Computer-Assisted Content Analysis: Topic Models for Exploring Multiple Subjective Interpretations

Jason Chuang[*]
Allen Institute for Artificial Intelligence
Seattle, Washington
jason@chuang.ca

John D. Wilkerson[†]
Political Science
University of Washington
jwilker@u.washington.edu

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

Dustin Tingley[†]
Government
Harvard University
dtingley@gov.harvard.edu

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

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

Forough Poursabzi-Sangdeh[†]
Computer Science
University of Colorado, Boulder
forough.poursabzi@gmail.com

Justin Grimmer[†]
Political Science
Stanford University
jgrimmer@stanford.edu

Leah Findlater[†]
College of Information Studies
University of Maryland, College Park
leahkf@umd.edu

Jordan Boyd-Graber[†]
Computer Science
University of Colorado, Boulder
jbg@boydgraber.org

Jeffrey Heer
Computer Science and Engineering
University of Washington
jheer@uw.edu

Abstract

Content analysis, a labor-intensive but widely-applied research method, is increasingly being supplemented by computational techniques such as statistical topic modeling. However, while the discourse on content analysis centers heavily on reproducibility, computer scientists often focus more on increasing the scale of analysis and less on establishing the reliability of analysis results. The gap between user needs and available tools leads to justified skepticism, and limits the adoption and effective use of computational approaches. We argue that enabling *human-in-the-loop* machine learning requires establishing users' trust in computer-assisted analysis. To this aim, we introduce our ongoing work on analysis tools for interactively exploring the space of available topic models. To aid tool development, we propose two studies to examine how a computer-aided workflow affects the uncovered codes, and how machine-generated codes impact analysis outcome. We present our prototypes and findings currently under submission.

1 Introduction

Content analysis — the examination and systematic categorization of written texts [1] — is a fundamental and widely-applied research method in the humanities and social sciences, as well as engineering fields such as computer-supported cooperative work. Researchers codify the political agendas of U.S. legislators to predict their voting patterns [18], analyze open-ended survey responses

[*]Research done while at Computer Science and Engineering, University of Washington
[†]Listed in reverse alphabetical order

to assess how experimental conditions alter participants' views on immigration [41, 43], and measure the effects of N.I.H. funding on active biomedical research areas [46]. About one third of all articles published in major mass communication journals employ content analysis [47].

Initial reading and coding, two labor-intensive steps in the analysis process, are increasingly replaced [18, 32, 37, 41, 46] by automated techniques such as statistical topic modeling, to encompass growing text corpora that can exceed a billion documents at times. However, while discourse about content analysis [27, 29, 31] overwhelmingly centers around the generalizability and reproducibility of a coding scheme, computer scientists tend to focus more on increasing the scale of analysis and less on coding reliability.

Many computational methods and tools falsely assume all machine-generated coding schemes are valid and useful, but findings from applying a single model to analysis often cannot be sustained under scrutiny [11]. Model output can vary greatly, even among multiple runs of an identical topic modeling algorithm [42]. Minor changes in corpus pre-processing are known to cause replication failures [16], but their effects are rarely documented. This disconnect between user needs and available validation tools leads to justified skepticism [21] and hampers continued adoption [20].

We argue that the true strength of computer-assisted content analysis should be enabling users to **explore multiple subjective interpretations** about their source documents. In order to support real-world deployment of and engage users in *human-in-the-loop* machine learning, we must first establish users' trust in the techniques.

To achieve these aims, we propose the following three steps to develop effective text analysis tools. These steps represent both work currently under submission as well as ongoing research. First, we are conducting human subject experiments to determine how a **computer-aided coding process** affects coding quality, compared to manual approaches. Second, we are measuring the **reliability of computer-generated codes**, to identify key factors that contribute to model variations. Finally, we are developing **interactive visual analysis tools** to support three classes of tasks: (1) assess coding reliability by exposing model sensitivity, (2) construct high-quality coding schemes by exploring the space of available models and through interactive topic modeling, and (3) support reasoning by mapping the uncovered codes onto phenomena of interest through an end-to-end system. Achieving these goals will require interdisciplinary efforts from the machine learning, human-computer interaction, and social science communities. We present our current prototypes and early findings, and welcome feedback from participants in this workshop.

