TopicCheck: Interactive Alignment for Assessing Topic Model Stability

Jason Chuang[*]
jason@chuang.ca

Margaret E. Roberts[†]
Political Science
U. California, San Diego
meroberts@ucsd.edu

Brandon M. Stewart[†]
Government
Harvard University
bstewart@fas.harvard.edu

Rebecca Weiss[†]
Communication
Stanford University
rjweiss@stanford.edu

Dustin Tingley
Government
Harvard University
dtingley@gov.harvard.edu

Justin Grimmer
Political Science
Stanford University
jgrimmer@stanford.edu

Jeffrey Heer
Computer Science & Eng.
University of Washington
jheer@uw.edu

Abstract

Content analysis, a widely-applied social science research method, is increasingly being supplemented by topic modeling. However, while the discourse on content analysis centers heavily on reproducibility, computer scientists often focus more on scalability and less on coding reliability, leading to growing skepticism on the usefulness of topic models for automated content analysis. In response, we introduce TopicCheck, an interactive tool for assessing topic model stability. Our contributions are threefold. First, from established guidelines on reproducible content analysis, we distill a set of design requirements on how to computationally assess the stability of an automated coding process. Second, we devise an interactive alignment algorithm for matching latent topics from multiple models, and enable sensitivity evaluation across a large number of models. Finally, we demonstrate that our tool enables social scientists to gain novel insights into three active research questions.

1 Introduction

Content analysis — the examination and systematic categorization of written texts (Berelson, 1952) — is a fundamental and widely-applied research method in the social sciences and humanities (Krippendorff, 2004a), found in one third of all articles published in major communication journals (Wimmer and Dominick, 2010). Initial reading and coding, two labor-intensive steps in the analysis process, are increasingly replaced by computational approaches such as statistical topic modeling (Grimmer, 2013; McFarland et al., 2013; Roberts et al., 2014a).

However, while the discourse on content analysis overwhelmingly centers around the reproducibility and generalizability of a coding scheme (Krippendorff, 2004b; Lombard et al., 2002), computer scientists tend to focus more on increasing the scale of analysis and less on establishing coding reliability. Machine-generated latent topics are often taken on faith to be a truthful and consistent representation of the underlying corpus, but in practice exhibit significant variations among models or modeling runs. These unquantified uncertainties fuel growing skepticism (Schmidt, 2012) and hamper the continued adoption (Grimmer and Stewart, 2011) of topic models for automated content analysis.

In response, we introduce TopicCheck, an interactive tool for assessing the stability of topic models. Our threefold contributions are as follows.

First, from established guidelines on reproducible content analysis, we distill a set of design requirements on how to computationally assess the stability of an automated coding process. We advocate for the use of multiple models for analysis, a user-driven approach to identify acceptable levels of coding uncertainty, and providing users with the capability to inspect model output at all levels of detail.

Second, we devise an interactive *up-to-one* alignment algorithm for assessing topic model stability. Through repeated applications of a topic model to generate multiple outputs, our tool allows users to inspect whether the model consistently uncover the

[*] Work completed while at Stanford University and the University of Washington, and submitted while at the Allen Institute for Artificial Intelligence.

[†] These authors contributed equally to this paper.

Human Language Technologies: The 2015 Annual Conference of the North American Chapter of the ACL, pages 175–184,
Denver, Colorado, May 31 – June 5, 2015.

same set of concepts. We allow users to interactively define groupings of matching topics, and present the aligned topics using an informative tabular layout, so that users can quickly identify stable topical groupings as well as any inconsistencies.

Finally, in three case studies, we demonstrate that our tool allows social scientists to gain novel insights into active and ongoing research questions. We provide an in-depth look at the multi-modality of topic models. We document how text pre-processing alters topical compositions, causing shifts in definitions and the removal of select topics. We report on how TopicCheck supports the validity of newly-proposed communication research methods.

2 Background

Manual approaches to extract information from textual data — reading the source documents and codifying notable concepts — do not scale. For example, Pew Research Center produces the News Coverage Index (2014) to measure the quality of news reporting in the United States. Intended to track 1,450 newspapers nationwide, their purely manual efforts only cover 20 stories per day. Researchers stand to lose rich details in their data when their attention is limited to a minuscule fraction of the available texts.

Critical of approaches that "*[make] restrictive assumptions or [are] prohibitively costly*," Quinn et al. (2010) discuss the use of topic models (Blei et al., 2003) to enable large-scale text analysis by using machine-generated latent topics to approximate previously manually-crafted codes. Automated content analysis has enabled groundbreaking massive studies (Grimmer, 2013; McFarland et al., 2013; Roberts et al., 2014a). While this initial uptake of topic models is encouraging, an over-emphasis on scalability and the use of a single model for analysis invites skepticism and threatens continued adoption.

2.1 Coding Reliability & Growing Skepticism

Coding reliability is critical to content analysis. When social scientists devise a coding scheme, they must clearly articulate the definition of their codes in such a way that any person can consistently apply the given codes to all documents in a corpus.

