Comparative Discourse Analysis Using Topic Models: Contrasting Perspectives on China from Reddit

Zachary Kimo Stine
Department of Information Science, University of Arkansas at Little Rock, AR, United States
zkstine@ualr.edu

Nitin Agarwal
Department of Information Science, University of Arkansas at Little Rock, AR, United States
nxagarwal@ualr.edu

ABSTRACT

In this study, we conduct a comparative analysis of the linguistic features that differentiate two China-focused discussion communities with contrasting perspectives from Reddit. We utilize probabilistic topic modeling to represent submissions from both communities as distributions of latent patterns of word-usage. Using information theoretic measures, we conduct a series of quantitative comparisons between the language patterns of each community and identify salient features that distinguish the two communities relative to each other. We describe the rhetorical techniques and discursive frames implied by these features and how they are utilized by each community in discussions surrounding the Hong Kong protests during 2019. Additionally, we contribute a novel method for representing collections of documents that preserves interdependencies between topics at the document level.

CCS CONCEPTS

• **Human-centered computing** → Collaborative and social computing.

KEYWORDS

Comparative discourse analysis, topic models, computational social science, information theory

ACM Reference Format:
Zachary Kimo Stine and Nitin Agarwal. 2020. Comparative Discourse Analysis Using Topic Models: Contrasting Perspectives on China from Reddit. In *International Conference on Social Media and Society (SMSociety '20), July 22–24, 2020, Toronto, ON, Canada.* ACM, New York, NY, USA, 12 pages. https://doi.org/10.1145/3400806.3400816

1 INTRODUCTION

The notion that a given discourse constitutes a particular way of understanding some aspect of the world, whose meaning is contingent and socially constructed, is fundamental to prevailing theories and methods of discourse analysis [15]. Viewed this way, competing discourses surrounding the same entity can be understood as competing constructions of that entity. In this study, we analyze two competing English-language discourses surrounding China. Understanding the range of popular Western perspectives on China is critical given China's importance on the global stage and the tensions that sometimes exist between China and Western countries such as the United States.

The competing discourses we analyze are produced by two communities from the discussion platform Reddit: *r/China* and *r/Sino*.[1] While specific discussions in either community may reflect a range of perspectives on China, the general discursive constructions of China produced by each community are fundamentally at odds with each other, especially in their views of the Communist Party of China, or CPC.[2] Whereas *r/China* generally tends to be highly critical of the CPC, *r/Sino* tends to defend the CPC against criticism and engages in much more positive discourse around the party. Understanding these conflicting discursive constructions of China is important to understand the perspectives and representations of China that English-language Reddit users may encounter. Given Reddit's popularity within the broader social media ecology, it is also reasonable to think that these discourses have some degree of salience for understanding popular perspectives on China within the English-speaking world more broadly.

While English-language discussions about China may occur within a variety of Reddit communities, called subreddits, we limit our analysis to *r/China* and *r/Sino* for three reasons. First, *r/China* and *r/Sino* have an ideal discursive scope that is neither overly broad nor narrow when compared with many other subreddits that may include discussions about China. For example, the subreddit *r/worldnews* includes discussions relevant to China, but the breadth of its discursive scope extends far beyond China. In order to isolate the discourses pertinent to China, it would be necessary to sample only the appropriate submissions and to ensure that the particularities of this sampling process did not introduce undesirable biases into the analysis.

Second, *r/China* and *r/Sino* represent concerted discourses about China that do not occur within some other primary context. For example, the subreddit *r/taiwan* may include discussions about China, but those discussions are likely to be too contextually specific, focusing only on aspects of China that are relevant within a Taiwan-specific context. Since *r/China* and *r/Sino* are directly focused on China, the users active in these subreddits are primarily there to discuss China, independent of some other primary context.

SMSociety '20, July 22–24, 2020, Toronto, ON, Canada

ACM ISBN 978-1-4503-7688-4/20/07…$15.00
https://doi.org/10.1145/3400806.3400816

[1] We follow the convention of using the prefix "r/" to denote the names of Reddit communities, or subreddits. Additionally, we write subreddit names as they are stylized by the subreddit itself, hence some subreddit names are not capitalized even when they refer to proper nouns.
[2] The Communist Party of China is also commonly referred to in English as the Chinese Communist Party, abbreviated as CCP. We follow the party's own English-language convention of using the Communist Party of China and the abbreviation CPC.

Third, *r/China* and *r/Sino* produce their respective constructions of China with documented awareness of the opposition between them. Notably, the discourse of *r/Sino* can be read as a reaction to the discourse of *r/China*. One of the earliest submissions in *r/Sino* serves as a welcome and explanation of the subreddit by one of the subreddit's moderators. In the submission text, the author explains that other China-related subreddits exist, but that they are "hateful" and "spread misinformation." The submission text itself does not explicitly mention *r/China*, but *r/China* is explicitly referenced 18 times within the submission's comments. References to each other occur throughout both subreddits, further validating the view that the two discourses are in competition with each other. *r/China* is the older subreddit, created in 2008 compared to 2015 for *r/Sino*. It also has wider exposure on Reddit with almost four times as many subscribers as *r/Sino* (at the time of this writing). Therefore, we can view the construction of China provided by *r/Sino* as an intended corrective to the alleged inaccuracies of the representation provided by the more dominant *r/China*.

Our goal in this study is to not only characterize the two competing discourses, but to characterize what makes each stand out most from the other. We accomplish this by first identifying the discursive features underlying the discussions of the two communities by training topic models on the discussion text, which represent each discussion thread as a mixture of latent word-usage patterns, or topics [3]. Second, we map each discussion thread to a categorical feature representation that includes individual topics or combinations of topics. To our knowledge, the process for mapping documents to categorical combinations of topics is a novel methodological contribution of this study. We then calculate how frequent these features are within each community to obtain community-specific feature distributions. Using an information theoretic quantity, we then calculate how conspicuous each feature is in one community when juxtaposed with the other. Finally, we provide qualitative interpretations of the discursive features that emerge as salient and what discursive frames and strategies they indicate.

Our theoretical framework for interpreting results is informed by the computational approach we take. Our use of topic models to learn the primary discursive features that we analyze constitutes a "distant reading" of the discourses that trades fine-grained, nuanced interpretations for access to large-scale patterns that would not be otherwise observable. In other words, we are interested in the broad tendencies of the discourses, which necessarily obscures more specific aspects of the discourses that might also be illuminating. While we do conduct manual readings of documents from each subreddit, we do so in a way that is guided and constrained by the topic models we obtain. The theoretical framework within which we compare the two discourses is motivated by our desire to make the comparisons objective (at least, as much as possible). Therefore, rather than qualitatively compare quantitative features, we rely on information theory to quantitatively interrogate the relationships between the two discourses and to quantify how salient each feature is for distinguishing between each discourse. This approach assumes that what is most interesting about these competing discourses is what most juxtaposes them.

We apply this methodology within two cases. The first consists of a general comparison between all discussions from the two subreddits over a roughly four-year period during which both subreddits were active. In the second case, we focus specifically on submissions from the two subreddits that are from 2019 and discuss Hong Kong. While the first case provides an overarching comparison of how the two communities conceive of China, the second allows us to see how these two ways of understanding China lead to different ways of understanding the protests occurring in Hong Kong throughout much of 2019. In analyzing discussions about the protests, the two discourses are brought into sharp contrast, providing a glimpse of how real-world events are constructed and interpreted differently by *r/China* and *r/Sino*.