2 Background

2.1 Manual Codification and Scalability Issues

Content analysis is frequently applied in the humanities and social sciences [23], journalism [40], psychology [6], communication [27], human-computer interaction [30], and computer-supported collaborative work [33]. Wimmer et al. [47] report that one third of all articles published in major mass communication journals employ quantitative content analysis. However, the first step in extracting information from textual data — manually reading the source documents and codifying notable concepts — is labor intensive and time consuming; its cost often prohibits the application of text analysis at scale.

Researchers often respond to time and resource limitations by aggressively subsampling their source documents. For example, Pew Research Center's Journalism Project produces the News Coverage Index [36] to measure the quality of news reporting in the United States. Intended to track all 1,450 newspapers published across the country, their purely manual approach only covers the front page of selected newspapers, resulting in the examination of 20 stories per day. While appropriate sampling strategies can lead to meaningful studies, researchers stand to lose rich details in their data when their attention is limited to only a minuscule fraction of the available texts.

Meticulous manual codification can produce high-quality data, but such studies are few and far between. The Guardian and the London School of Economics [34, 35] examined the causes of mass rioting and looting that hit England in August 2011. By manually codifying 2.5 million tweets, the authors were able to track personal communication at a national level, and identify how misinformation spread and was rebuffed. The initial coding pass, for tracking this large-scale social phenomenon, took the team of social scientists three months.

2.2 Statistical Topic Models

An active area of machine learning research, *statistical topic modeling* refers to latent Dirichlet allocation (LDA) [5] and its variants [2, 3, 24, 38, 43]. When applied to a document collection, topic models uncover weighted groups of words that frequently co-occur in the same documents. Chang et al. [8] find that users presented with such word groupings, or *latent topics*, recognize them as semantically meaningful concepts. Though some word groupings can also appear nonsensical, Griffiths et al. [17] find that LDA outperforms previous automatic techniques for extracting semantic dimensions from unstructured text.

Critical of manual approaches that "*make several restrictive assumptions or [are] prohibitively costly*," Quinn et al. [37] discuss the use of topic models to enable large-scale text analysis, by extracting initial codes from texts and utilizing latent topics to facilitate human deep reading.

Topic models have enabled groundbreaking massive studies in numerous fields. In the social sciences, McFarland et al. [32] tracked language use across 1 million Ph.D. dissertations, and mapped the shifts in research activities over 30 years in the United States. In biomedical science, Talley et al. [46] extracted biomedical research topics from 400,000 N.I.H. grant proposals, and studied the effects of funding on research activities. In political science, Grimmer et al. [18] examined 100,000 press releases to identify the political agendas of U.S. legislators, and predicted their voting patterns.

While this initial uptake of topic models is encouraging, an overemphasis on scalability among computer scientists and the use of a single model for analysis invite skepticism and hamper continued adoption. The machine learning and visualization communities have produced tools such as Termite [12] and the Topic Model Visualization Engine [7] aimed at making topic models accessible to non-technical users, but the output of these tools is almost always presented without any measure of uncertainty. In many cases, findings from these one-off modeling efforts cannot be sustained under scrutiny [11]. Without the capability to establish reproducibility and generalizability, two cornerstones of academic research, these tools are of little value to researchers and practitioners.

Recognizing the potential of computational approaches but concerned by the growing disconnect with user needs, Guiliano et al. [21] voiced in the call for their 2011 Workshop on Topic Modeling for Humanities Research: "*[T]he most promising work in topic modeling is being done not by humanists exploring literary or historical corpora but instead by scholars working in natural language processing and information retrieval*."

