、缴费问题有些误解，我的回答主要解释这个问题，也作为题主对握梦了解的一个方式吧。第一，握梦从来没有向“ 刘一燊 ”所描
99489	0.73	推荐，同高中党，参加握梦会议两次，组委与主席团一直坚持“ 自主会议 ”的理念维护到每位代表的参会体验，会场设计多元有趣，并且会有多个预选方案，会根据代表志愿情况确立最终会场，并且对于某刘姓先生的回复，OU

FIGURE C.37: Specificity: News (An incidence about Taiwan)



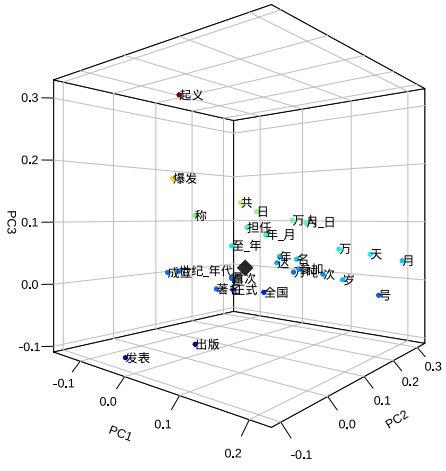
ID	θ	text
38941	0.73	如何看待 Facebook 上大陆网民用表情包回复台独网民？ - 王伊森的回答 - 知乎 如何看待 Facebook 上大陆网民用表情包回复台独网民？ - 王伊森的回答
79064	0.71	谢邀。大陆网友扫脸书，不是为了侵略性的输出文化，而是在受到挑衅之后的回应。这件事情的起因在于，苹果日报等渣媒借'周子瑜事件'恶意挑衅大陆网民，，这才引起了脸书大战。为什么大陆网友没有在扫脸书的同时扫t
208741	0.71	百分百的大陆人都认为台湾是中国不可分割的一部分，凭什么只要我们去尊重他们的政治立场，而他们就可以随意践踏我们的立场？说实话一开始根本没人care她举旗子的事，我作为twice的粉丝当初还维护她，可是台
26540	0.67	公知们洗地说，这是因为大量举报所以被封号。如果说大量举报就会被删除，那么我请大量律师起诉某公知，是不是就可以判处该公知死刑了？公知们不是FB管理员，也不是FB管理员肚子里的蛔虫，是怎么知道被封杀的反港

FIGURE C.39: Specificity: Place



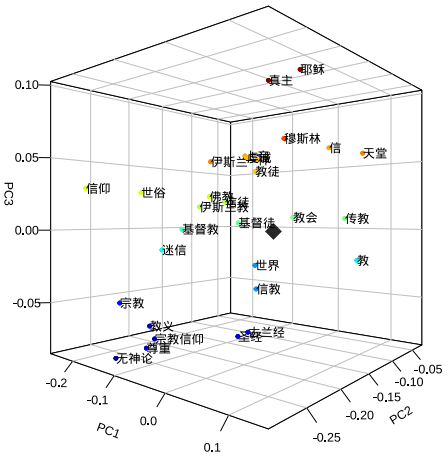
ID	θ	text
297537	0.93	当然是北京。首先，河北省由数个差异较大的亚文化区组成，宋代的河北路以及之前这一地区的一级政区基本是目前河北省位于华北平原的部分，张北，冀东等地区并入这一区划是后来北京建都以后的事情，这些城市在自然地理
297535	0.89	【中国地名系列1同音地名】1.yuncheng：运城市（山西西南角地市）、郓城县（山东西南）、云城区（广东云浮市辖区）2.tongguan：潼关县（陕西关中东部）、同官县（今铜川旧称，陕西关中北缘）、
692335	0.82	注：1. 因数据庞大，故统计的为现用名，未统计古用名；2. 数据限于大陆的一级行政区（省、自治区、直辖市、特别行政区），二级行政区（地区、盟、自治州、地级市）及三级行政区（县、自治县、旗、自治旗、县级
659697	0.82	很简单，因为广州的地铁里程没有北京上海多，但实际每日混迹在大广州的人口，可能比上海还要多，只略低于号称4000万人的北京。北上广深里面，广州的外来人口是最多的（而且绝大部分都没登记），外籍人口也应该是