We find that the most frequently observed features are similar across the two communities, but that *r/Sino* can be best distinguished from *r/China* by negative discourse around other countries—most often the United States—while more general stylistic features and discussions about living and working in China most differentiate *r/China*. In the case of the 2019 Hong Kong protests, we find that discussions about the politics underlying the protests are salient in *r/China*, while discussions about violence in the protests alongside criticism of the United States are salient in *r/Sino*. We also find that our method for constructing community-level feature distributions with combinations of topics allows us to look beyond isolated topics and better appreciate important interactions between multiple topics within discussions.

2 BACKGROUND

In order to better understand our analyses, we provide some background and review selections of relevant work covering Reddit scholarship, topic modeling, and the use of information theoretic measures for comparative purposes.

2.1 Reddit

Data collected from Reddit have been analyzed within a variety of research contexts. Network approaches have been used to characterize common user roles in subreddits [5] and user loyalty to subreddits [14]. Prior work on Reddit has also focused on specific behaviors of users including the usage of hate speech [6] and norm violations [7].

Reddit is an especially useful source for data given that each community has a well-defined focus. For example, the subreddit *r/ChangeMyView* has been analyzed in order to better understand persuasion [25] and to characterize user susceptibility [17]. Additionally, birth narratives from the subreddit *r/BabyBumps* have been computationally analyzed to better understand the experiences of people who have given birth [1].

A contribution of the present study to Reddit scholarship is the comparative approach taken to understand the relationship between the language used in two subreddits.

2.2 Latent Dirichlet allocation

In order to identify the word-usage patterns underlying discussions from *r/China* and *r/Sino*, we train probabilistic topic models via latent Dirichlet allocation (LDA) [3]. LDA results in two kinds of distributions being inferred: some number of distributions over the vocabulary present in the corpus and a mixture of those distributions that best represent each document. The distributions over vocabulary are referred to as topics, though this is not always the

most appropriate way to think about what they represent. Topics comprise a pattern of word-usage and may correspond to certain rhetorical styles as well as actual topics. Additionally, topics learned though LDA arguably reflect certain concepts from the sociology of culture, including framing, polysemy, heteroglossia, and a relational approach to meaning [9].

LDA requires the number of learned topics, k, to be specified. Because it is unlikely that a uniquely "true" number of topics exists underlying any non-trivial corpus, the selection of k is often based on more pragmatic grounds, primarily how useful the resulting topics are for the researchers making sense of the corpus. While several quantitative methods exist for evaluating topic models [8, 18, 29], qualitative evaluation is necessary [23]. Different selections of k may result in slightly different though potentially equally plausible sets of topics with k influencing the specificity of the topics [20].

Topic models have been used in a variety of contexts including comparative philosophy [21], literary scholarship [11], cultural evolution [2], and in comparing Twitter data from different sources [19]. While Nichols et al. [21] and Morstatter et al. [19] both use topic models within a comparative context, their approaches differ from ours in key ways. In the case of Morstatter et al. [19], two document collections are compared by training two separate topic models for each collection and calculating the similarity between matched topics from each model. The two models reflect two distinct feature spaces. Their approach discovers whether a feature from one model has an analogous feature in the other model and, if so, how similar the two features are. Thus, their interest is in finding the extent to which two separate data sets produce similar features as a proxy for understanding how well a sample of data represents a much larger data set. While this is an appropriate method for the questions being asked in that study, having two distinct sets of features for the two subreddits we are comparing would complicate our ability to calculate how distinguishing a feature is in either of the subreddits. In other words, we are not interested in comparing the overall similarity of two distinct feature spaces, but rather the characteristics in a common set of features that are strong signals of one subreddit relative to the other.

The comparative approach taken by Nichols et al. [21] uses a single topic model to compare documents within a shared feature space, which the present study more closely mirrors. However, in that study, the authors compare three philosophical works by treating the ten highest probability topics within each of the three texts as sets. The texts are then compared based on the topics within the intersections of each set. By comparing the sets of each text's top ten topics, useful information from the probabilities of these topics within a text are mostly discarded (outside of determining which topics should be included in a text's topic set). Information theory provides a set of tools for making more rigorous comparisons between probability distributions, which we use as the basis of our quantitative comparisons.

Prior work exists which attempts to use LDA within the context of discourse analysis. LDA was combined with the theoretical framework of critical discourse analysis in order to examine how Muslims and Islam are discursively constructed within Swedish social media [27] as well as the discursive relationship between Islamophobia and anti-feminism [28]. In both studies, the topics produced by LDA were found to be useful in representing the discursive context under analysis. Brooks and McEnery [4] provide a less favorable view of LDA within discourse analysis, criticizing it on the grounds that it lacks linguistic theoretical grounding, and that the topics produced by LDA from their data were difficult to interpret from lists of high-probability words and often lacked thematic coherence across documents. In the present study, we did not find similar problems with our topic models after a combined analysis of the high-probability words for each topic along with manual readings of exemplar documents for each topic.

2.3 Information theory and measures of divergence

Because the features of interest from topic models are in the form of probability distributions, they lend themselves to the use of information theoretic measures for rigorously interrogating relationships between objects within the inferred topic space. In this study, we are specifically interested in the relationship between distributions of topics among two collections of documents, each representing a Reddit community. We follow the usage of the partial Kullback-Leibler divergence by Klingenstein et al. [16] who use the measure to identify features that were most salient for distinguishing between violent and non-violent trials in England over time. Here, we are interested in how well a given feature acts as a signal of one community over the other.

Other relevant uses of divergence measures include comparisons of hashtag usage between protestors and counterprotestors on Twitter [10] and comparisons of proceedings from natural language processing conferences [13]. Notably, Hall et al. [13] also use LDA to represent the documents being compared, but the method used for creating collection-level topic distributions amounts to calculating an average topic distribution to represent each collection. While this is a reasonable approach, it results in the loss of document-level topic interactions, which we preserve in this study.

3 METHODS

In the following sections we describe the data collected and the methods used to analyze them. After data collection, we trained topic models on the combined data from the two subreddits. We then constructed feature representations for the two document collections in two ways: first, by counting the dominant topic for each document within a collection and second, by counting combinations of topics for each document within a collection, using a threshold value to determine which topics to combine. Information theoretic measures of divergence were then used to identify the most distinguishing topics or combinations of topics between the two communities' collection-level topic distributions. While these methods are used in the context of Reddit discussions, they are likely to be useful in any context in which collections of text are compared and where the size of these data sets is too large to feasibly make sense of them through manual reading alone.