2.3 Coding Reliability

Coding reliability is critical to content analysis, and considered fundamental to reproducible research. When researchers devise a coding scheme, they must clearly articulate the definition of their codes in such a way that any human following the coding procedure will consistently apply the given codes to all documents in a corpus.

The proper application of reliability measures is heavily discussed and debated in the literature. While inter-rater agreement is one measure of coding reliability, other statistics [13, 26, 28] can be more suitable depending on the underlying data types, acceptable levels of agreement, and other factors. For example, Lombard et al. [31] analyzed and critiqued coding reliability in 200 content analysis studies in mass communication. Krippendorff [29] responded by providing additional guidelines, standards, and conditions on the use of seven popular reliability measures.

However, computational tools rarely account for reliability in their design. Statistical topic models eliminate many issues associated with human coding, but also introduce new factors that affect the resulting codes. While practitioners are aware of fluctuations in the model output, we find very few comprehensive studies on model sensitivity. What are the key factors that affect the stability of a topic model? How might variations in the model output change the outcome of subsequent analyses?

Writing for a special issue of the Journal of Digital Humanities in 2012, Schmidt [44] summarized the view amongst digital humanists, a group of early adopters of topic models, on the experience of working with these 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. . . . [E]xcitement about the use of topic models for discovery needs to be tempered with skepticism about how often the unexpected juxtapositions LDA creates will be helpful, and how often merely surprising.*"

2.4 Alternative Coding Schemes

Examining multiple interpretations of a corpus — and resolving their differences — is another critical step toward establishing the quality and external ecological validity of a chosen coding scheme.

Speaking at the Digital Humanities Conference in 2014 on designing visualizations to support text analysis, Klein [15, 25] emphasized the importance of allowing users to "*recombine preliminary analysis to test theories and develop arguments*" so that they may "*ask what about the data is exposed, and what remains obscured from view*." Later in his invited article, Schmidt argued that computer-aided text analysis should incorporate competing models or "*humanists are better off applying zero computer programs.*"

In a comprehensive survey of automatic content analysis methods, Grimmer et al. [20] highlight the need to validate models through close reading and model comparison, in order to ensure they are conceptually valid and useful. In the same survey, however, the authors also find that available software often "*simply provide the researcher with output*" without providing the capability to validate whether the output is optimal. In our own work on the use of topic models for text analysis, the lack of support for model comparisons is also repeatedly identified as a bottleneck to analysis, and the subsequent communication and publication of analysis findings.

Topic modeling has greatly changed the cost of text codification. Whereas manually constructing a single scheme used to take weeks or months, hundreds of models can now be generated in parallel. With computer-aided content analysis, we expect an even greater need to compare the quality of alternative coding schemes — in addition to validating code reliability within a scheme.

As an additional note, while supervised machine learning techniques [45] have been applied to content analysis, they represent the application of a user's pre-defined coding scheme to a text corpus, which is different from the task of defining a coding scheme and assessing its quality. Such techniques can be useful if the analysis categories are known a priori, but do not directly help with the exploration of alternative coding schemes.

2.5 Model Sensitivity and Model Designs

Existing literature suggests that topic model output can vary due to at least the following factors — both by design and from perturbations that stem from user or algorithm actions.

Intra-Model Sensitivity. Roberts et al. [42] examine how multiple runs of an identical topic modeling algorithm may converge to different solutions, due to the underlying optimization being non-convex. They demonstrate that a topic model's output across multiple runs may exhibit multi-modal distributions. Their paper is rare in that the authors not only characterize the uncertainty around a model's output, but also establish that the spread of topics can lead to contradictory results in subsequent analyses. Training a topic model also requires users to select a set of parameters and hyper-parameters. Chuang et al. [9] examine LDA trained using over 10,000 settings. When compared to expert-curated topics, the authors identify parameterizations that produce fused concepts or duplications. Models with poor settings uncover significantly fewer expert concepts.