Despite high labor cost, content analysis is typically conducted with multiple coders in order to establish coding reliability; the proper application of reliability measures is heavily discussed and debated in the literature (Krippendorff, 2004b; Lombard et al., 2002). In contrast, software packages (McCallum, 2013; Řehůřek and Sojka, 2010) and graphical tools (Chaney and Blei, 2014; Chuang et al., 2012b) have made topic models accessible, cheap to compute, easy to deploy, but they almost always present users with a single model without any measure of uncertainty; we find few studies on topic model sensitivity and no existing tool to support such analyses.

Schmidt (2012) summarizes the view among digital humanists, a group of early adopters of topic models, on the experience of working with uncertain modeling results: "*A poorly supervised machine learning algorithm is like a bad research assistant. It might produce some unexpected constellations that show flickers of deeper truths; but it will also produce tedious, inexplicable, or misleading results. . . . [Excitement] about the use of topic models for discovery needs to be tempered with skepticism about how often the unexpected juxtapositions. . . will be helpful, and how often merely surprising.*"

Researchers increasingly voice skepticism about the validity of using single models for analysis. In a comprehensive survey of automatic content analysis methods, Grimmer et al. (2011) highlight the need to validate models through close reading and model comparison, and advise against the use of software that "*simply provide the researcher with output*" with no capability to ensure the output is conceptually valid and useful. Chuang et al. (2012a) report that findings from one-off modeling efforts may not sustain under scrutiny. Schmidt (2012) argues that computer-aided text analysis should incorporate competing models or "*humanists are better off applying zero computer programs.*"

2.2 Uncertainties in Topic Models

While topic models remove some issues associated with human coding, they also introduce new sources of uncertainties. We review three factors related to our case studies: multi-modality, text pre-processing, and human judgment of topical quality.

Roberts et al. (2014b) examine the multi-modal distributions of topic models that arise due to the non-convex nature of the underlying optimization. They characterize the various local solutions, and demonstrate that the spread of topics can lead to contradictory analysis outcomes. The authors note that optimal coding may not necessarily correspond to models that yield the highest value of the objective function, but there is currently a paucity of computational tools to inspect how the various modes differ, help researchers justify why one local mode might be preferred over another on the basis of their domain knowledge, or for an independent researcher to validate another's modeling choices.

Fokkens et al. (2013) report widespread reproducibility failures in natural language processing when they replicate — and fail to reproduce — the results reported on two standard experiments. The authors find that minor decisions in the modeling process can impact evaluation results, including two factors highly relevant to topic modeling: differences in text pre-processing and corpus vocabulary.

The word intrusion test (Chang et al., 2009; Lau et al., 2014) is considered the current state-of-the-art approach to assess topical quality, and captures human judgment more accurately than other topical coherence measures (Stevens et al., 2012; Wallach et al., 2009). However, in this approach, users inspect only a single latent topic at a time without access to the overall set of topics. As a part of this paper, we investigate whether exposure to multiple competing models affects human judgment, and whether model consistency impacts topical coherence.

2.3 Reproducibility of a Coding Process

While no single definition exists for the process of content analysis, a frequently-cited and wide-applied template is provided by Krippendorff (1989; 2004b) who recommends four steps to safeguard the reproducibility of a coding process. Practitioners must demonstrate *coder reliability*, *a decisive agreement coefficient*, *an acceptable level of agreement*, and test *individual variables*.

To the best of our knowledge, our paper is the first to convert guidelines on reproducible human coding into software design requirements on validating automated content analysis. Our interactive alignment algorithm is the first implementation of these guidelines. Our case studies represent the first reports on the impact of computationally quantifying topic model uncertainties, situated within the context of real-world ongoing social science research.

Much of the research on topic modeling focuses on model designs (Blei et al., 2004; Blei and Lafferty, 2006; Rosen-Zvi et al., 2004) or inference algorithms (Anandkumar et al., 2012). Our tool is complementary to this large body of work, and supports real-world deployment of these techniques. Interactive topic modeling (Hu et al., 2014) can play a key role to help users not only verify model consistency but actively curate high-quality codes; its inclusion is beyond the scope of a single conference paper. While supervised learning (Settles, 2011) has been applied to content analysis, it represents the application of a pre-defined coding scheme to a text corpus, which is different from the task of devising a coding scheme and assessing its reliability.

3 Validation Tool Design Requirements

A measure of coding reproducibility is whether a topic model can consistently uncover the same set of latent topics. We assume that users have a large number of topic model outputs, presumed to be identical, and that the users wish to examine unexpected variations among the outputs. To guide tool development, we first identify software design requirements, to meet the standards social scientists need to demonstrate producible coding.

3.1 Topical Mapping & Up-to-One Alignment

A key difference exists between measuring inter-coder agreement and assessing topic model variations. In a manual coding process, human coders are provided code identifiers; responses from different coders can be unambiguously mapped onto a common scheme. No such mapping exists among the output from repeated runs of a topic model. Validation tools must provide users with **effective means to generate topical mapping**.

