Internet censorship as a social process: Do governments censor content pertaining to specific topics and issues of interest?

Pranathi Iyer

MACSS, University of Chicago

Prompt 1

Research question

This project is an extension of my first project which looks at censorship on the Internet by government institutions as a social process.

Given that governments intervene in the process of regulating content online, are there certain pertinent topics or events of interest that they censor? Do institutions suppress discussions and conversations around certain issues? That is to say that are there certain broader level sensitive topics that emerge from the kinds of content that are reported by the government?

This is the research question that I aim to address in my project using data collected from accounts that were reported specifically by the government of India. The URLs of some such accounts between 2019 and 2021 is the archive that I had created for my first project, and will be employed for research in this project.

Prompt 2

Introduction

Censorship as a concept is not alien for countries across the world. While contexts, avenues, and channels of censorship might vary, with the advent of internet and social media, these avenues have only expanded. Social media platforms allow for conversation and discourse online, and consequentially are increasingly becoming tools used by nations to curtail free

speech and communication on the internet (Levin, n.d.). In a recent study, censorship was found to increase in 103 countries, including relatively democratic countries such as Norway, Japan, and India (Sundara Raman et al., 2020). This widespread prevalence of the internet and its censorship emphasizes the need to study online censorship, its mechanisms, and ramifications.

Having said this, operationalization and measurement of censorship in itself can be an arduous task. Data once redacted, more often than not, cannot be retrieved, and this makes it extremely challenging to analyse and understand the nuances of the process of censorship. However, owing to its increasing relevance, several researchers in the recent past have adopted various methods to study censorship across the world. Two of the most elaborate and large scale projects on censorship were carried out by Gary King and his fellow researchers. They were able to use supervised machine learning (King et al., 2013) and build on this by employing reverse engineering (King et al., 2014) to decipher the mechanism of Chinese censorship. A few recent studies have expanded such large scale approaches to the global level. The Censored Planet at the University of Michigan was able to use remote censorship measurement techniques to collect observations for 20 months and find increased censorship activity in 100 countries (Sundara Raman et al., 2020). While these are only a few examples, the common thread amongst these studies is the complexity of the method used. This speaks for the difficulty of studying censorship, but also sheds light on the potency that these approaches could bring to a topic of research as relevant and meaningful as online censorship.

However, these methods might vary depending on the country under consideration. Historically, different countries have practiced different forms and extents of censorship (Subramanian, n.d.), and these continue to change with time. As mentioned earlier, internet censorship in India has witnessed an increase in the recent past, despite it being the largest democracy in the world. Interestingly, India was among the top five countries to make removal

requests to twitter between 2012 and 2020 (Twitter, 2020). Despite this, not many studies have employed sophisticated methods to study censorship in the Indian context. A few have utilized tools such as the OONI (*OONI: Open Observatory of Network Interference | OONI*, n.d.) to study mechanisms of web censorship in India (Kumar Yadav et al., 2018). Yet, there remains immense potential to exploit abilities of computational social science to study complex processes such as censorship. This project is a stepping stone towards the larger attempt to understand mechanisms of internet censorship in India, and breakdown pertinent aspects that the government of the largest democracy in the world finds problematic in the larger realm of discourse online. Akin to other research studies on censorship, research of this kind can be instrumental in understanding intentions of the state, and extent of freedom of speech of its citizens.

Prompt 3

Data

I use two forms of data. The first is an archive of URLs of reported accounts that I had created as part of my first project. These URLs were collected from Lumen database at Harvard University, and were used to access accounts against whom action was taken by the government. Since some of these accounts were suspended, and content of a few were withheld, in order to maintain consistency, I further collect only replies to tweets that were reported by the government. While some of the original tweets were not withheld by twitter, taking them into account into the sample, along with replies to other withheld posts would create inconsistency in representation. Around 200 accounts were checked, and owing to heavy censorship, replies to 51 tweets were gathered. An average of 15 replies per tweet was collected. It is important to note here that the data collected was textual, which implies all images and videos were not taken into consideration.

Why is the data valid for the project?

Studying censorship can be expensive and disparate. The data selected is a direct part of content censored by the Indian government on a large platform like twitter. Moreover, the details of requests are collected from Lumen Database, a large organization that keeps track of data removed from the internet.

As I mentioned earlier, censorship is an extremely difficult construct to operationalize, and this data is only one method of doing so. There could be several others — perhaps better ones— that can help study censorship in the Indian context.