Pre-Processing of Text Corpus. Researchers and practitioners frequently pre-process a text corpus by removing common or rare terms prior to topic modeling. However, Fokkens et al. [16] find widespread reproducibility failures in natural language processing, when they replicate — and fail to reproduce — the results reported in numerous papers that were evaluated on two standard experiments. The authors find that minor decisions in the modeling process can impact evaluation results. While the authors' chosen tasks do not include statistical topic modeling, their five identified factors are still relevant, including corpus *pre-processing*. Phrases are sometimes fed into a topic model, creating differences in the *system output*. Other factors include *experimental setup*, software and dataset *versioning*, and the *treatment of ties*.

Post-Processing of Topic Models. Chuang et al. [11] point out that the output of topic models frequently needs to be manually validated and refined prior to subsequent analysis [22, 46]. This step introduces two additional factors. First, validation can introduce significant changes; experts have rejected up to two thirds of the machine-generated topics [22]. Second, while topic modeling is often said to *replace* the initial coding pass, in reality, validation *introduces a new step* into the computer-aided coding process; the validation step has no natural correspondence in manual codification.

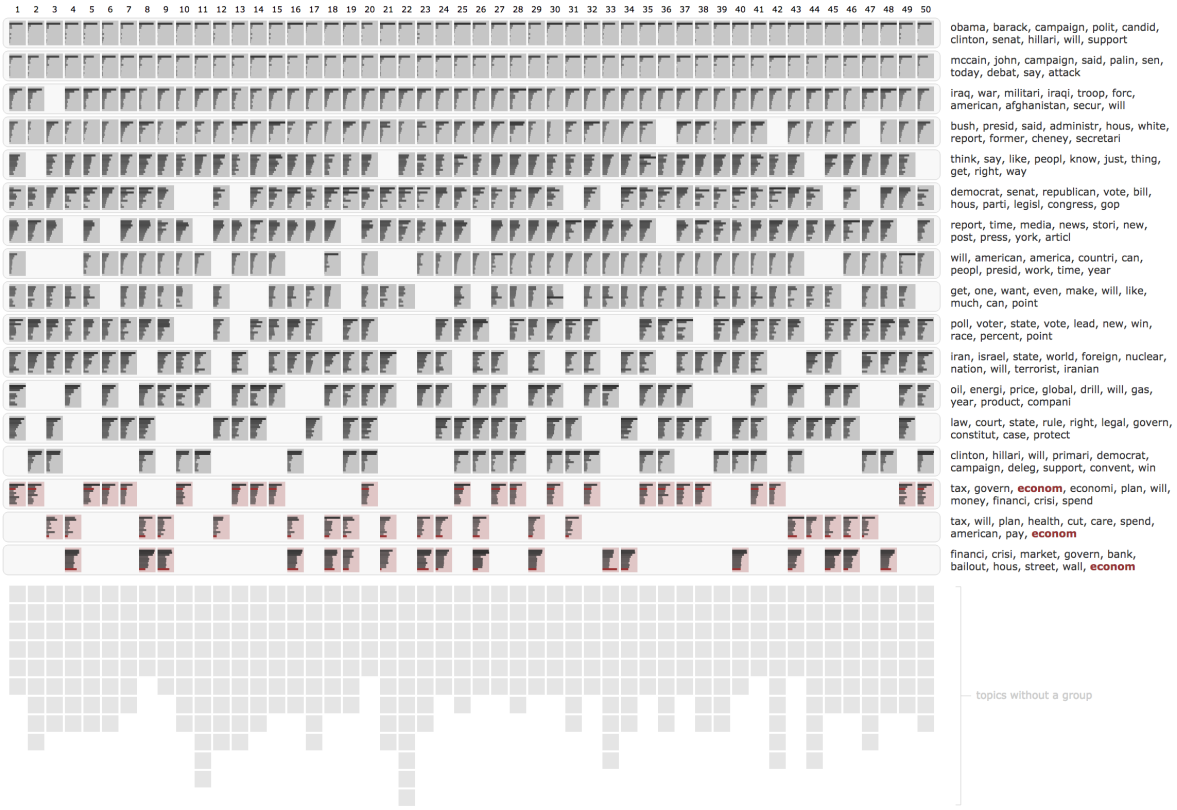

Figure 1: **Coding Reliability**: This TopicCheck visualization shows the political topics extracted from 13,250 political blogs [14] collected in 2008 and generated by 50 topic model runs. Each rectangle in the visualization represents a latent topic. The horizontal bars within each rectangle represent the frequencies of top words. Hovering over a word reveals other occurrences of the same word (highlighted in red). Each column represents topics belonging to a single topic model. Each row represents a group of similar topics from different models. The list of words on the right represents the top word belonging to each group of topics. Words are stemmed, so that different forms of a word (e.g., *military, militaries*) are collapsed onto the same root form of the word (e.g., *militari*). A live version of this visualization is at `http://content-analysis.info/nips2014`

Auxiliary Measures. Users often depend on auxiliary measures — such as cosine, information divergence, top-ranked words, or top-ranked documents — to calculate the similarity or distance between topics, in order to generate aggregated statistics for their analyses. Chuang et al. [9] examine three classes of distance measures and find that their scores deviate from how human raters assign topical similarity. Roberts et al. [42] examine the correlations between various similarity, top-word, and top-document measures, and find that their accuracies depend on the distribution of probability mass of individual latent topics.