However, the general alignment problem of optimally mapping *multiple* topics from one model to *multiple* topics in another model is both ill-defined and computationally intractable. Since our tool is to support the comparison of similar — and supposedly identical — model output, we impose the following constraint. A latent topic belonging to a model may align with *up to one* latent topic in another model. We avoid the more restrictive constraint of *one-to-one* alignment. Forcing a topic to always map onto another topic may cause highly dissimilar topics to be grouped together, obscuring critical mismatches. Instead, up-to-one mapping allows for two potential outcomes, both of which correspond directly to the intended user task: recognize consistent patterns across the models (when alignment occurs) and identify any deviations (when alignment fails).

3.2 Guidelines Adapted for Topic Models

We synthesize the following four requirements from Krippendorff's guidelines (2004b).

To calculate the equivalent of *coder reliability*, we advocate the **use of multiple models to determine modeling consistency**, which may be determined from the repeated applications of the same topic model, a search through the parameter space of a model, or the use of multiple models.

Selecting an appropriate *agreement coefficient* depends on the underlying data type, such as binary, multivariate, ordered, or continuous codes (Cohen, 1960; Holsti, 1969; Krippendorff, 1970; Osgood, 1959; Scott, 1995). No widely-accepted similarity measure exists for aligning latent topics, which are probability distributions over a large vocabulary. We argue that validation tools must be sufficiently modular, in order to **accept any user-defined topical similarity measure** for aligning latent topics.

Acceptable level of agreement depends on the purpose of the analysis, and should account for the costs of drawing incorrect conclusions from a coding scheme. For example, do "*human lives hang on the results of a content analysis?*" (Krippendorff, 2004b). Validation tools must **allow users to set the appropriate acceptable level of agreement**, and help users determine — rather than dictate — when topic models match and what constitutes reasonable variations in the model output.

Finally, Krippendorff points out that aggregated statistics can obscure critical reliability failures, and practitioners must test *individual variables*. We interpret this recommendation as the need to **present users with not a single overall alignment score but details at all levels**: models, topics, and constituent words within each latent topic.

4 Interactive Topical Alignment

We introduce TopicCheck, an implementation of our design specifications. At the core of this tool is an interactive topical alignment algorithm.

4.1 Hierarchical Clustering with Constraints

Our algorithm can be considered as hierarchical agglomerative clustering with up-to-one mapping constraints. As input, it takes in three arguments: a list of topic models, a topical similarity measure, and a matching criterion. As output, it generates a list of topical groups, where each group contains a list of topics with at most one topic from each model.

At initialization, we create a topical group for every topic in every model. We then iteratively merge the two most similar groups based on the user-supplied topical similarity measure, provided that the groups satisfy the user-specified matching criterion and the mapping constraints. When no new groups can be formed, the algorithm terminates and returns a sorted list of final topical groups.

During the alignment process, the following two invariants are guaranteed: Every topic is always assigned to exactly one group; every group contains at most one topic from each model. A topic model m consists of a list of latent topics. A latent topic t is represented by a probability distribution over words. A topical group g also consists of a list of latent topics. Let $|m|$, $|t|$, and $|g|$ denote the number of models, topics, and groups respectively. We create a total of $|g| = |m| \times |t|$ initial topical groups. Although $|g|$ decreases by 1 after each merge, $|g| \geq |t|$ at all times. At the end of alignment, $|g| = |t|$ if and only if perfect alignment occurs and every group contains exactly one topic from each model.

Users may supply any topical similarity measure that best suits their analysis needs. We select cosine similarity for our three case studies, though our software is modular and accepts any input. As a first implementation, we apply single-linkage clustering criteria when comparing the similarity of two topical groups. Single-linkage clustering is computationally efficient (Sibson, 1973), so that users may interact with the algorithm and receive feedback in real-time; our procedure generalizes to other linkage criteria such as complete-linkage or average-linkage.

At each merge step, the most similar pair of topical groups are identified. If they meet the matching criteria and the mapping constraints, the pair is combined into a new group. Otherwise, the algorithm iteratively examines the next most similar pair until either a merge occurs or when all pairs are ex-