3.1 Data

For each community, we collected all submission identifiers from the community's date of creation up to December of 2019 using

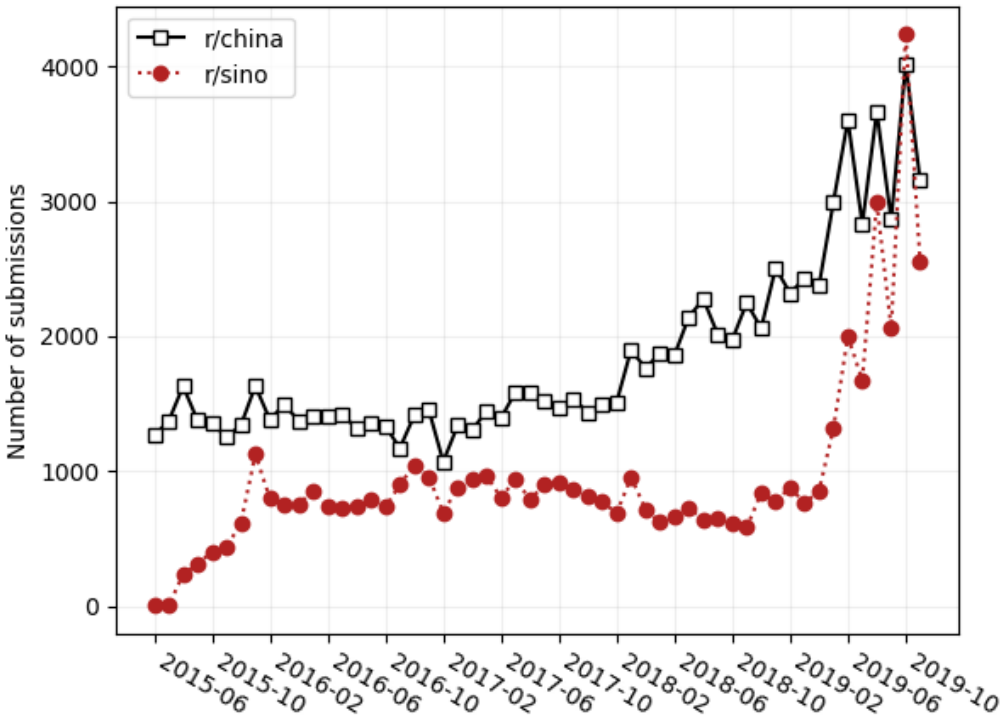

Figure 1: Monthly submission frequency of *r/China* and *r/Sino* from June 2015 through November 2019. Month labels are formatted as YYYY-MM.

the service PushShift.io. We then used Reddit's application programming interface (API) to collect the text of each submission along with all comments from the submission's discussion thread. Submissions are available for *r/China* going back to January of 2008, while submissions are only available for *r/Sino* as early as June 2015, with only five available for *r/Sino* in its first two months of existence (Figure 1). Given the more recent creation of *r/Sino*, submissions considered when training topic models for either community were posted no earlier than August 2015. We consider a document to be the text of a submission and the comments from its discussion thread.

From these submissions, we performed basic preprocessing. After tokenizing, only tokens consisting of at least three characters were kept. Common words that occur in over 25% of all submissions were removed. Rare words that occur in fewer than five submissions were also removed. While both *r/China* and *r/Sino* are predominantly English-language communities, Chinese characters (*hanzi*) are sometimes used. All Chinese characters were identified based on their Unicode values and removed. We did not stem tokens, as this has been shown to have minimal or even negative effects on topic modeling [24]. Prior to preprocessing, there were 261,555 unique word types, which were reduced to 65,176 word types after processing. Documents were discarded if they contained fewer than 20 post-processing tokens, resulting in 97,619 total documents (down from 147,681 documents).

All of the data we collected are public and did not require IRB approval. Both *r/China* and *r/Sino* are publicly accessible and do not require a Reddit account to access. While these data are considered public, we avoid linking to specific submissions, direct quoting, and mentioning any user names in order to avoid bringing any unwanted attention to the individuals whose comments we analyze.

3.2 Collection-level topic distributions

In order to get a range of topic specificity, we trained LDA models with 30, 90, and 150 topics, which we refer to throughout the paper as models A, B, and C, respectively. With these models, each document can be represented as a distribution over 30, 90, or 150 topics, where each topic is a distribution over the vocabulary of 65,176 word types. When referring to a topic, we include the model name to distinguish between two different topics that happen to share the same topic number (e.g., A.10 and B.10 are two different topic features from models A and B respectively). We used the Gensim Python package for LDA model training [22].

We considered multiple methods for constructing topic distributions that reflect a collection of documents. While an LDA model provides topic distributions for each document, we would like to construct a topic distribution that reflects all documents within a collection. Existing methods for combining document-level topic distributions into a collection-level topic distribution include calculating an average topic distribution from the document-level topic distributions for all documents in a collection, such as in [13]. Another possible method would be to assign each word in a document to a topic based on the document's topic distribution and then count these word-topic assignments within each document

in the collection to make a collection-level topic distribution, such as in [26]. This method also incorporates the length of each document in the collection-level topic distribution—longer documents will have greater influence over the resulting collection-level topic distribution.

Both of these methods for constructing collection-level topic distributions result in the loss of information about potential interdependencies between topics that are salient within the same document and thus provide important context. For example, a document that contains language primarily about both the United States and the Hong Kong protests is better represented by a combination of these two topics that is lost if we consider each of the two topics in isolation. Additionally, a feature may be prevalent in both collections, but may be combined with other features differently in the two collections.

The loss of potential topic relationships in the methods just discussed results from both methods being ways of calculating the frequency of each topic within a collection, whether by counting topic proportions in each document and then normalizing or by counting word-topic assignments and then normalizing. In order to capture topic interdependencies within documents, we propose expanding the feature space of topics to also include combinations of topics. By mapping each document's topic distribution to a single categorical feature consisting of either an individual topic or a combination of topics, we can construct collection-level topic distributions that preserve topic relationships from the document level within the broader collection-level distribution. Below, we describe two kinds of collection-level topic distributions constructed from mapping each document's topic distribution to a categorical representation: dominant topic distributions and topic tuple distributions.

3.2.1 Dominant topic distributions. As a baseline, we first calculated collection-level distributions by assigning each document to the topic with the highest probability in the document's topic distribution. After assigning each document to a topic in this way, a community's collection-level topic distribution can be formed from the relative frequencies of these document-topic assignments. We refer to collection-level topic distributions created in this way as dominant topic distributions.

Dominant topic distributions can be thought of as a special case of the topic tuple distributions described below where the threshold value is zero. This method treats each document as equally important regardless of length, in contrast to the word-topic assignment method discussed above. However, relationships between topics are necessarily lost in dominant topic distributions, since documents will always be assigned to an individual topic. Therefore, we use dominant topic distributions as a baseline with which to compare the results found using topic tuple distributions in order to see what, if anything, is gained from the combining topics.

3.2.2 Topic tuple distributions. As we will see, interesting findings can be made from analyzing dominant topic distributions. However, dominant topic distributions necessarily obscure potentially interesting interdependencies between topics. In order to preserve potential interdependencies between multiple topics within a single document, we propose a method for mapping a document's topic distribution to an ordered tuple of topics where the number of elements in the tuple is based on a threshold parameter and is therefore flexible. To the best of our knowledge, this is a novel method for representing documents as categorical topic features based on their topic distributions as inferred through LDA.

To construct a topic tuple distribution, we first define a threshold parameter, t, within a range from 0 to 1. For each document belonging to a community, that document's topic tuple consists of the ordered topic indices that, when their corresponding proportions in the topic distribution are summed together, equal or exceed the specified threshold. A document's topic tuple must have at least one element and only the minimum number of elements necessary to meet the threshold condition.

As an example, consider the following topic probabilities for some document with four topics: 0.01, 0.49, 0.41, and 0.09 corresponding to the proportions of topics T.0, T.1, T.2, and T.3, respectively, and which sum to one. If we define the threshold to be 0, then the document's topic tuple only includes the dominant topic, T.1, which has the largest probability of 0.49. However, if we define the threshold to be 0.5, then topic T.1 is no longer sufficient to meet the threshold. Instead, the topic with the next highest proportion, T.2, must be combined with T.1 to form the topic tuple, (T.1, T.2). The summed probability of these two topics in the document is 0.9, which satisfies the threshold of 0.5.