Prompt 4

Method used

My objective of this research is ultimately to identify if there are certain pertinent issues of interest against which the government of India takes action. To do this, I employed topic modelling to understand if certain topics emerge across replies of tweets, with replies to one tweet being one document, over a course of two years. While topic modelling in general seems suitable to understand clusters of domains within a certain text, it factors out very differently on social media data, twitter in this case. I elaborate further on this when I discuss potential limitations of this project.

Topic modelling can help understand the underlying thematic structure and context of a large amount of data over a certain period of time. In the context of internet censorship, it could help understand what topics of data are censored, how the nature of this content varies over time, and how it is affected by sensitive events occurring in countries at specific times.

However, utilizing topic models to understand social media data is not a new idea. A study that was discussed in class, used topic modelling on Facebook data (Bonacchi et al., 2018) to

understand how ideas from early Iron Age to Early Medieval Britain influenced discussions around Brexit. While I use a small sample of twitter data which is shorter than text one might find on Facebook, I believe the ideas, methodologies, and intent for research employed in the research that studies Brexit are broadly applicable to topic modelling studies based on data gathered from social media platforms.

The step wise methodology that I followed is given below.

Step 1	Collect documents of tweet replies using Octoparse
Step 2	Synthesise all tweets into a csv file
Step 3	Pre-process and tokenize data
Step 4	Convert tokens into gensim dictionary and create bag of words
Step 5	Compute coherence values for LDA models with different topics
Step 6	Create visuals for the top words within topics of the topic with the highest coherence score
Step 7	Create pyLDAvis visuals to understand interaction between topics

Prompt 5

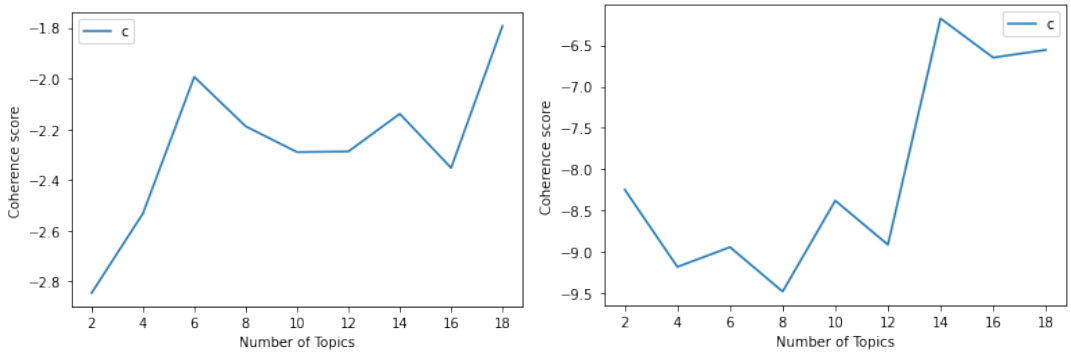
Findings

Text analysis using social media data and can provide several insightful insights.

At the pre-processing stage, I had to very astutely look at the data via multiple iterations to understand what set of stop words might be required above the existing nltk collection. This took several iterations, since tweets and their replies use various kinds of abbreviations and references which must carefully be included or excluded from the dataset. For instance, initially

'pm' seemed trivial but after inspection, I realized that it was actually referring to the prime minister of the country, and did not include it in the stop words. Moreover, the fact that people in India post in different languages, made the pre-processing task even harder since I had to eliminate words that were not in english for this model. However, some words might still remain.

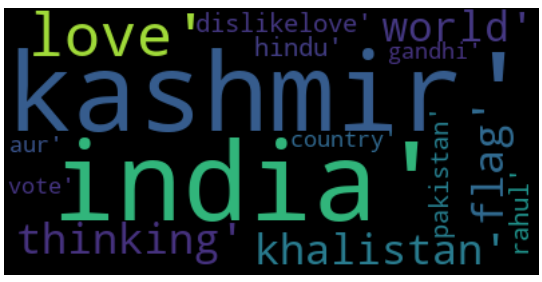
This step proved to be extremely integral in my analysis. At first, without removing custom stop words, and words from other languages, the coherence score of the model was extremely high (left in the figure below). However, it was meaningless since most words did not offer any meaning. After cleaning, the coherence score dropped but the model was more meaningful.

