Computational Text Analysis for Social Science: Model Assumptions and Complexity

Brendan O'Connor* **David Bamman†** **Noah A. Smith†***
*Machine Learning Department
†Language Technologies Institute
Carnegie Mellon University
{brenocon,dbamman,nasmith}@cs.cmu.edu

December 8, 2011
Second Workshop on Computational Social Science
and Wisdom of the Crowds (NIPS 2011)

Abstract

Across many disciplines, interest is increasing in the use of computational text analysis in the service of social science questions. We survey the spectrum of current methods, which lie on two dimensions: (1) computational and statistical model complexity; and (2) domain assumptions. This comparative perspective suggests directions of research to better align new methods with the goals of social scientists.

1 Use cases for computational text analysis in the social sciences

The use of computational methods to explore research questions in the social sciences and humanities has boomed over the past several years, as the volume of data capturing human communication (including text, audio, video, etc.) has risen to match the ambitious goal of understanding the behaviors of people and society [1]. Automated content analysis of text, which draws on techniques developed in natural language processing, information retrieval, text mining, and machine learning, should be properly understood as a class of quantitative social science methodologies. Employed techniques range from simple analysis of comparative word frequencies to more complex hierarchical admixture models. As this nascent field grows, it is important to clearly present and characterize the assumptions of techniques currently in use, so that new practitioners can be better informed as to the range of available models.

To illustrate the breadth of current applications, we list a sampling of substantive questions and studies that have developed or applied computational text analysis to address them.

- Political Science: How do U.S. Senate speeches reflect agendas and attention? How are Senate institutions changing [27]? What are the agendas expressed in Senators' press releases [28]? Do U.S. Supreme Court oral arguments predict justices' voting behavior [29]? Does social media reflect public political opinion, or forecast elections [12, 30]? What determines international conflict and cooperation [31, 32, 33]? How much did racial attitudes affect voting in the 2008 U.S. presidential election [34]?
- Economics: How does sentiment in the media affect the stock market [2, 3]? Does sentiment in social media associate with stocks [4, 5, 6]? Do a company's SEC filings predict aspects of stock performance [7, 8]? What determines a customer's trust in an online merchant [9]? How can we measure macroeconomic variables with search queries and social media text [10, 11, 12]? How can we forecast consumer demand for movies [13, 14]?
- Psychology: How does a person's mental and affective state manifest in their language [15]? Are diurnal and seasonal mood cycles cross-cultural [16]?

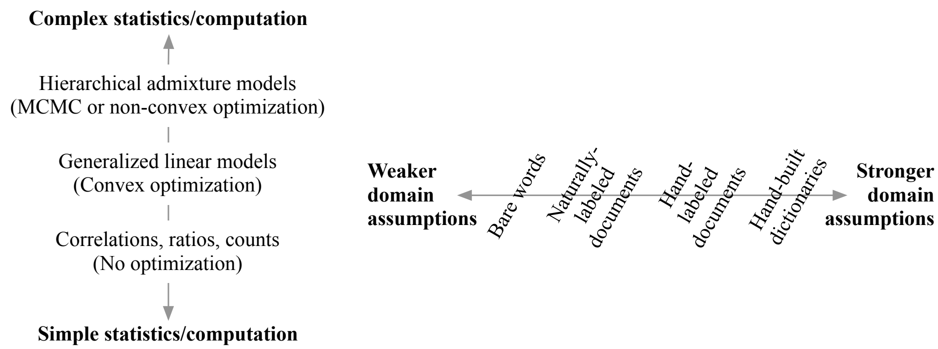

Figure 1: Schematic of model complexity versus domain assumptions for various computational text analysis methods. Statistical models are listed with their respective inference/training algorithms; computational expense increases with model expressiveness.

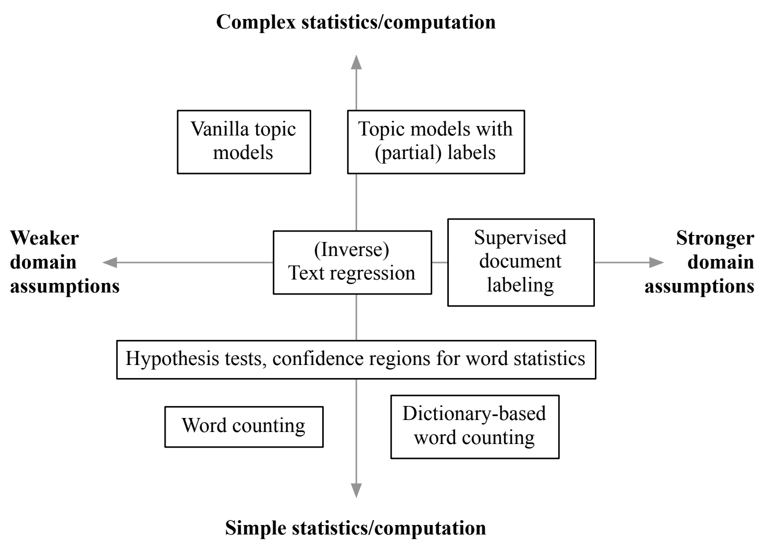

Figure 2: Typical methods used in computational text analysis. Compare to Table 1 in [27].