Model Designs. Machine learning and social science researchers have designed numerous topic models to account for temporal variations [3], hierarchical organization [2], correlated topics [4], user-defined labels [38], partially known labels [39], and structures inferred from known metadata [43]. Many of these models also offer additional parameters such as changing the number of topics to uncover. Grimmer et al. [19] describe a theoretically infinite space of potential models, where each one represents a possible conceptualization of the corpus. Indeed, Hu et al. [24]'s work on interactive topic modeling allows users to specify any number of word-level constraints and iteratively craft new topic models throughout their analysis.

We divide all possible model variations into two types. We refer to unexpected model variations as **model sensitivity**, and the active intentional construction of topic models as **model design**. We consider sensitivity as analogous to coding reliability, where the models are expected to produce identical reproducible results but differ in practice. We consider model design as analogous to

generating alternative coding schemes, where users are actively seeking an optimal solution. As users have very different intents for the two tasks, we propose two classes of tools to help users assess model sensitivity and accelerate model design, respectively. To facilitate tool development, however, we first investigate how machine-generated codes differ from manually-crafted ones, and how the above factors affect model output.

3 User Experiments and Studies

3.1 Computer-Assisted Coding Process

Manual validation alters the coding workflow; this post-processing step may introduce human subjectivity. In addition, a common measure of scalability is reporting on the number of documents analyzed, but we believe the scale of analysis should refer to the scope of generated codes, not merely the input bandwidth to an algorithm.

We formulate two canonical content analysis workflows: a baseline manual approach in which people read and annotate a corpus vs. a computer-aided approach in which a topic model generates candidate concepts that are then manually verified. We are quantifying changes to the uncovered coding schemes in the two conditions.

Selection Bias. Suppose people are predisposed to pick one coding scheme over another when they manually examine a text corpus. Does seeing topic model output help the subjects select more representative codes, or does such subjectivity carry over to the same degree in the aided condition? We hypothesize the former, and conjecture that removing selection bias requires exposing people to multiple interpretations and asking the users to actively resolve their differences.