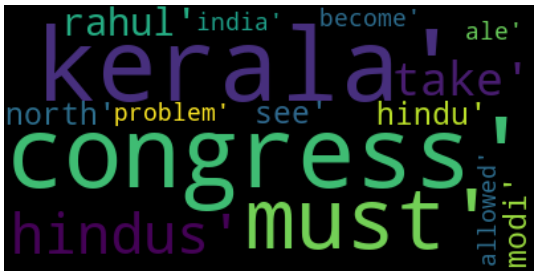


Before pre-processing

After pre-processing

Moving on to the coherence score, topic number 14 seemed to have the highest coherence, and upon creating word clouds for top words within this, these are the visuals I was able to develop.

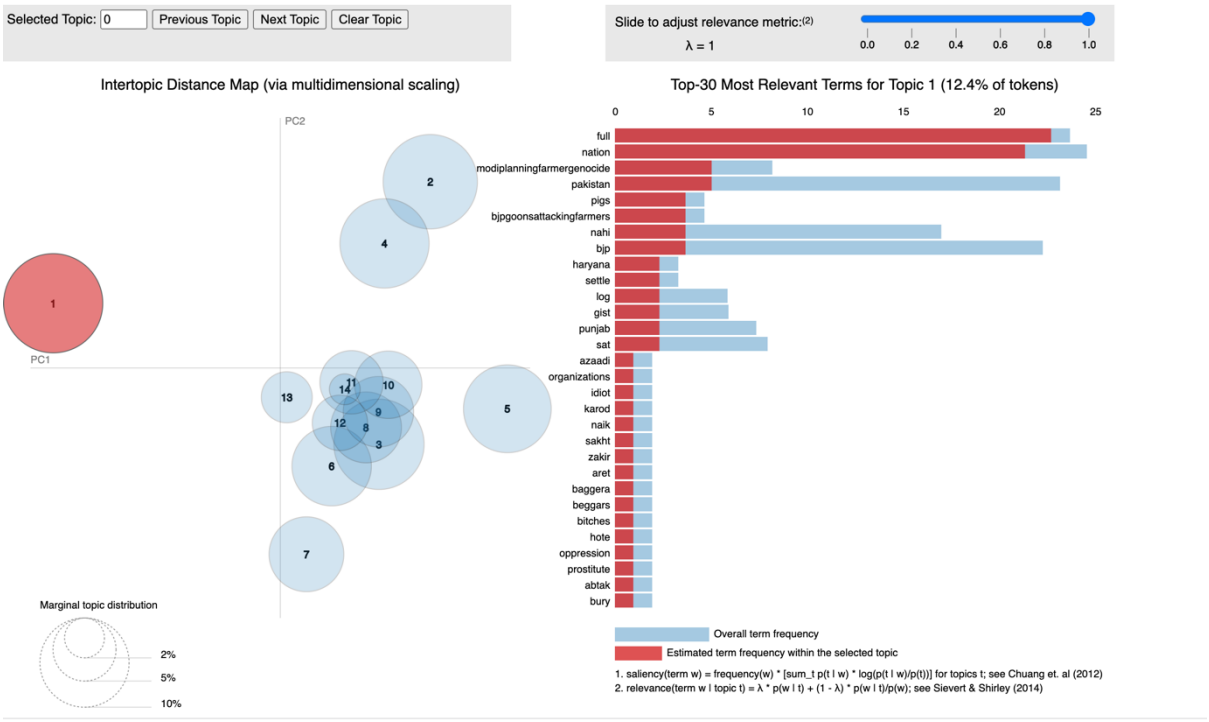




I only discuss 4 of these visuals here to avoid making the section lengthy and verbose. Figure 1 mentions ‘Kashmir’, ‘Khalistan’, ‘India’ amongst other words. Kashmir is an Indian state that shares its border with Pakistan and has been a sensitive issue since India was partitioned. ‘Khalistan’ is a separate state that a certain religious community in India has been demanding for a while. Similarly, ‘Sangh’ in figure 2 refers to people belonging to a parent faction of the BJP, which is the party currently in power in India. Figure 3 mentions ‘Pakistan’ and ‘modiplanningfarmergenocide’ amongst others. The latter has been a matter of great significance over the past few months with farmers protesting against the privatization of agricultural markets and removal of minimum support price. Lastly, figure 4 mentions ‘Congress’ and ‘Kerala’, which are the major opposition party and key state where BJP is not the party in power, respectively.

Of course, this is a very preliminary analysis and claims regarding the context in which these words were used cannot be made without in depth analysis. However, even with such a small data size, the model was able to present useful results. This only reiterates the potency it could offer to a larger and more representative dataset.

I finally use pyLDAvis to create visuals to look at the interaction amongst topics.