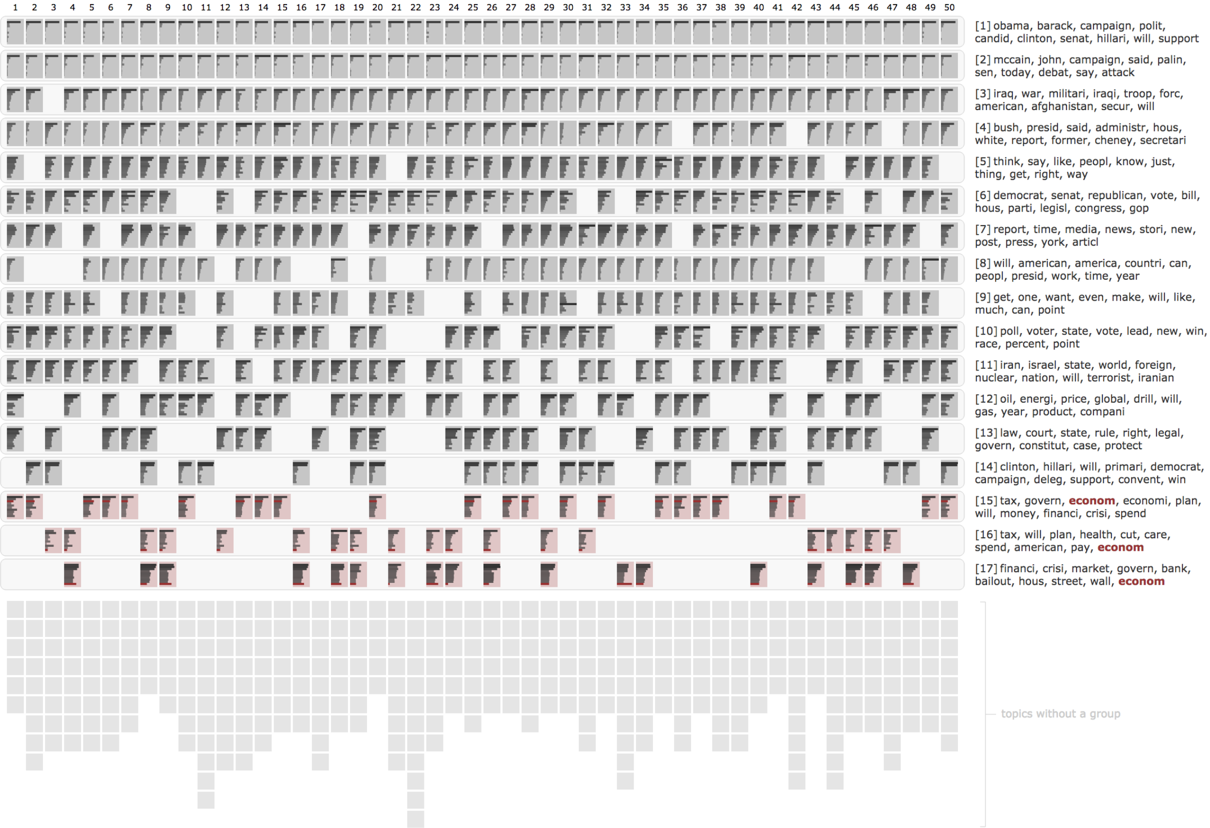

Figure 1: This chart shows topics uncovered from 13,250 political blogs (Eisenstein and Xing, 2010) by 50 structural topic models (Roberts et al., 2013). Latent topics are represented as rectangles; bar charts within the rectangles represent top terms in a topic. Topics belonging to the same model are arranged in a column; topics assigned to the same group are arranged in a row. This chart is completely filled with topics only if perfect alignment occurs. When topics in a model fail to align with topics in other models, empty cells appear in its column. Similarly, when topics in a group are not consistently uncovered by all models, empty cells appear in its row. Hovering over a term highlights all other occurrences of the same term. Top terms belonging to each topical group are shown on the right; they represent the most frequent words over all topics in the group, by summing their probability distributions.

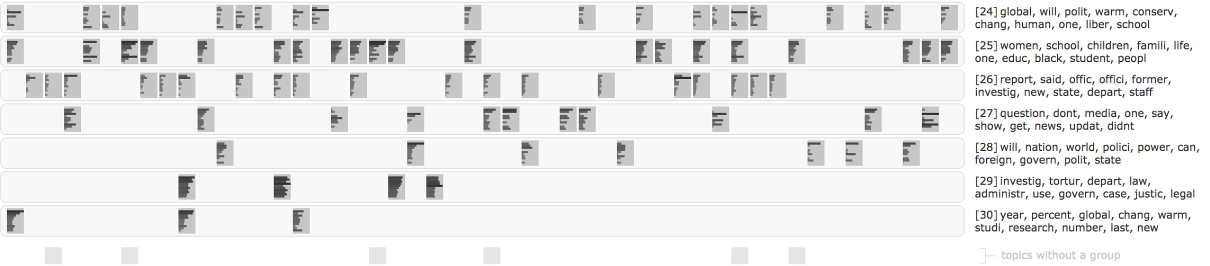

Figure 2: Continued from Figure 1, users may decrease the similarity threshold to generate additional groupings of topics that are less consistent, uncovered by as few as 3 of the 50 modeling runs.

hausted, at which point the procedure terminates.

Users can specify a similarity threshold, below which topical groups are considered to differ too much to be matched. Two groups are allowed to merge only if both of the following conditions are met: their similarity is above the user-defined similarity threshold *and* every topic in the combined group belongs to a different model.

4.2 Tabular Layout and User Interactions

We devise a tabular layout to present the alignment output at all levels of detail: groups, models, topics, and words. Users can interact with the algorithm, redefine matching criteria, and inspect the aligned models interactively in real-time.

We arrange topical groups as rows and topic models as columns as shown in Figure 1. A topic assigned to group g_i and belonging to model m_j is placed at the intersection of row i and column j. Our up-to-one mapping ensures at most one topic per each cell. A table of size $|g| \times |m|$ will only be completely filled with topics if perfect alignment occurs. When topics in model m_j fail to align with topics in other models, empty cells appear in column j. Similarly, when topics in group g_i are not consistently uncovered by all models, empty cells appear in row i. Within each topic, we show the probability distribution of its constituent words as a bar chart.