This example illustrates the kind of interdependencies between topics that can be preserved using this method. Topics T.1 and T.2 are both similarly salient in the document (based on their similar proportions in the document's topic distribution), which is reflected by the topic tuple containing both. If instead, the proportion of T.1 was 0.85 and the proportion of T.2 0.05, then only T.1 would be needed to meet the threshold of 0.5. In this case, T.1 is uniquely salient within the topic distribution and so additional topics are not needed in the document's topic tuple to meet the threshold.

It should be noted that the same threshold will make higher demands when used on topic distributions with larger number of topics. For example, a threshold of 0.5 may result in a topic tuple of 2 elements if k is 90, but may result in a topic tuple of 3 elements for the same document when represented in a distribution where k is 150. This is simply due to the probability mass having to be spread out over a larger number of elements in the case of 150 topics versus 90 topics. Additionally, increasing the threshold value may result in a larger number of features that constitute the collection-level topic distribution (see Table 1). Arbitrarily increasing the number of features in this way may have undesirable effects by decreasing the ability to meaningfully discriminate between two topic tuple distributions.

We constructed topic tuple distributions using topics from each of the three topic models described above for threshold values of 0.1, 0.3, 0.5, and 0.7. We limited our qualitative analysis to threshold values of 0.3 and 0.5 in order to avoid the potential problems of having too many features.

3.3 Levels of analysis

We compared collection-level topic distributions representing *r/China* and *r/Sino* at two different levels of analysis. First, we compared distributions reflecting all documents used in training

Table 1: Number of features present in collection-level topic distributions.

Model	Dominant topic	Topic tuple (t=0.3)	Topic tuple (t=0.5)
A (k=30)	30	1,232	10,515
B (k=90)	89	6,786	32,231
C (k=150)	138	12,087	54,618

our topic models, from August 2015 through November 2019. Second, we narrowed our focus to documents from 2019 that contain language about Hong Kong as a more focused case study of the different perspectives on the Hong Kong protests. These documents were selected after identifying six topics relating to Hong Kong—one from model A (A.11), two from model B (B.30 and B.44), and three from model C (C.6, C.70, and C.118). Any document which has one of these six topics as its most dominant or second-most dominant topic and that has a submission date between January through November of 2019 was included in this set.

For both levels of analysis, we filtered out all documents with dominant topic A.18, as they correspond to submissions that are dominated by one of several boilerplate moderation comments, typically due to the submission not following the community rules. This filtering resulted in 5,114 documents being discarded in the subsequent analysis.

3.4 Feature comparisons

For each level of analysis, we calculated the Jensen-Shannon divergence between a given collection-level topic distribution representing *r/China* and a distribution over the same features representing *r/Sino*. In order to measure how strongly each feature of the distribution functions as a signal of a community, we calculated the partial Kullback-Leibler divergence, KL_i, for each element of the distribution, which reflects how much each feature individually contributes to the Jensen-Shannon divergence [16]. The Jensen-Shannon divergence can be formulated in the following manner as the symmetrized version of the asymmetric Kullback-Leibler divergence:

$$JSD(\mathbf{p},\ \mathbf{q}) = \frac{1}{2}\left[KLD(\mathbf{p},\ \mathbf{m}) + KLD(\mathbf{q},\ \mathbf{m})\right] \tag{1}$$

where $\mathbf{p}$ and $\mathbf{q}$ are the distributions being compared and $\mathbf{m} = {}^1/_2\ (\mathbf{p} + \mathbf{q})$. The Kullback-Leibler divergence with an expectation based on $\mathbf{p}$ is given by

$$KLD(\mathbf{p},\ \mathbf{m}) = \sum_i \mathbf{p}_i \log_2 \frac{\mathbf{p}_i}{\mathbf{m}_i} \tag{2}$$

from which the partial Kullback-Leibler divergence for the i^{th} feature in the distribution is simply

$$KL_i(\mathbf{p},\ \mathbf{m}) = \mathbf{p}_i \log_2 \frac{\mathbf{p}_i}{\mathbf{m}_i} \tag{3}$$

The partial Kullback-Leibler divergence measures how strongly feature i acts as a signal of the expectation distribution ($\mathbf{p}$ as written in equation 3) [16]. Thus, knowing the partial Kullback-Leibler divergence for each feature with an expectation based on a distribution representing *r/China* tells us how conspicuous that feature is for *r/China* against a background based on a distribution representing *r/Sino*. We calculated the partial KL for each feature using *r/China* as the expectation distribution to rank features in order of relative salience in *r/China* and then did the same with *r/Sino* as the expectation distribution.

For each comparison, we then examined the ten most frequent features in each community as well as the ten most distinguishing features of each community (based on the partial KL values). We conducted these comparisons for each combination of LDA model, feature type, and level of analysis described above. In order to understand the significance of each feature in context, we manually read multiple documents that possess the feature. This is necessary since interpreting topic features based only on some number of highly probable terms in a topic can be problematic.

4 RESULTS

In this section, we describe our findings from the broader level of analysis followed by findings from documents discussing Hong Kong during 2019. At each level of analysis, we first examine dominant topic features followed by topic tuple features. We report both highly frequent and highly distinguishing features in the tables below. Highly frequent features are reported with their relative frequency within a community's document collection (e.g., Table 2). While highly frequent features within a community provide a general characterization of that community's discourse, the features which are most distinguishing reflect which aspects of one discourse are comparatively salient in that discourse relative to the other. The most distinguishing features of each community's discourse are reported with their partial KL values, given in bits (e.g., Table 3).

4.1 Broad comparison of discourse

In order to get a broad sense of how the discursive constructions of China differ between the two communities, we first consider findings that arise when comparing documents from August 2015 through November 2019. In each of the three topic models, a topic emerges that is prevalent in both communities. These topics—A.24, B.31, and C.123—reflect a general rhetorical style that tends to be negative and critical, based on our manual reading of exemplar documents that feature these topics with high probability. These stylistic topics are the most frequently observed dominant topics in both communities at this level of analysis. Several other features exist within each of the topic models, so reporting them within each model would be redundant. We limit reporting results at this level of analysis to the features from model B (k=90), as they are interpretable, but not overly specific.

4.1.1 Results from dominant topic distributions. When examining the most frequent features in *r/China* from the dominant topic distribution, the features that appear most frequently (aside from the stylistics topic B.31) concern practical matters—asking questions and seeking advice (e.g., how to ship a package to China) (B.76), personal aspects of life in China as a foreign national (B.21), discussions about jobs and working in China (B.87), the use of VPNs for accessing websites (B.18), etc. In addition to these more practical topics, discussions about trade with the United States (B.75) occurs as the fifth most frequent feature.

