Supplemental Information:
Pick Your Language: How Riot Reporting Differs Between English and Hindi Newspapers in India

The supplemental information contains details about the data collection process for both the *Times of India* (TOI) and *Hindustan* (SI.1), a brief discussion on selecting the optimal number of topics for topic modeling (SI.2), and a discussion on best practices for using topic models to compare two different sources (SI.3).

SI.1: Detailed Description of Data Collection

I collected 10,244 articles mentioning the word 'riot' from the TOI (7,365) and *Hindustan* (2,859) from May 27, 2009 to April 29, 2018. A number of issues prevented collection of a longer time period or for different newspapers. Initially, I conducted a comprehensive search of the top circulating newspapers in India including all languages. I was primarily interested in comparing Hindi and English newspapers because some of the hypotheses I present are specific to these two languages, but I was also open to collecting data from other newspapers. Apart from the TOI, no other Indian newspapers were available on subscription archives in a comprehensive format. LexisNexis translated some articles from vernacular newspapers, but not the entire set. Acquiring as close as possible to the full sample of riot related stories was very important; this meant that web scraping was necessary. Most vernacular newspapers did not have an online archive or one that was easily searchable. This excluded some of the largest Hindi newspapers like *Dainik Jagran*, *Dainik Bhaskar*, *Rajasthan Patrika*, and *Amar Ujala*. Though other English language newspapers had better archives, the TOI has by far the largest circulation.

I collected the TOI articles by hand using Factiva. I searched over all TOI editions during this period for the term 'riot,' excluding identical duplicates. In Factiva terms, this means that instances of the identical story being run in different editions/locations should be excluded, though this mechanism was not perfect. A story with the same text except for one or two lines changed would be included as a 'new' story in the sample. However, more restrictive search methods on Factiva exclude articles similar, but not the same as those already presented. The story text from the TOI does not include the author's name or the location of the story. This byline information was rarely printed in Indian newspapers and when included, Factiva placed it separately from the main article text.

Hindustan articles were scraped from https://livehindustan.com/archive using Python and Selenium. Replication code is available on the author's website. Hindi text was scraped from the website and fed into Google Translate using the Google API. Aiken and Balan (2011) find that translation between Hindi and English on Google Translate is quite effective. Because Python has difficulty dealing with non-English characters, only the English translation was stored. *Hindustan* archives go back only to 2009, and before May 27, all articles were given the same date. I searched for all articles mentioning the Hindi word for 'riot.' Unfortunately, no other Hindi newspapers had archives that could be searched or scraped over a long period of time, so this comparison only covers the TOI and *Hindustan*. The end date of April 29, 2018 represents the day that data collection started for this project.

One initial idea was to collect historical newspaper archives in order to match riot data with fine grained, district level data on riot reports. Only a few Hindi newspapers are available in microfilm for periods before 2010. I scanned microfilm from *Hindustan* and translated them

using Tessaract, but the image quality of the microfilm itself meant that Hindi characters were almost unreadable, no matter the quality of the scan.

I estimated the total number of articles per newspaper per year in order to determine the percentage of all stories discussing riots in each newspaper. This corrected for the length of the newspaper and any differences in the number of stories being posted on the *Hindustan* website compared to on the Factiva page of the TOI. For the TOI, I searched for articles containing the word 'the,' the most common word in English. I then estimated the percentage of identical duplicate articles using the percentage of identical duplicate articles from the 'riot' search. For *Hindustan*, I searched for articles containing the word 'of,' the most common word in Hindi. I estimated the number of articles based on the number of pages of search results by year on https://livehindustan.com/archive.

SI.2: Selecting the Number of Topics

The topic modelling approach is introduced in the main text. Topic models do have problems (Grimmer and Stewart, 2013). I address the major issue, selecting the optimal number of topics, below.

There is no agreed upon method to select the optimal number of topics, but several groups of scholars have proposed solutions. I chose three of them: Arun et al. (2010), Cao et al. (2009), and Deveaud et al. (2014). Each group of scholars propose a statistic that can be used to measure how well a certain number of topics fully explains the variety of topics in the dataset. The number of topics that best explains the dataset is known as the optimal number of topics. For example, these statistics suggest that classifying all of the articles in the dataset into one topic loses a lot of important variation in the themes discussed in the articles. Similarly, having twenty

or more topics starts breaking themes into such specific categories that few articles fit into each category.

I computed the statistics for each of these three measures for between two and twenty topics. The Deveaud et al. (2014) measure suggested that fourteen topics are optimal. The Arun et al. (2010) and Cao et al. (2009) measures suggested between fourteen and twenty topics. I ran the topic model for twenty topics and found it difficult to differentiate some topics from others. For this reason, I used the lower bound of the Arun et al. (2010) and Cao et al. (2009) measures and selected the Deveaud et al. (2014) optimal number of fourteen topics.