Users define three parameters in our tool. First, they may set the matching criteria, and define how aggressively the topics are merged into groups. Second, users may alter the number of topical groups to reveal. Rather than displaying numerous sparse groups, the tool shows only the top groups as determined by their topical weight. Topics in all remaining groups are placed at the bottom of the table and marked as *ungrouped*. Third, users may adjust the number of top terms to show, as a trade-off between details vs. overview. Increasing the number of terms allows users to inspect the topics more carefully, but the cells take up more screen space, reducing the number of visible groups. Decreasing the number of terms reduces the size of each cell, allowing users to see more groups and observe high-level patterns.

The tabular layout enables rapid visual assessment of consistency within a model or a group. We further facilitate comparisons via brushing and linking (Becker and Cleveland, 1987). When users hover over a word on the right hand side or over a bar within the bar charts, we highlight all other occurrences of the same word. For example, in Figure 1, hovering over the term *econom* reveals that the word is common in three topical groups.

5 Deployment and Initial Findings

We implemented our alignment algorithm and user interface in JavaScript, so they are easily accessible within a web browser; topical similarity is computed on a Python-backed web server. We report user responses and initial findings from deploying the tool on three social science research projects. Interactive versions of the projects are available at `http://content-analysis.info/naacl`.

5.1 A Look at Multi-Modal Solutions

We deployed TopicCheck on topic models generated by Roberts et al. (2014b) to examine how model output clusters into local modes. As the models are produced by 50 runs of an identical algorithm with all pre-processing, parameters, and hyper-parameters held constant, we expect minimal variations.

As shown in Figure 1, we observe that the top two topical groups, about Barack Obama and John McCain respectively, are consistently uncovered across all runs. The third topical group, about the Iraqi and Afghani wars (defined by a broader set of terms) is also consistently generated by 49 of the 50 runs.

Toward the bottom of the chart, we observe signs of multi-modality. Topical groups #15 to #17 represent variations of topics about the economy. Whereas group #15 is about the broader economy, groups #16 and #17 focus on taxes and the financial crisis, respectively. Half of the runs produced the broader economy topic; the other runs generated only one or two of the specialized subtopics. No single model uncovered all three, suggesting that the inference algorithm converged to one of two distinct local optimal solutions. In Figure 2, by lowering the matching criteria and revealing additional groups, we find that the model continues to produce interesting topics such as those related to global warming (group #24) or women's rights (group #25), but these topics are not stable across the multiple modes.

5.2 Text Pre-Processing & Replication Issues

We conducted an experiment to investigate the effects of rare word removal using TopicCheck. As a part of our research, we had collected 12,000 news reports from five different international news sources over a period of ten years, to study systematic differences in news coverage on the rise of China, between western and Chinese media.

While many modeling decisions are involved in our analysis, we choose rare word removal for two reasons. First, though the practice is standard, to the best of our knowledge, we find no systematic studies on how aggressively one should cull the vocabulary.

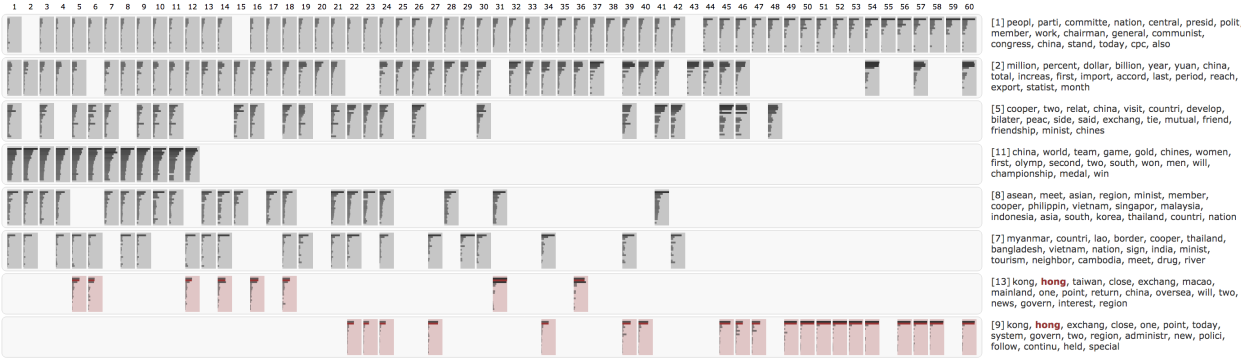

Figure 3: While rare word removal is generally considered to have limited impact on topic model output, we find evidence to the contrary. By varying the removal threshold, for this corpus of international news reports on the rise of China, we observe that topics such as group #11 on the Beijing Olympics begin to disappear. Topics about Hong Kong appear sporadically. On top of the inconsistency issues, different pre-processing settings lead to drifts in topic definitions. For milder removal thresholds (toward the left), group #13 discusses Hong Kong within the context of Taiwan and Macau. With more aggressive filtering (toward the right), group #14 shifts into discussions about Hong Kong itself such as *one country two systems* and the *special administrative region*. Unchecked, these seemingly minor text pre-processing decisions may eventually lead researchers down different paths of analysis.

Second, as latent topics are typically defined through their top words, filtering words that occur only in a small fraction of the documents is generally considered to have limited impact on model output.