Table 2: Most frequent dominant topic features.

r/China features	Proportion of collection	*r/Sino* features	Proportion of collection
B.31 Critical stylistics	0.612	B.31 Critical stylistics	0.458
B.76 Questions	0.064	B.24 International partnerships	0.072
B.21 Being a foreign national	0.041	B.75 Trade with the US	0.052
B.87 Jobs and working life	0.024	B.19 Military/engineering	0.045
B.75 Trade with the US	0.019	B.9 Political ideology	0.029
B.18 VPNs/website access	0.017	B.37 Technological growth	0.029
B.9 Political ideology	0.016	B.86 Economic standing	0.023

Table 3: Most distinguishing dominant topic features.

r/China features	Partial KL (bits)	*r/Sino* features	Partial KL (bits)
B.31 Critical stylistics	0.1186	B.24 International partnerships	0.0611
B.76 Questions	0.0583	B.19 Military/engineering	0.0418
B.21 Being a foreign national	0.0329	B.75 Trade with the US	0.0283
B.87 Jobs and working life	0.0182	B.37 Technological growth	0.0195
B.18 VPNs/website access	0.0159	B.25 Financial reporting	0.0133
B.72 Comm. Applications	0.0088	B.9 Political ideology	0.0112
B.33 Purchasing products	0.0088	B.86 Economic standing	0.0109
B.64 Water sanitation	0.0066	B.54 Scientific research	0.0107

After the stylistics topic B.31, the most frequent dominant topics observed in *r/Sino* include discussions around China establishing partnerships with other countries (B.24), trade with the United States (B.75), military and engineering innovation in China (B.19), political ideology and systems (B.9), technological innovation and growth (B.37), China's international economic standing (B.86), etc. An overview of these features for both communities can be seen in Table 2

While we can see some interesting differences between the most frequent dominant topic features of the two communities, calculating the partial KL values of each feature provides a rigorous way of ranking the comparative salience of each feature. For *r/China*, the ordering of distinguishing dominant topics resembles its most frequent dominant topics with a few changes. After topics B.31, B.76, B.21, and B.87, the next most distinguishing topic for *r/China* is B.18 (VPNs), followed by several less frequent topics reflecting discussions about needing help with communication applications (most often, the messaging application, WeChat) (B.72), purchasing products (B.33), and water sanitation (B.64).

The most distinguishing dominant topic features for *r/Sino* similarly reflect several of its highest-frequency features with some changes. Topic B.24 (international partnerships) is the most distinguishing topic, followed by B.19 (military and engineering innovation), B.75 (trade relations with the United States), B.37 (technological innovation and growth), B.25 (financial reporting), B.9 (political ideology), B.86 (economic standing), and B.54 (scientific research). See Table 3 for an overview of the most distinguishing dominant topics.

Notably, the discourses that emerge from *r/China* over this broad period of time tend to reflect the experiences of individuals—their experiences as foreign nationals living in China (B.21), their working lives (B.87), and practical concerns (e.g., B.18, B.72, and B.33). In juxtaposition to this, the discourses that emerge from *r/Sino* focus on China at the state level with respect to its relationships and relative standings with other countries (B.24, B.75, and B.86) as well as its own internal growth, development, and power (B.19, B.37, B.25, and B.54).

4.1.2 Results from topic tuple distributions. While the analysis of dominant topic distributions has yielded interesting results, the most frequent feature in either community is the somewhat vague stylistic topic B.31, which accounts for over 60 percent of the *r/China* collection and over 45 percent of the *r/Sino* collection, and which is the strongest signal of *r/China* relative to *r/Sino*. When we compare collection-level distributions of topic tuples, we find that several interesting features emerge in which B.31 is dominant but interdependent with an additional topic.

Using a threshold value of 0.3, we find that the most frequent topic tuples in *r/China* involve the same topics seen in Table 2, but now in more contextually informative combinations, including (B.31, B.21) reflecting discourse that features critical stylistic elements in combination with life as a foreign national in China, (B.21, B.31) representing the same combination but with life as a foreign national given primacy, and (B.31, B.76) combining critical stylistics with questions and advice. An interesting picture begins to emerge from these co-salient topic tuples that is borne out when reading the source documents with these features—while *r/China* may often invoke critical language that is untethered from more specific discursive foci, the emergence of (B.31, B.21) and (B.21, B.31) as relatively frequent reflects the tendency of *r/China* users

Table 4: Most frequent topic tuple features (t=0.3).

r/China features	Proportion of collection	*r/Sino* features	Proportion of collection
(B.31) Critical stylistics	0.341	(B.31) Critical stylistics	0.219
(B.31, B.21) Stylistics & foreigner	0.033	(B.24) International partnerships	0.032
(B.76) Questions	0.023	(B.75) Trade with the US	0.025
(B.21, B.31) Foreigner & stylistics	0.017	(B.31, B.75) Stylistics & US trade	0.025
(B.31, B.76) Stylistics & questions	0.017	(B.19) Military/engineering	0.024

Table 5: Most distinguishing topic tuple features (t=0.3).

r/China features	Partial KL (bits)	*r/Sino* features	Partial KL (bits)
(B.31) Critical stylistics	0.0968	(B.24) International partnerships	0.0298
(B.31, B.21) Stylistics & foreigner	0.0259	(B.19) Military/engineering	0.0235
(B.76) Questions	0.0218	(B.75) Trade with the US	0.0146
(B.21, B.31) Foreigner & stylistics	0.0141	(B.31, B.88) Stylistics & Western hypocrisy	0.0128
(B.31, B.76) Stylistics & questions	0.0132	(B.31, B.75) Stylistics & US trade	0.0111

to discuss their lives in China as foreign nationals in ways that are often negative.

Likewise, we see several of the same features occur with high frequency in *r/Sino* at a threshold of 0.3 which we saw as dominant topics, but now including the topic tuple (B.31, B.35) representing the combination of the critical stylistics feature with discussions about trade with the United States. Many of these documents include discussions that heavily criticize the United States in relation to the so-called trade war between the two countries that began in 2018. See Table 4 for an overview of the five most frequent topic tuples for this threshold value.

For *r/China*, the same five most frequent features are also the five most distinguishing features relative to *r/Sino*. However, a few interesting changes are present within the five most distinguishing features of *r/Sino* relative to *r/China*. The first three most distinguishing features correspond to those described for the dominant topic features from Table 3. Additionally, we see the topic tuple (B.31, B.88), corresponding to a combination of the critical stylistics topic with B.88, which represents discourse about the "West" (typically used to refer to the United States), most often as accusations of hypocrisy (e.g., perceived double standards regarding the state's treatment of Uyghurs in light of the United States' treatment of those seeking refugee status there) and more general charges of propaganda and anti-Chinese bias in Western media. This feature demonstrates the usefulness of examining topic tuples in this manner: When only analyzing dominant topics, B.88 is obscured by the prevalence of highly critical language that often accompanies discussions of the West by *r/Sino*. By allowing for the possibility that more than one topic is needed to adequately represent a document, we can see the different ways in which *r/Sino* uses the critical stylistics topic, whether in criticism of perceived Western hypocrisy or in criticism of the China-US trade war. See Table 5 for an overview of the most distinguishing topic tuples from each community at a threshold of 0.3.

From analyzing the feature distributions that characterize the collection of documents from *r/China* and that of *r/Sino* over a period of over four years, we see that both communities employ a generally similar way of using language that involves being highly critical (topic B.31). From an analysis of the dominant topic distributions of each community, we see that (aside from B.31), *r/China* submissions are often concerned with the experiences of individuals—most often as foreign nationals navigating their lives in China. When we look more deeply into the topic relationships that may occur within documents (by constructing distributions of topic tuples rather than dominant topics), we find that the critical stylistics topic frequently pairs with these other topics, providing greater context for understanding the discourses.