Coverage. Manual reading of documents can yield significant insights that statistical modeling may not capture. Does the aided condition — when we account for codes rejected in the validation step — actually produce better and/or more comprehensive coding schemes? We hypothesize that attaining high topical coverage requires well-tuned models, otherwise manual coding prevails. We believe that the utility of topic modeling will increase significantly when users, through the exploration of the model space, can identify or iteratively craft high-quality models suited for their analyses.

3.2 Reliability of Machine-Generated Codes

We are replicating previously published work, to determine the relative significance of the aforementioned impact factors (§2.5) on computer-assisted content analysis. We are examining not just how the latent topics differ but also how the differences affect analysis outcome.

Relative Importance of Impact Factors. Grimmer et al. [18] show that their hierarchical topic model, trained on press releases of U.S. legislators, can be used to predict their voting patterns. We plan to alter their model designs, as well as perturb the list of known factors in the modeling process to determine how the combination of factors affect the final prediction accuracy.

Reliability Measures. We plan to correlate observed differences with established reliability measures, the majority of which are originally designed for human coders. Our study results may help researchers devise novel reliability measures for machine-generated codes, and contribute to social sciences research. Our findings may also inform users of good modeling practices.

4 Interactive Visual Analysis Tools

4.1 TopicCheck: Coding Reliability

We develop TopicCheck, an interactive visualization to help users assess model sensitivity, in scenarios where the users expect a set of models to produce identical output. We first devise a constrained agglomerative clustering algorithm to identify groupings of similar topics, by aligning up to one topic from every model. We then design a corresponding visualization to highlight the deviations among the aligned topics. Figure 1 shows political topics extracted from 13,250 political blogs [14] collected in 2008 and generated by 50 runes of a structural topic model [43].

In a seminal article in mass communication, Krippendorff [29] establishes four recommendations to safeguard the reproducibility of a coding process: (a) the use of multiple coders for analysis; (b) the

selection of a suitable inter-rater agreement statistic; (c) setting an appropriate level of agreement; and (d) allowing users to inspect individual distinctions in the data.

In our paper currently under submission, we adapt these rules originally designed for human coders for a topic model-backed coding process. We demonstrate that available user interactions in TopicCheck support the following four actions: (a) the use of multiple topic models for analysis; (b) modular software design for exploring multiple topical similarity measures; (c) setting an appropriate topical grouping criteria; and (d) allowing users to inspect model output at all levels of details.

We present three case studies in our paper, to demonstrate the impact on real-world social science research when statistical topic modeling is paired with suitable validation tools.

We first provide an in-depth view on the multi-modality of topic models [42] to help political science methods research. As shown in Figure 1, we observe that the top two topics (about Barack Obama and John McCain respectively) are consistently uncovered across all runs. The third topical group (about the Iraqi and Afghani Wars as identified by the more diverse vocabulary of *iraq, afghanistan, war, military(-ies), troop(s), force(s)*) is also consistently generated by 49 of the 50 runs. However, 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 (*tax(es), government spending, economy(-ies), financial crisis*), 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 a second case study, we document how rare word removal can drastically alter topical compositions in an analysis of ten years of news reports. Finally, by examining 200,000 hours of broadcast news, we find that identifying *consistent topics across multiple models* may be more critical in capturing diverse political view points than enforcing *topical coherence within a single model*.

4.2 Interactive Codification User Interface

We are designing tools to support the exploration of alternative coding schemes, by exposing multiple model output and enabling interactive construction of topic models. In this scenario, users may wish to perform additional comparisons such as identifying how a broad concept divides into more specialized subtopics based on (mis)alignment measures [9]. Users may also wish to use other statistical properties such as topical correlations [42] to evaluate coding quality.

To account for multi-modality, we expect TopicCheck be used as a subroutine within the model design tools, so that when comparing multiple models, users see the "average" aligned output of a model rather than a single noisy instance.