We trained structural topic models (Roberts et al., 2013) based on a subset of the corpus with 2,398 documents containing approximately 20,000 unique words. We applied 10 different settings where we progressively removed a greater number of rare terms beyond those already filtered by the default settings while holding all other parameters constant. The number of unique words retained by the models were 1,481 (default), 904, 634, 474, 365, . . ., down to 124 for the 10 settings. We generated 6 runs of the model at each setting, for a total of 60 runs. Removed words are assigned a value of 0 in the topic vector when computing cosine similarity.

We observe significant changes to the model output across the pre-processing settings, as shown in Figure 3. The six models on the far left (columns 1 to 6) represent standard processing; rare word removal ranges from the mildest (columns 7 to 12) to the most aggressive (columns 55 to 60) as the columns move from left to right across the chart.

While some topical groups (e.g., #1 on the communist party) are stable across all settings, many others fade in and out. Group #11 on the Beijing Olympics is consistent under standard processing and the mildest removal, but disappears completely afterward. We find two topical groups about Hong Kong that appear sporadically. On top of the instability issues, we observe that their content drifts across the settings. With milder thresholds, topical group #13 discusses Hong Kong within the context of Taiwan and Macau. With more aggressive filtering, topical group #14 shifts into discussions about Hong Kong itself such as *one country two systems* and the *special administrative region*. Unchecked, these minor text pre-processing decisions may lead researchers down different paths of analysis.

5.3 News Coverage & Topical Coherence

Agenda-setting refers to observations by McCombs et al. (1972) that the media play an important role in dictating issues of importance for voters, and by Iyengar et al. (1993) that news selection bias can determine how the public votes. Studying agenda-setting requires assessing the amount of coverage paid to specific issues. Previous manual coding efforts are typically limited to either a single event or subsampled so thinly that they lose the ability to consistently track events over time. Large-scale analysis (e.g., for an entire federal election) remains beyond the reach of most communication scholars.

As part of our research, we apply topic modeling to closed-captioning data from over 200,000 hours of broadcasts on all mainstream news networks, to track the full spectrum of topics across all media out-

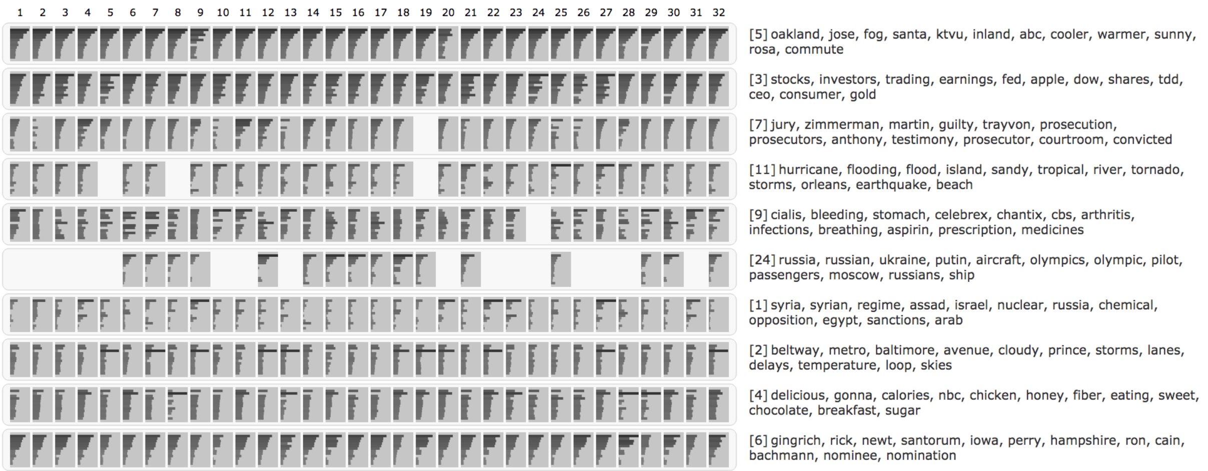

Figure 4: To enable large-scale studies of agenda-setting, we applied topic modeling to closed-captioning of over 200,000 hours of broadcasts, to estimate coverage in mainstream news networks. Through TopicCheck, the researchers find consistent topical groups that correspond to known major news categories. Group #9 represents topics about advertisements and valuable data to study the relationships between broadcasters and advertisers.

lets. We conduct word intrusion tests (Chang et al., 2009) on Amazon Mechanical Turk, and obtain over 50,000 user ratings to identify high quality topics. However, to establish topic modeling as a valid research method, we must demonstrate the reliability of how we include or exclude topics in our analyses.

By applying TopicCheck to 32 runs of the same topic model, as shown in Figure 4, we confirm that the consistent topical groupings capture at least four major known news categories: weather (such as group #5), finance (group #3), major events (group #7 on the Trayvon Martin shooting), and natural disasters (group #11 on Hurricane Katrina). We find additional evidence supporting the use of topic models, including the consistent appearance of advertising topics (group #9 on the sales of prescription medicine to senior citizens, a major demographic of the broadcast news audience). These topics may enable studies on the relationship between broadcasters and advertisers, an important but difficult question to address because few previous studies have the resources to codify advertisement content.