4.2 Comparison of discourse concerning Hong Kong in 2019

When we compare discourse surrounding Hong Kong during 2019, we again see that the critical stylistics topic B.31 is the most frequently occurring feature in each community, both when analyzing dominant topics and topic tuples with a threshold of 0.3.

4.2.1 Results from dominant topic distributions concerning Hong Kong in 2019. In looking at the top five most frequent dominant topics from the two communities, they share much in common in terms of the features' rankings (see Table 6). Notably, the topic B.35 may represent two kinds of language. On the one hand, B.35 appears in submissions which include a string of phrases intended to provoke censorship. These phrases typically include references to Tibet, the Tiananmen Square massacre of 1989, "democratization," "independence," and "freedom" among others. Among the *r/Sino* collection of documents concerning Hong Kong, the appearance of B.35 almost always indicates this usage within a submission title, which are flagged as violating of the subreddit rules (it is likely considered a form of trolling). However, the appearance of B.35 within *r/China* may also include language that shares some words

Table 6: Most frequent dominant topic features concerning Hong Kong in 2019.

r/China features	Proportion of collection	*r/Sino* features	Proportion of collection
B.31 Critical stylistics	0.649	B.31 Critical stylistics	0.677
B.44 Protest violence	0.109	B.44 Protest violence	0.142
B.30 Protest politics	0.102	B.30 Protest politics	0.052
B.35 Sensitive phrases	0.054	B.35 Sensitive phrases	0.041
B.75 Trade with the US	0.011	B.88 Western hypocrisy	0.012

Table 7: Most distinguishing dominant topics concerning Hong Kong in 2019.

r/China features	Partial KL (bits)	*r/Sino* features	Partial KL (bits)
B.30 Protest politics	0.0411	B.44 Protest violence	0.0257
B.35 Sensitive phrases	0.0106	B.31 Critical stylistics	0.0206
B.76 Questions	0.0054	B.65 Automatic summary bot	0.0098
B.83 CPC policy	0.0033	B.88 Western hypocrisy	0.0091
B.49 Subreddit rules	0.0023	B.45 Media criticism	0.0027

in common with the trolling usage just described. Submissions that feature actual discussions invoking Tiananmen Square, Tibet, or democracy may also have this dominant topic.

Examining which dominant topics most distinguish each community yields interesting differences (see Table 7). Topic B.30 is highly conspicuous in *r/China* and represents general discussions around the protests, typically framed as political tensions between mainland China and Hong Kong. The most distinguishing dominant topic of *r/Sino*, B.44, also reflects language about the Hong Kong protests, but more specifically concerns violence occurring during protests. Such submissions from *r/Sino* tend to focus on violence alleged to have been committed by the protestors (though in *r/China* this topic may reflect violence carried out against protestors by police in addition to violence committed by protestors). This marks an interesting change in the discursive strategies we previously described for *r/Sino*: While *r/Sino* broadly tends to focus on states as actors, rather than individuals, as described in section 4.1 above, in discourse around the Hong Kong protests, the community emphasizes the negative actions of individuals.

The other dominant topics that distinguish *r/China* at this level of analysis include the previously described sensitive phrases topic (B.30), question asking (B.76), language about reports or announcements from the CPC (B.83), and the enforcing of subreddit rules (B.49). The critical stylistics topic is more conspicuous within *r/Sino* at this level of analysis whereas, at the broad level of analysis described in section 4.1, this topic served as a stronger signal of *r/China*. While B.65 reflects submissions that include automatically constructed summaries by a self-declared bot account, we also see the appearance of B.88, denoting accusations of Western hypocrisy, and B.45, criticizing media outlets for reporting alleged falsehoods.

4.2.2 Results from topic tuple distributions concerning Hong Kong in 2019. When we analyze the topic tuples representing each community's collection of documents, we again see that the critical stylistics topic is often co-salient with other relevant features, which are obscured when only considering the dominant topic of each document. Here, those features include language about the protests in relation to their political underpinnings (B.30) and to protest-related violence (B.44). See Table 8 for an overview of the most frequent topic tuples.

Notably, language about protest-related violence does not occur within any of the five most distinguishing topic tuples for *r/China*. However, three of the five most distinguishing features for *r/Sino* feature language about violence, almost always as carried out by protestors. Instead, *r/China*'s distinguishing features concerning Hong Kong deal more with the underlying politics, both as reflected by B.30 and some of the discussions related to B.35 that sometimes invoke language about democracy. See Table 9 for an overview of the most distinguishing topic tuples for each community at a threshold of 0.3.

Interestingly, the topic tuple (B.31, B.88) appears as the third most distinguishing feature for *r/Sino*, despite these documents being required to have a Hong Kong-related topic as its first or second most probable topic. This is a case where using more specific topics can be helpful as these documents have a Hong Kong-related topic in the 150-topic model as the first or second most dominant topic that does not appear in the 90-topic model.

If we increase the threshold to 0.5, we do see that there is a connection between *r/Sino*'s usage of critical stylistics (B.31), charges of Western hypocrisy (B.88) and Hong Kong-related topics within the 90-topic model. See Table 10 for the distinguishing features of each community when examining topic tuples with a threshold of 0.5. At this threshold, we see that distinguishing discussions on *r/Sino* often combine discussions of the protests with charges of Western hypocrisy (B.88). Importantly, the connection that *r/Sino* forges between the Hong Kong protests and Western hypocrisy becomes clear when topic tuples are examined. These results suggest two dominant discursive strategies employed by users of *r/Sino* when discussing the protests—to foreground alleged violence committed by protestors and to shift discursive focus onto the hypocrisy of the West.

Table 8: Most frequent topic tuples (t=0.3) concerning Hong Kong in 2019.

r/China features	Proportion of collection	*r/Sino* features	Proportion of collection
(B.31) Critical stylistics	0.401	(B.31) Critical stylistics	0.393
(B.31, B.30) Stylistics & protests	0.087	(B.31, B.44) Stylistics & violence	0.115
(B.31, B.44) Stylistics & violence	0.073	(B.31, B.30) Stylistics & protests	0.080
(B.44) Protest violence	0.047	(B.44) Protest violence	0.066
(B.35) Sensitive phrases	0.037	(B.44, B.31) Violence & stylistics	0.048

Table 9: Most distinguishing topic tuples (t=0.3) concerning Hong Kong in 2019.

r/China features	Partial KL (bits)	*r/Sino* features	Partial KL (bits)
(B.31, B.35) Stylistics & sensitive phrases	0.0116	(B.31, B.44) Stylistics & violence	0.0331
(B.30) Protest politics	0.0112	(B.44) Protest violence	0.0148
(B.30, B.31) Protests & stylistics	0.0092	(B.31, B.88) Stylistics & Western hypocrisy	0.0109
(B.35, B.31) Sensitive phrases & stylistics	0.0073	(B.44, B.31) Violence & stylistics	0.0104
(B.31) Critical stylistics	0.0056	(B.65, B.31) Summary bot & stylistics	0.0025

Table 10: Most distinguishing topic tuples (t=0.5) concerning Hong Kong in 2019.

r/China features	Partial KL (bits)	*r/Sino* features	Partial KL (bits)
(B.31, B.30) Protests & stylistics	0.0248	(B.31, B.44) Stylistics & violence	0.0238
(B.31) Critical stylistics	0.0151	(B.31, B.44, B.88) Stylistics & violence & Western hypocrisy	0.0226
(B.31, B.30, B.35) Stylistics & protests & sensitive phrases	0.0111	(B.44, B.31) Violence & stylistics	0.0179
(B.31, B.44, B.35) Stylistics & violence & sensitive phrases	0.0099	(B.31, B.88) Stylistics & Western hypocrisy	0.0099
(B.31, B.35) Stylistics & sensitive phrases	0.0094	(B.31, B.88, B.30) Stylistics & Western hypocrisy & protests	0.0093

5 DISCUSSION

The results described in section 4 reflect the discursive tendencies that are both prevalent in *r/China* and *r/Sino* and that best differentiate the two communities. The most frequently observed features in the two communities tend to overlap. By calculating which features serve as the strongest signals of one community relative to the other, we can see beyond the features they have in common and identify the frames and discursive strategies that are conspicuous in one community in light of the other.