Beyond exploring pre-computed models, we plan to incorporate interactive topic modeling [24], so that users may create custom user-defined topic models within the visualization. We envision our tools allowing users to inspect how an updated model differ from all previous models they've constructed. We previously documented [10] that users frequently request the ability to track the history of their model designs for two reasons. First, such a "lab notebook" can aid their own sensemaking after they've constructed dozens or hundreds of models. Second, a detailed notebook allows their work to be scrutinized in greater details, and may help better communicate their modeling decisions within the context of other alternative designs.

4.3 Mapping Uncovered Codes onto Phenomena of Interest

Finally, we are also developing tools to map the uncovered latent topics directly onto interpretable domain-specific variables that may be of interest to analysts, so that the users can draw on their background knowledge to validate and refine the uncovered codes — rather than assessing coding quality in isolation of their analysis task.

Such an end-to-end system can help users more rapidly iterate between codification, coding scheme selection, and their eventual text analysis. We are applying supervised machine learning techniques (i.e., support vector machines) to connect the latent topics with observable variables (e.g., time) and relevant artifacts (e.g., documents). Our system differs from standard classification tasks, in that our goal is to optimize task-specific coding quality, rather than classifier accuracy.

5 Conclusion

We highlight the critical roles of establishing coding reliability and validating coding quality, in order to effectively apply statistical topic models to accelerate content analysis. We are conducting user studies to improve our understanding of machine-aided coding. We are developing tools that will not only aid the task of text codification but also improve user trust in machine learning techniques. We anticipate that these tools will bridge the gap between available computational methods and the needs of scholars seeking reproducible content analysis — and help bring greater acceptance to computer-assisted content analysis approaches.

References

[1] Bernard Berelson. *Content analysis in communication research*. Free Press, 1952.

[2] David M. Blei, Thomas L. Griffiths, Michael I. Jordan, and Joshua B. Tenenbaum. Hierarchical topic models and the nested chinese restaurant process. In *NIPS*, 2004.

[3] David M. Blei and John D. Lafferty. Dynamic topic models. In *ICML*, pages 113–120, 2006.

[4] David M. Blei and John D. Lafferty. A correlated topic model of science. *The Annals of Applied Statistics*, 1(1):17–35, 2007.

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

[6] Virginia Braun and Victoria Clarke. Using thematic analysis in psychology. *Qualitative Research in Psychology*, 3(2):77–101, 2006.

[7] Allison Chaney. Topic model visualization engine. `https://code.google.com/p/tmve`, 2013.

[8] Jonathan Chang, Jordan Boyd-Graber, Chong Wang, Sean Gerrish, and David M. Blei. Reading tea leaves: How humans interpret topic models. In *NIPS*, pages 288–296, 2009.

[9] Jason Chuang, Sonal Gupta, Christopher D. Manning, and Jeffrey Heer. Topic model diagnostics: Assessing domain relevance via topical alignment. In *ICML*, 2013.

[10] Jason Chuang, Yuening Hu, Ashley Jin, John D Wilkerson, Daniel A McFarland, Christopher D Manning, and Jeffrey Heer. Document exploration with topic modeling: Designing interactive visualizations to support effective analysis workflows. In *NIPS Workshop on Topic Models*, 2013.

[11] Jason Chuang, Christopher D. Manning, and Jeffrey Heer. Interpretation and trust: Designing model-driven visualizations for text analysis. In *CHI*, pages 443–452, 2012.

[12] Jason Chuang, Christopher D. Manning, and Jeffrey Heer. Termite: Visualization techniques for assessing textual topic models. In *Advanced Visual Interfaces*, 2012.

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

[14] J. Eisenstein and E. Xing. *The CMU 2008 Political Blog Corpus*. Carnegie Mellon University, 2010.

[15] Jacob Eisenstein, Iris Sun, and Lauren F. Klein. Exploratory thematic analysis for historical newspaper archives. In *Digital Humanities*, 2014.