However, event-specific topics tend to appear less consistently (such as group #24 on Russia, its conflict with Ukraine, and the Sochi Olympics). We note the lack of consistent topics on supreme court cases, an expected but missing news category, which warrants more in-depth investigations.

We compare human judgment of topical quality when examining multiple models and those based on word intrusion tests. We calculate the aggregated topical coherence scores for each topical grouping. We find that consistent topical groups tend to receive higher coherence scores. However, topics about natural disasters receive low scores with a high variance (avg 0.5371; stdev 0.2497); many of them would have previously been excluded from analysis.

6 Discussions

To many social scientists, statistical models are measurement tools for inspecting social phenomena, such as probing recurring language use in a text corpus with topic models. In this light, instruments with known performance characteristics — including well-quantified uncertainties and proper coverage — are more valuable than potentially powerful but inconsistent modeling approaches.

Our initial findings suggest that a single topic model may not capture all perspectives on a dataset, as evident in the multiple local solutions about the economy, Hong Kong, and natural disasters in the three case studies respectively. By exposing model stability, our tool can help researchers validate modeling decisions, and caution against making too general a claim about any single modeling result.

We hypothesize that the low coherence scores for topics about natural disasters might derive from two causes. First, news media might cover an event differently (e.g., focusing on economic vs. humanitarian issues during Hurricane Katrina). Second, unfolding events may naturally have less stable vocabularies. In both cases, detecting and pinpointing reporting bias is central to the study of agenda-setting. These observations suggest that for certain applications, identifying consistent topics across multiple models may be equally critical as, if not more than, enforcing topical coherence within a single model.

Increasingly, text analysis relies on data-dependent modeling decisions. Rare word removal can substantively alter analysis outcomes, but selecting an appropriate threshold requires inspecting the content of a text corpus. TopicCheck can help archive the exact context of analysis, allowing researchers to justify — and readers to verify and challenge — modeling decisions through access to data.

Finally, topic modeling has dramatically lowered the costs associated with content analysis, allowing hundreds of models to be built in parallel. The current intended user task for TopicCheck is to validate the stability of presumably identical models. We plan to develop additional tools to help social scientists design better models, and actively explore the effects of alternative coding schemes.

7 Conclusion

We present TopicCheck for assessing topic model stability. Through its development, we demonstrate that existing research on reproducible manual codification can be transferred and applied to computational approaches such as automated content analysis via topic modeling. We hope this work will help computer scientists and social scientists engage in deeper conversations about research reproducibility for large-scale computer-assisted text analysis.

Acknowledgments

This research was supported in part by a grant from the Brown Institute for Media Innovation.

References

Anima Anandkumar, Yi kai Liu, Daniel J. Hsu, Dean P Foster, and Sham M Kakade. 2012. A spectral algorithm for latent dirichlet allocation. In *Neural Information Processing Systems (NIPS)*, pages 917–925.

Richard A. Becker and William S. Cleveland. 1987. Brushing scatterplots. *Technometrics*, 29(2):127–142.

Bernard Berelson. 1952. *Content analysis in communication research*. Free Press.

David M. Blei and John D. Lafferty. 2006. Dynamic topic models. In *International Conference on Machine Learning (ICML)*, pages 113–120.

David M. Blei, Andrew Y. Ng, and Michael I. Jordan. 2003. Latent Dirichlet allocation. *Journal of Machine Learning Research*, 3(1):993–1022.

David M. Blei, Thomas L. Griffiths, Michael I. Jordan, and Joshua B. Tenenbaum. 2004. Hierarchical topic models and the nested chinese restaurant process. In *Neural Information Processing Systems (NIPS)*.

Allison June-Barlow Chaney and David M. Blei. 2014. Visualizing topic models. In *International Conference on Weblogs and Social Media (ICWSM)*, pages 419–422.

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

Jason Chuang, Christopher D. Manning, and Jeffrey Heer. 2012a. Interpretation and trust: Designing model-driven visualizations for text analysis. In *Conference on Human Factors in Computing Systems (CHI)*, pages 443–452.

Jason Chuang, Christopher D. Manning, and Jeffrey Heer. 2012b. Termite: Visualization techniques for assessing textual topic models. In *Advanced Visual Interfaces (AVI)*.

Jacob Cohen. 1960. A coefficient of agreement for nominal scales. *Educational and Psychological Measurement*, 20:37–46.

Jacob Eisenstein and Eric Xing. 2010. *The CMU 2008 Political Blog Corpus*. Carnegie Mellon University.

Antske Fokkens, Marieke van Erp, Marten Postma, Ted Pedersen, Piek Vossen, and Nuno Freire. 2013. Offspring from reproduction problems: What replication failure teaches us. In *Annual Meeting of the Association for Computational Linguistics (ACL)*, pages 1691–1701.

Justin Grimmer and Brandon M. Stewart. 2011. Text as data: The promise and pitfalls of automatic content analysis methods for political texts. *Political Analysis*, 21(3):267–297.

Justin Grimmer. 2013. Appropriators not position takers: The distorting effects of electoral incentives on congressional representation. *American Journal of Political Science*, 57(3):624–642.

Ole R. Holsti. 1969. *Content analysis for the social sciences and humanities*. Addison-Wesley Publishing Company.