FIGURE C.41: Specificity: Time



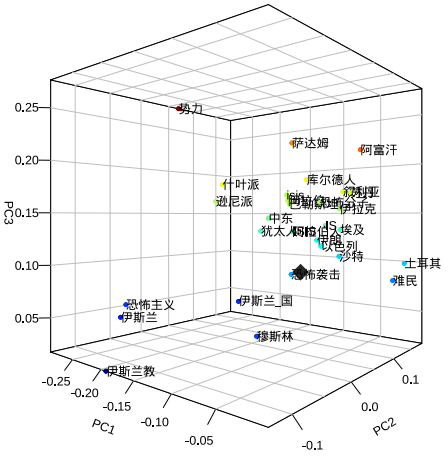
ID	θ	text
107224	1.00	中華民國台灣地區黨派實錄及成立時間:資料截止至2011年5月，源自中華民國內政部 「黨派關係請參考各個黨派的名字」编号 政党名称 成立日期 1 中国国民党 1894年11月24日2 中国青年党 1923
329823	0.99	国家税：1. 关税2. 盐税3. 统税 即货物出厂税1930年，国民政府决定裁撤厘税，年收入损失8000万元，再加上裁撤的常关税，复进口税等税务，每年损失1亿元1931年，开办棉纱、水泥、火柴统税。1
334752	0.95	1991年9月15日，C-17原型机首飞。1992年5月18日，C-17第一架生产型首飞。1993年2月5日，C-17正式服役。——1971年3月25日，IL-76原型机首飞。1975年，服役。——1
660216	0.91	提主，好心劝慰一句：自己动手，丰衣足食。以下为提主准备的资料： （努力修正编辑中....）秦公元前221（统一后始）——207共15年，历三帝，嬴姓，建都：咸阳。 1、始皇帝政 前221—前

FIGURE C.43: Religion: General



ID	θ	text
692059	0.89	逻辑是这样的：现世是短暂的，后世是永恒的，后世你要么进入火狱永受痛苦或者是奶与蜜的乐园，而这一审判完全由全知全能的真主来定度，审判的条件就写在真主降世给最后一个先知的古兰经中，按古兰经的方式生活以获取
207031	0.82	所谓”不要试探你的上帝“，其实是给上帝一个台阶下，别让上帝尴尬。以基督徒的祷告为例：基督徒在祷告中怕“试探上帝”，于是为在祷告结束时补上，“然而，不要成就我的意思，只要成就你的意思。”【路22:42
309134	0.81	中国是无神论最顽固的堡垒！可是基督教却攻破了这个堡垒雨后春笋般的在中国发展起来了！共产党人最相信实践是检验真理的标准。实践已经证实马克思主义的从来就没有什么救世主论断是错误的。基督教传播的是-----
127648	0.80	基督徒来传教，你就说你是天主教徒；天主教徒来传教，你就说你是佛教徒；佛教徒来传教，你就说你是伊斯兰教徒.....

FIGURE C.45: Religion: Muslim



ID	θ	text
281435	0.81	跟随沙特与伊朗断交的都是逊尼派穆斯林国家。逊尼派和什叶派是世仇，伊朗伊斯兰革命之后，逊尼派各国一直担心伊朗搞伊斯兰革命输出。萨达姆死后伊拉克已经变成什叶派的天下了，茉莉花之后埃及的国力也一落千丈，而今
271995	0.73	我简单来说一下，主要有四点 1.资金：一些穆斯林富豪，尤其以海湾国家为甚，他们会向圣战组织捐助资金，ISIS代替了基地组织的恐怖大亨地位，所以现在来自穆斯林富豪的主流资金捐助不再流向基地组织，而是流向
591927	0.72	逊尼派和什叶派从教义分歧上敌对，ISIS是逊尼派，但因为手段过于极端，被逊尼派默许（严格来说伊斯兰教就是如此极端，但是明面上还是要文明些）。犹太人和逊尼派和什叶派,从宗教差异上敌对，和ISIS不仅是宗
319704	0.69	背后是两大超级大国的对峙，沙特打的不是也门，是也门的反政府武装。也门反政府武装夺取了首都，并占领了总统府，同时控制了南部的空军基地。反政府武装气焰越烧越旺。周三一架战机轰炸了也门总统所在的官邸，所幸被

D How Sentiment of Answers Explains Anonymity

TABLE D.1: SENTIMENT ANALYSIS

	DV: Anonymous Answer
Anger	−0.046
	(0.057)
Sadness	−0.063
	(0.044)
Fear	−0.075
	(0.059)
Antipathy	0.049**
	(0.023)
Boredom	0.025
	(0.044)
Shame	0.449***
	(0.145)
Guilt	0.387***
	(0.109)
Worry	−0.059
	(0.126)
Disappointment	0.016
	(0.064)
Jealousy	−0.073
	(0.199)
Suspicion	0.095***
	(0.031)
Blame	0.083***
	(0.014)
Happiness	0.069**
	(0.027)
Like	0.042
	(0.032)
Surprise	0.058
	(0.084)
Respect	−0.244***
	(0.043)
Peace of mind‘	−0.041
	(0.058)
Trust	−0.104***
	(0.030)
Praise	−0.137***
	(0.017)
Wish	0.039
	(0.053)
Constant	−1.538***
	(0.005)
Observations	508,202
Log Likelihood	−236,023.600
Akaike Inf. Crit.	472,089.100

Note: Heteroskedastic standard errors clustered at prefecture level are reported in parentheses. * $p < 0.05$, ** $p < 0.01$