[16] Antske Fokkens, Marieke van Erp, Marten Postma, Ted Pedersen, Piek Vossen, and Nuno Freire. Offspring from reproduction problems: What replication failure teaches us. In *ACL*, pages 1691–1701, 2013.

[17] Thomas L. Griffiths and Mark Steyvers. Prediction and semantic association. In *NIPS*, 2002.

[18] Justin Grimmer. A bayesian hierarchical topic model for political texts: Measuring expressed agendas in senate press releases. *Political Analysis*, 18(1):1–35, 2010.

[19] Justin Grimmer and Gary King. General purpose computer-assisted clustering and conceptualization. *Proceedings of the National Academy of Sciences*, 108(7):2643–2650, 2011.

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

[21] Jennifer Guiliano, Sayan Bhattacharyya, Travis Brown, and Kirsten Keister. Topic modeling for humanities research. `http://mith.umd.edu/topicmodeling`, 2011.

[22] David Hall, Daniel Jurafsky, and Christopher D Manning. Studying the history of ideas using topic models. In *EMNLP*, pages 363–371, 2008.

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

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

[25] Lauren F. Klein. Exploratory thematic analysis for historical newspaper archives. `http://lklein.com/2014/07/talk-at-digital-humanities-2014/`, 2014.

[26] Klaus Krippendorff. Bivariate agreement coefficients for reliability data. In E. R. Borgatta and G. W. Bohrnstedt, editors, *Sociological methodology*, pages 139–150. 1970.

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

[28] Klaus Krippendorff. *Content analysis: An introduction to its methodology*. Sage, 2 edition, 2004.

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

[30] Jonathan Lazar, Jinjuan Heidi Feng, and Harry Hochheiser. *Research Methods in Human-Computer Interaction*. John Wiley & Sons, 2010.

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

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

[33] David Newman, Brian Webb, and Clive Cochrane. A content analysis method to measure critical thinking in face-to-face and computer supported group learning. *Interpersonal Computing and Technology Journal*, 3(2):56–77, 1995.

[34] Rob Procter, Farida Vis, Alex Voss, Marta Cantijoch, Yana Manykhina, Mike Thelwall, Rachel Gibson, Andrew Hudson-Smith, and Steven Gray. Reading the riots study to examine causes and effects of August unrest. `http://www.theguardian.com/uk/2011/sep/05/reading-riots-study-guardian-lse`, 2011.

[35] Rob Procter, Farida Vis, Alex Voss, Marta Cantijoch, Yana Manykhina, Mike Thelwall, Rachel Gibson, Andrew Hudson-Smith, and Steven Gray. Riot rumours: How misinformation spread on Twitter during a time of crisis. `http://www.guardian.co.uk/uk/interactive/2011/dec/07/london-riots-twitter`, 2011.

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

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

[38] Daniel Ramage, David Hall, Ramesh Nallapati, and Christopher D. Manning. Labeled LDA: A supervised topic model for credit attribution in multi-label corpora. In *EMNLP*, pages 248–256, 2009.

[39] Daniel Ramage, Christopher D. Manning, and Susan T. Dumais. Partially labeled topic models for interpretable text mining. In *KDD*, pages 457–465, 2011.

[40] Daniel Riffe and Alan Freitag. A content analysis of content analyses: Twenty-five years of journalism quarterly. *Journalism and Mass Communication Quarterly*, 74(3):515–524, 1997.

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

[42] Margaret E. Roberts, Brandon M. Stewart, and Dustin Tingley. 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.

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

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

[45] Burr Settles. Closing the loop: Fast, interactive semi-supervised annotation with queries on features and instances. In *EMNLP*, pages 1467–1478, 2011.

[46] Edmund M. Talley, David Newman, David Mimno, Bruce W. Herr, Hanna M. Wallach, Gully A. P. C. Burns, A. G. Miriam Leenders, and Andrew McCallum. Database of NIH grants using machine-learned categories and graphical clustering. *Nature Methods*, 8(6):443–444, 2011.

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