SI.3: Combined Versus Separate Topic Models

Although topic models have been used in many applications to determine underlying topics in one corpus, comparing two corpora using topic models is quite rare. A number of computer scientists including Crossno et al. (2011) and Zhao et al. (2011) have developed sophisticated methods to make this comparison, but these methods have only been used in their specific computer science applications. I identified two simple approaches to comparing topic models across corpora. One was to run two separate topic models, one for each corpus. This technique is appropriate if we believe that the topics contained in the corpora are extremely different because no topic from one corpus is forced to appear in the other corpus. However, a major problem is that running two different topic models means that topics are not directly comparable. It is possible to try to match topics to each other based on the words most associated with the two sets of topics, but this method introduces a good deal of human intervention and analysis. Further, the two sets of topics are never truly the same even if the defining words for the two sets of topics match.

A second approach is to merge the two corpora and find a common set of topics. Once the topic model is complete, we can split the topic proportions out by newspaper source and compare. This method ensures that the topics in both corpora are directly comparable, but it means that the two corpora must be focused on similar events or else a common topic in one corpus may be forced to appear in the other corpus. I implemented the second method in the main text because the newspaper articles I examined were all related to riots. This method would not be appropriate if I was comparing riot articles with all the other articles appearing in the two newspapers because the newspapers would cover different types of content.

We still might be concerned that running two separate topic models may uncover dramatically different results. I ran individual topic models on each newspaper corpus, and I found that at most two of the fourteen topics differed from the original list of topics and differed across newspapers. This suggested that there was a common set of topics across both newspapers and that the most appropriate strategy was to combine corpora and to run a single topic model.

Finally, some might be concerned with the somewhat ad-hoc manner of assessing which topics in the topic model are substantive. I mostly relied on the maximum difference between the two newspapers in a given year. Another way of telling the difference between topics is to calculate the mutual information of the TOI and *Hindustan* for each topic. Mutual information is defined as the Kullback-Leibler divergence between the probability mass functions of the two newspapers. A substantive interpretation of this value is the number of bits you will need to store the topic proportions for one newspaper given that you already know the topic proportions for the other newspaper. As such, higher values mean that there is less similarity between the two newspapers for a given topic. Table SI.3.1 shows the topics in the main text, adding the correlation and mutual information. Riots and Senses, Police Control, and Political Parties have

much higher mutual information, meaning they were much more dissimilar compared to other topics. The mutual information for Official Statements was, however, low. This was most likely due to the high correlation between the two newspapers; information storage is low because one newspaper is just a vertical shift from the other. This highlights a difficulty with using the Kullback-Leibler divergence to measure differences in topic proportions. Because vertical shifts use little information, a large vertical shift could go undetected by measuring mutual information. In the main text, it made more sense to interpret the maximum difference instead of discussing the efficiency of information storage.

Table SI.3.1: Similarity Between TOI and Hindustan Topics

Topic	Correlation	Mutual Information
Description of Riot Events	0.905	0.0031
Official Statements	0.884	0.0016
Police Arrests	0.861	0.0008
Court Statements	0.861	0.0040
Riots as a Metaphor	0.748	0.0009
Political Parties	0.744	0.0097
Women & Children	0.700	0.0011
World Without Riots	0.534	0.0038
Government Report	0.281	0.0021
Sikhs	0.253	0.0057
Police Control	0.246	0.0111
Riots in Film/Music	0.028	0.0038
Communal Issues	-0.389	0.0051
Riots and Senses	-0.810	0.0137

Topic names are interpreted from the most influential words in each topic. The Correlation column is a Pearson correlation. Mutual Information is a measure of similarity between topics in the TOI and topics in *Hindustan*. Higher values mean less similarity, lower values mean more similarity.

References

Aiken, M., and Balan, S. (2011). An Analysis of Google Translate Accuracy. *Translation Journal* 16(2).

Arun, R., Suresh, V., Veni Madhavan, C., and Narasimha Murty, M. (2010). On Finding the Natural Number of Topics with Latent Dirichlet Allocation: Some Observations. In Zaki, M.J., Yu, J.X., Ravindran, B., and Pudi, V. (Eds.), *Advances in Knowledge Discovery and Data Mining* (pp. 391–402). Berlin: Springer.

Cao, J., Xia, T., Li, J., Zhang, Y., and Tang, S. (2009). A density-based method for adaptive LDA model selection. *Neurocomputing,* 72(7-9): 1775–1781.

Crossno, P.J., Wilson, A.T., Shead, T.M., and Dunlavy, D.M. (2011). Topicview: Visually comparing topic models of text collections. *Proceedings of the ICCC International Conference on Artificial Intelligence*. IEEE.

Deveaud, R., SanJuan, E., and Bellot, P. (2014). Accurate and effective latent concept modeling for ad hoc information retrieval. *Document Numerique,* 17(1): 61–84.

Grimmer, J., and Stewart, B.M. (2013). Text as Data: The Promise and Pitfalls of Automatic Content Analysis Methods for Political Texts. *Political Analysis,* 21(3): 267–297.

Zhao, W.X., Jiang, J., Weng, J., He, J., Lim, E.P., Yan, H., and Li, X. (2011). Comparing Twitter and Traditional Media Using Topic Models. In Clough, P., Foley, C., Jones, G.J.F., Kraaij, W., Lee, H., and Murdoch, V. (Eds.), *Advances in Information Retrieval* (pp. 338–349). Berlin: Springer.