At the broad level of analysis that includes documents from August 2015 through November 2019, we see that *r/China* is most distinguished from *r/Sino* by its focus on individual concerns and experiences, often on the part of foreign nationals working or studying in China. These more practical, mundane, and individual-focused discourses exist in contrast to the discussions on *r/Sino* that distinguish it from *r/China*. The primary actors in *r/Sino* discussions are not the individual users, but rather, states. Rather than describing life in China, *r/Sino* describes China in terms of accomplishments—international partnerships, economic standing, technological innovation, etc. The United States also appears as an important character in *r/Sino* discourse, serving as a foil to China.

The discursive tendencies of the two communities concerning the Hong Kong protests during 2019 show that, while *r/China* is most distinguished by discussions of the political underpinnings of the protests, *r/Sino* is most distinguished by its focus on violence committed by protestors. Here we see an interesting reversal from the discursive foci that distinguished the communities more broadly. The actions of individuals are here salient in *r/Sino*, while political tensions between Hong Kong and the rest of China are salient in *r/China*. In *r/Sino*, the West (typically the US) continues to be an important part of distinguishing discourse wherein discussions about the Hong Kong protests are often nested within discourse about the hypocrisy of the West. In other words, when *r/Sino* discusses the Hong Kong protests, its users often end up discussing the West, pointing a finger back to perceived critics. This forms an important discursive strategy of *r/Sino* alongside the focus on violence committed by protestors: the flaws of the US and the Hong Kong protests are emphasized, often through concrete language about the experiences of individuals (e.g., refugee-seekers in the US and innocent victims of protestor violence in Hong Kong), while such concrete language is less prevalent when discussions are about China, which

is discussed at a more abstract, and therefore idealized, level. Aspects of this discursive strategy can also be seen in *r/China* through the focus on concrete, negative experiences of individuals living in China, while discussing the Hong Kong protests primarily in terms of more abstract, idealized entities. These discursive strategies echo the "Fallacy of the Misguided Comparison" as described by Hall and Ames [12] within the context of cross-cultural communication between the West and China. The authors describe this fallacy as the comparison of "the ideals of one society or culture with the practices of another" [12]. The implications of these findings are that popular conceptions of China from Reddit are likely to reflect such misguided comparisons, by either privileging the ideals of China (as in *r/Sino*) or its flawed realities (as in *r/China*), leaving a gap where more even-handed cross-cultural understanding between the West and China might exist.

Many of these findings come into clearer view when analyzing topic tuples as document features rather than single dominant topics. This method permits us to see topic combinations that provide important context. For example, it is not just the case that *r/Sino* uses critical language stylistics, but rather, it pairs critical language stylistics with features like protest violence and Western hypocrisy, whereas *r/China* uses the same topic in combination with describing experiences as foreign nationals.

While this analysis has yielded interesting insights, there are limitations present in the current study. Our analyses focus on the frequency of certain features at the document level, treating each document equally. However, various kinds of metadata are available from Reddit that are likely to be of interest when combined with these features. One kind of potentially interesting metadata is a document's score, which is derived from the number of positive and negative votes it received (known as upvotes and downvotes, respectively). Correlating document scores with discursive features might provide additional information on which features are not only frequent, but broadly endorsed by the community. Additionally, our focus has been on two important China-related subreddits, but there are other communities whose analysis would contribute to a larger understanding of the various China-related discourses that are active within the English-speaking world of Reddit, but with the caveats we noted in section 1.

6 CONCLUSIONS

The subreddits *r/China* and *r/Sino* represent two popular and distinct sets of English-language discursive constructions of China. For a number of reasons, understanding popular modes of discourse around China are important owing to China's international importance more broadly. Using latent word-usage patterns underlying discussions from both communities, we have examined the word-usage patterns that are most frequent in each community and that most distinguish them against a backdrop informed by the other. We find that *r/China* is broadly distinguished by a focus on the (often negative) experiences of individuals, whereas *r/Sino* is broadly distinguished by a focus on states. When we focus our analysis on discussions related to Hong Kong during 2019, we find that *r/China* is distinguished by discussions of the political underpinnings of the protests deriving from tensions between China and Hong Kong as abstract primary characters, while *r/Sino* is distinguished by a focus on violence committed by protestors (a reversal from the lack of focus on individuals more broadly) and by the tendency for discussions about the protests to foreground accusations of Western hypocrisy. These findings contribute to a broader understanding of the popular Western perspectives surrounding China.

ACKNOWLEDGMENTS

This research is funded in part by the U.S. National Science Foundation (OIA-1920920, IIS-1636933, ACI-1429160, and IIS-1110868), U.S. Office of Naval Research (N00014-10-1-0091, N00014-14-1-0489, N00014-15-P-1187, N00014-16-1-2016, N00014-16-1-2412, N00014-17-1-2605, N00014-17-1-2675, N00014-19-1-2336), U.S. Air Force Research Lab, U.S. Army Research Office (W911NF-16-1-0189), U.S. Defense Advanced Research Projects Agency (W31P4Q-17-C-0059), Arkansas Research Alliance, the Jerry L. Maulden/Entergy Endowment at the University of Arkansas at Little Rock, and the Australian Department of Defense Strategic Policy Grants Program (SPGP) (award number: 2020-106-094). Any opinions, findings, and conclusions or recommendations expressed in this material are those of the authors and do not necessarily reflect the views of the funding organizations. The researchers gratefully acknowledge the support.

REFERENCES

[1] Maria Antoniak, David Mimno, and Karen Levy. 2019. Narrative paths and negotiation of power in birth stories. In *Proc. ACM Hum.-Comput. Interact. 3, CSCW,*Article 88 (November 2019), 27 pages. DOI: https://doi.org/10.1145/3359190

[2] Alexander T. J. Barron, Jenny Huang, Rebecca L. Spang, and Simon DeDeo. 2018. Individuals, institutions, and innovation in the debates of the French Revolution. PNAS 115, 18 (May 2018), 4607–4612. DOI: https://doi.org/10.1073/pnas.1717729115

[3] David M. Blei, Andrew Y. Ng, and Michael I. Jordan. 2003. Latent Dirichlet allocation. *Journal of Machine Learning Research* 3, 993–1022.

[4] Gavin Brooks and Tony McEnery. 2019. The utility of topic modelling for discourse studies: A critical evaluation. *Discourse Studies* 21, 1 (Feb. 2019), 3–21. DOI: https://doi.org/10.1177/1461445618814032.