The figure above depicts topic 1, which has top words such as ‘nation’, ‘modiplanningfarmergenocide’, ‘Pakistan’, ‘pigs’, and ‘bjpgoonsattackingfarmers’. These are all sensitive matters that could be closely related to certain events related to Pakistan and farmers that took place in India in the beginning of 2021. Topics 2 and 4 have an intersection because they both speak of Kashmir, perhaps one in the context of Muslims, and the other in the context of Khalistan (mentioned above) having s special state status like Kashmir. However, more analysis might be needed to make a more confident claim. Lastly, all the topics clustered in the middle are related because they all have the words ‘India’, certain common words of the English language, and use words such as ‘BJP’, ’country’, ’sad’ in different contexts. Moreover, a lot of words such as ‘shameless’, ‘save’, ‘country’, are used in several matters such as women’s safety, Khalistan, and use of ‘evms’ during elections. Hence, naturally, several of the topics are not distinct in their word use, but differ in their contexts.

This research is probably as rudimentary as it can get to understand topics of censorship on social media. However, more advanced research can be built on this, which can further contribute towards creating an ecosystem of understanding censorship online, which might only increase as use and prevalence of social media platforms increase.

Limitations of the project

This project comes with certain limitations that one must be cognizant of in order to meaningfully interpret the results.

- Studying censorship is challenging, and despite having a Twitter API, it was extremely difficult to extract replies to withheld posts using it, since the API does not allow for smooth extraction of replies for withheld posts.
- I was able to extract replies to reported tweets using Octoparse, however, after checking over 200 tweets, I was able to extract only replies to 51 tweets since most of the accounts or posts were suspended by twitter. This means that in order to use topic modelling meaningfully for a problem like this, it would require to go through an extremely large number of reported accounts to have a decent sample size of replies. I tried to extract all URLs of reported accounts between May 2019 and October 2021 from the Lumen database, constrained by their throttling.
- Owing to the diverse languages in India, some of the tweets collected were not in English, and this made data cleaning and interpretation harder.
- The data does not account for any form of non-textual data.
- The above restrictions made the data sample to be relatively small for using topic modelling meaningfully, and hence might not be representative of the results that might emerge from a larger, and more meaningful dataset.

Conclusion

This project is not the end and only a means to understanding a potential method of deciphering internet censorship by the government of India. Despite the extremely small sample, the results conveyed are powerful because each of the topics speak of issues that have been extremely sensitive issues at some point in time, or continue to do so. With enough processing power, and more efficient means, extracting replies and using topic models can prove to be an interesting an innovative way to study censorship and augment the existing literature on mechanisms to operationalize censorship.

References

Bonacchi, C., Altaweel, M., & Krzyzanska, M. (2018). The heritage of Brexit: Roles of the past in the construction of political identities through social media: *Https://Doi.Org/10.1177/1469605318759713*, *18*(2), 174–192. https://doi.org/10.1177/1469605318759713

King, G., Pan, J., & Roberts, M. E. (2013). How censorship in China allows government criticism but silences collective expression. *American Political Science Review*, *107*(2), 326–343. https://doi.org/10.1017/S0003055413000014

King, G., Pan, J., & Roberts, M. E. (2014). Reverse-engineering censorship in China: Randomized experimentation and participant observation. *Science*, *345*(6199). https://doi.org/10.1126/science.1251722

Kumar Yadav, T., Sinha, A., Gosain, D., Kumar Sharma, P., & Chakravarty, S. (2018). *Where The Light Gets In: Analyzing Web Censorship Mechanisms in India*. https://doi.org/10.1145/3278532.3278555

Levin, D. (n.d.). *Challenges in Measuring and Evading Nation-state Censors*.

OONI: Open Observatory of Network Interference | OONI. (n.d.). Retrieved November 11, 2021, from https://ooni.org/

Subramanian, R. (n.d.). The Growth of Global Internet Censorship and Circumvention: A Survey. *Communications of the IIMA*, *11*(2). Retrieved November 6, 2021, from https://scholarworks.lib.csusb.edu/ciimaAvailableat:https://scholarworks.lib.csusb.edu/ciima/vol11/iss2/6

Sundara Raman, R., Shenoy, P., Kohls, K., & Ensafi, R. (2020). Censored Planet: An Internet-wide, Longitudinal Censorship Observatory. *Proceedings of the ACM Conference on Computer and Communications Security*, 49–66. https://doi.org/10.1145/3372297.3417883

Twitter. (2020). Removal requests. Twitter. Retrieved October 18, 2021, from https://transparency.twitter.com/en/reports/removal-requests.html#2020-jul-dec