Yuening Hu, Jordan Boyd-Graber, Brianna Satinoff, and Alison Smith. 2014. Interactive topic modeling. *Machine Learning*, 95(3):423–469.

Shanto Iyengar and Adam Simon. 1993. News coverage of the gulf crisis and public opinion: A study of agenda-setting, priming, and framing. *Communication Research*, 20(3):365–383.

Klaus Krippendorff. 1970. Bivariate agreement coefficients for reliability of data. In E. R. Borgatta and G. W. Bohrnstedt, editors, *Sociological methodology*, pages 139–150. John Wiley & Sons.

Klaus Krippendorff. 1989. Content analysis. In E. Barnouw, G. Gerbner, W. Schramm, T. L. Worth, and L. Gross, editors, *International encyclopedia of communication*. Oxford University Press.

Klaus Krippendorff. 2004a. *Content analysis: An introduction to its methodology*. Sage, 2nd edition.

Klaus Krippendorff. 2004b. Reliability in content analysis: Some common misconceptions and recommendations. *Human Communication Research*, 30(3):411–433.

Jey Han Lau, David Newman, and Timothy Baldwin. 2014. Machine reading tea leaves: Automatically evaluating topic coherence and topic model quality. In *Conference of the European Chapter of the Association for Computational Linguistics (EACL)*, pages 530–539.

Matthew Lombard, Jennifer Snyder-Duch, and Cheryl Campanella Bracken. 2002. Content analysis in mass communication: Assessment and reporting of intercoder reliability. *Human Communication Research*, 28(4):587–604.

Andrew McCallum. 2013. MALLET: A machine learning for language toolkit. `http://mallet.cs.umass.edu`.

Maxwell E. McCombs and Donald L. Shaw. 1972. The agenda-setting function of mass media. *Public Opinion Quarterly*, 36(5):176–187.

Daniel A. McFarland, Daniel Ramage, Jason Chuang, Jeffrey Heer, and Christopher D. Manning. 2013. Differentiating language usage through topic models. *Poetics: Special Issue on Topic Models and the Cultural Sciences*, 41(6):607–625.

C. E. Osgood. 1959. The representational model and relevant research. In I. de Sola Pool, editor, *Trends in content analysis*, pages 33–88. University of Illinois Press.

Ted Pedersen. 2008. Empiricism is not a matter of faith. *Computational Linguistics*, 34(3):465–470.

Pew Research Journalism Project. 2014. News coverage index methodology. `http://www.journalism.org/news_index_methodology/99/`.

Kevin M. Quinn, Burt L. Monroe, Michael Colaresi, Michael H. Crespin, and Dragomir R. Radev. 2010. How to analyze political attention with minimal assumptions and costs. *American Journal of Political Science*, 54(1):209–228.

Radim Řehůřek and Petr Sojka. 2010. Software framework for topic modelling with large corpora. In *LREC Workshop on New Challenges for NLP Frameworks*, pages 45–50.

Margaret E. Roberts, Brandon M. Stewart, Dustin Tingley, and Edoardo M. Airoldi. 2013. The structural topic model and applied social science. In *NIPS Workshop on Topic Models*.

Margaret E. Roberts, Brandon Stewart, Dustin Tingley, Chris Lucas, Jetson Leder-Luis, Bethany Albertson, Shana Gadarian, and David Rand. 2014a. Topic models for open-ended survey responses with applications to experiments. *American Journal of Political Science*. Forthcoming.

Margaret E. Roberts, Brandon M. Stewart, and Dustin Tingley. 2014b. Navigating the local modes of big data: The case of topic models. In R. Michael Alvarez, editor, *Data Science for Politics, Policy and Government*. In Press.

Michal Rosen-Zvi, Thomas Griffiths, Mark Steyvers, and Padhraic Smyth. 2004. The author-topic model for authors and documents. In *Conference on Uncertainty in Artificial Intelligence (UAI)*, pages 487–494.

Benjamin M. Schmidt. 2012. Words alone: Dismantling topic models in the humanities. *Journal of Digital Humanities*, 2(1).

William A. Scott. 1995. Reliability of content analysis:: The case of nominal scale coding. *Public Opinion Quarterly*, 19(3):321–325.

Burr Settles. 2011. Closing the loop: Fast, interactive semi-supervised annotation with queries on features and instances. In *Empirical Methods in Natural Language Processing (EMNLP)*, pages 1467–1478.

Robin Sibson. 1973. SLINK: an optimally efficient algorithm for the single-link cluster method. *The Computer Journal*, 16:30–34.

Keith Stevens, Philip Kegelmeyer, David Andrzejewski, and David Buttler. 2012. Exploring topic coherence over many models and many topics. In *Conference on Empirical Methods on Natural Language Processing and Computational Natural Language Learning (EMNLP-CoNLL)*, pages 952–961.

Hanna M. Wallach, Iain Murray, Ruslan Salakhutdinov, and David Mimno. 2009. Evaluation methods for topic models. In *International Conference on Machine Learning (ICML)*, pages 1105–1112.

Roger Wimmer and Joseph Dominick. 2010. *Mass Media Research: An Introduction*. Cengage Learning.