[5] Cody Buntain and Jennifer Golbeck. 2014. Identifying social roles in reddit using network structure. In *Proceedings of the 23rd International Conference on World Wide Web (WWW '14 Companion)*. Association for Computing Machinery, New York, NY, USA, 615–620. https://doi.org/10.1145/2567948.2579231

[6] Eshwar Chandrasekharan, Umashanthi Pavalanathan, Anirudh Srinivasan, Adam Glynn, Jacob Eisenstein, and Eric Gilbert. 2017. You Can't Stay Here: The Efficacy of Reddit's 2015 Ban Examined Through Hate Speech. *Proc. ACM Hum.-Comput. Interact.* 1, CSCW, Article 31 (December 2017), 22 pages. DOI: https://doi.org/10.1145/3134666

[7] Eshwar Chandrasekharan, Mattia Samory, Shagun Jhaver, Hunter Charvat, Amy Bruckman, Cliff Lampe, Jacob Eisenstein, and Eric Gilbert. 2018. The Internet's Hidden Rules: An Empirical Study of Reddit Norm Violations at Micro, Meso, and Macro Scales. *Proc. ACM Hum.-Comput. Interact.*2, CSCW, Article 32 (November 2018), 25 pages. DOI: https://doi.org/10.1145/3274301

[8] Jonathan Chang, Jordan Boyd-Graber, Sean Gerrish, Chong Wang, and David M. Blei. 2009. Reading tea leaves: How humans interpret topic models. In *Advances in Neural Information Processing Systems (NIPS '09)*. Vancouver, 288–296.

[9] Paul DiMaggio, Manish Nag, and David Blei. 2013. Exploiting affinities between topic modeling and the sociological perspective on culture: Application to newspaper coverage of U.S. government arts funding. *Poetics* 41, 6, 570–606. DOI: https://doi.org/10.1016/j.poetic.2013.08.004

[10] Ryan Gallagher, Andrew J. Reagan, Christopher M. Danforth, Peter Sheridan Dodds. 2018. Divergent discourse between protests and counter-protests: #BlackLivesMatter and #AllLivesMatter. *PLoS ONE* 13, 4 (Apr. 2018). DOI: https://doi.org/10.1371/journal.pone.0195644

[11] Andrew Goldstone and Ted Underwood. 2012. What can topic models teach us about the history of literary scholarship? *Journal of Digital Humanities* 2, 1 (Dec. 2012).

[12] David L. Hall and Roger T. Ames. 1999. *The Democracy of the Dead: Dewey, Confucius, and the Hope for Democracy in China* (1st. ed.). Open Court, Chicago and Lasalle, IL.

[13] David Hall, Daniel Jurafsky, and Christopher D. Manning. 2008. Studying the history of ideas using topic models. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing (EMNLP '08)*. Association for

Computational Linguistics, USA, 363–371.

[14] William L. Hamilton, Justine Zhang, Cristian Danescu-Niculescu-Mizil, Dan Jurafsky, and Jure Leskovec. 2017. Loyalty in online communities. In *Proceedings of the 11th International AAAI Conference on Web and Social Media (ICWSM '17)*. AAAI Press, 540–543. https://www.aaai.org/ocs/index.php/ICWSM/ICWSM17/paper/view/15710/14848

[15] Marianne Jørgensen and Louise J. Phillips. 2002. Discourse Analysis as Theory and Method (1st. ed.). Sage, London.

[16] Sara Klingenstein, Tim Hitchcock, and Simon DeDeo. 2014. The civilizing process in London's Old Bailey. *PNAS* 111, 26 (Jul. 2014), 9419–9424. DOI: https://doi.org/10.1073/pnas.1405984111

[17] Humphrey Mensah, Lu Xiao, and Sucheta Soundarajan. 2019. Characterizing susceptible users on Reddit's ChangeMyView. In *Proceedings of the 10th International Conference on Social Media and Society (SMSociety '19)*. Association for Computing Machinery, New York, NY, USA, 102–107. https://doi.org/10.1145/3328529.3328550

[18] David Mimno, Hanna M. Wallach, Edmund Talley, Miriam Leenders, and Andrew McCallum. 2011. Optimizing semantic coherence in topic models. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing (EMNLP '11)*. Association for Computational Linguistics, USA, 262–272.

[19] Fred Morstatter, Jürgen Pfeffer, Huan Liu, and Kathleen M. Carley. 2003. Is the sample good enough? Comparing data from Twitter's streaming API with Twitter's firehose. In *Proceedings of the 7th International AAAI Conference on Weblogs and Social Media,* AAAI Press, Cambridge, MA., 400–408.

[20] Dong Nguyen, Maria Liakata, Simon DeDeo, Jacob Eisenstein, David Mimno, Rebekah Tromble, and Jane Winters. 2019. How we do things with words: Analyzing text as social and cultural data. arXiv:1907.01468. Retrieved from https://arxiv.org/abs/1907.01468

[21] Ryan Nichols, Edward Slingerland, Kristoffer Nielbo, Uffe Bergeton, Carson Logan, and Scott Kleinman. 2018. Modeling the contested relationship between *Analects, Mencius,* and *Xunzi*: Preliminary evidence from a machine-learning approach. *The Journal of Asian Studies* 77, 1 (Feb. 2018), 19–57. DOI: https://doi.org/10.1017/S0021911817000973

[22] Radim Řehůřek and Petr Sojka. 2010. Software framework for topic modelling with large corpora. In *Proceedings of the LREC 2010 Workshop on New Challenges for NLP Frameworks*. ELRA, Malta, 45–50.

[23] Margaret E. Roberts, Brandon M. Stewart, and Dustin Tingley. 2016. Navigating the Local Modes of Big Data: The Case of Topic Models. In *Computational Social Science: Discovery and Prediction* Cambridge University Press, New York, NY, 51–97.

[24] Alexandra Schofield and David Mimno. 2016. Comparing apples to apple: The effects of stemmers on topic models. In *Transactions of the Association for Computational Linguistics* 4, 287–300. DOI: https://doi.org/10.1162/tacl_a_00099

[25] Chenhao Tan, Vlad Niculae, Cristian Danescu-Niculescu-Mizil, and Lillian Lee. 2016. Winning arguments: Interaction dynamics and persuasion strategies in good-faith online discussions. In *Proceedings of the 25th International Conference on World Wide Web (WWW '16)*. International World Wide Web Conferences Steering Committee, Republic and Canton of Geneva, CHE, 613–624. https://doi.org/10.1145/2872427.2883081

[26] Laure Thompson and David Mimno. 2018. Authorless topic models: Biasing models away from known structure. In *Proceedings of the 27th International Conference on Computational Linguistics*. Association for Computational Linguistics, USA, 3903–3914.

[27] Anton Törnberg and Petter Törnberg. 2016. Muslims in social media discourse: Combining topic modeling and critical discourse analysis. *Discourse, Context and Media* 13, 132–142. DOI: https://doi.org/10.1016/j.dcm.2016.04.003.

[28] Anton Törnberg and Petter Törnberg. 2016. Combining CDA and topic modeling: Analyzing discursive connections between Islamophobia and anti-feminism on an online forum. *Discourse & Society* 27, 4, 401–422. DOI: https://doi.org/10.1177/0957926516634546

[29] Hanna M. Wallach, Iain Murray, Ruslan Salakhutdinov, and David Mimno. 2009. Evaluation methods for topic models. In *Proceedings of the 26th Annual International Conference on Machine Learning (ICML '09)*. Association for Computing Machinery, New York, NY, USA, 1105–1112. DOI: https://doi.org/10.1145/1553374.1553515