- Scientometrics/Bibliometrics: What are influential topics within a scientific community? What determines a paper's citations [35, 36, 37, 38]?
- Sociolinguistics: What are the geographic determinants of lexical variation in social media [39]? Demographic determinants [12, 40]?
- Public Health: How can search queries and social media help measure levels of the flu and other public health issues [41, 42, 43]?
- History: How did modern English legal institutions develop over the 17th to 20th centuries [17]? When did concepts of religion, secularism, and social institutions develop over two millennia of Latin literature [18]? What can topical labels in Diderot's 18th Encyclopédie reveal about contemporary ways of thought [19]? Who were the true authors of a particular piece of historical text [20]? The last deserves mention as a classic (1964) work that analyzed pseudonymous Federalist papers and answered long-standing questions about their authorship—one of the earliest instances of automated, statistical *stylometry* and automated text analysis for social science in general.
- Literature: What are the textual allusions in Classical Latin poetry [21] and the synoptic gospels [22]? How do demographic determinants of fictional characters affect their language use [23]? Who is the true author of a work of literature [24]? Roberto Busa's work in digitizing and lemmatizing the complete works of Thomas Aquinas, begun in 1949, also deserves mention as one of the earliest efforts at creating a machine-readable *annotated* corpus [25, 26].

This list is incomplete, both in the works cited and in the range of areas that have used or could use these methods. These techniques are still in their infancy: while several of the works above thoroughly address substantive questions (and, in a few cases, there exist lines of work published in high-quality social science journals), most tend to focus on developing new methodologies or data sources. There are also more exploratory analyses not aimed at specific research questions [44, 45]. In most cases, automated text analysis functions as a tool for *discovery* and *measurement* of prevalent attitudes, concepts, or events in textual data.

2 Classes of methods

In most cases, the analysis is restricted to the frequencies of words or short phrases (n-grams) in documents and corpora.[1] Even so, there is still a rich variety of methods, with two important axes of variation: statistical model assumptions and domain assumptions (Figure 1).

Domain assumptions refer to how much knowledge of the substantive issue in question is used in the analysis. A purely exploratory, "bare words" analysis only considers the words of documents; for example, examining the most common words in a corpus, or latent topics extracted automatically from them. Next, non-textual metadata about the documents is almost always used; for example, tracking word frequencies by year of book publication [45], or spatial location of a microblogger [39]. We call these *naturally-labeled* documents; typically the labels take the form of continuous, discrete, or ordinal variables associated with documents or segments of them. In contrast, *manually-labeled* documents may be created in order to better understand particular quantities of interest; for example, annotating news articles with information about the events they describe [32]. Creating and evaluating the codebook (that carefully defines the semantics of the annotations a coder will produce) can be a laborious and iterative process, but is essential to understand the problem and create a psychometrically reliable coding standard [46]. Finally, another source of domain information can take the form of *dictionaries*: lists of terms of interest to the analysis, such as names, emotional words, topic-specific keywords, etc. Ideally, they may be custom-built for the problem (if done manually, a labor-intensive task similar to coding documents), or they may be reused or adapted from already-existing dictionaries (e.g., Freebase [47] for names or LIWC [15] for affect, though see [48]'s critical comments on the naïve use of affect dictionaries). Useful information can be revealed with just a handful of terms; for example, [34] analyzes the Google search query frequencies of one highly charged racial epithet as a proxy for racial attitudes.

The second dimension is **computational and statistical complexity**.[2] The simplest techniques count words. A typical analysis is to compare word frequencies between groups (e.g., listing the most common words per speaker in a debate). Note that any type of comparison requires some form of natural labels for text; generally, metadata is what links text to interesting substantive questions. In this case, the metadata is speaker identity, but time, social group membership, and others have also been considered—as in the common analysis of plotting a word's frequency over time.

Frequency ratios and correlations with response variables can be seen as parameters or hypothesis tests for simple two-variable models between text frequencies and response/metadata variables; in the case of a categorical metadata variable, words' conditional probabilities $p(x \mid y)$ correspond to parameters of the naïve Bayes model of text. A hallmark of this class is that they are computationally straightforward to calculate; typically they involve a single pass through the data to compute counts, sums, and other quantities.

Another popular set of techniques sits on the other side of the computational spectrum: hierarchical admixture models, specifically LDA-style topic models [49]. Here, documents undergo dimension reduction by being modeled as mixtures of multinomials, where each component is a distribution over words—called a topic. The output of a topic model can be used for exploratory analysis, or post-hoc compared across observed variables. With some work, these models can also be usefully customized for a variety of applications; typically, an important change is to incorporate the natural labels and structure of the domain. (Models that can incorporate reasonably generic types of labels, and in substantially different ways, include SLDA, DMR, and PLDA [50, 51, 36].)

[1] A few interesting exceptions: [18, 9, 31].

[2] We use the term "complexity" informally, not intending to imply any of its technical senses.

Another class of techniques is generalized linear models, and specifically regularized linear and logistic regression [52]. In *text regression*, the response variable is modeled as a conditional distribution given a linear combination of text features, $p(y \mid x)$. This model has often been used for the task of *text categorization*, to predict a document's category according to a training set of prespecified labeled documents; research has shown that linear models are state-of-the-art for this problem ([53] §14.7-8). A researcher can manually label documents and train a classifier to aid in the analysis of a large document collection; but these models can also be used for natural labels, to directly model a response variable of interest with text. We prefer the term *regression* for both discrete and continuous response variables, to emphasize these models' connections to the extensively developed statistical literature in GLMs and applied regression analysis [54, 55, 56].

An alternative is *inverse text regression*, where $p(x \mid y)$ is modeled as a multinomial logistic regression over the vocabulary, using the document labels [57] (or possibly latent variables [58]) as features. This direction of conditioning is more like naïve Bayes and (labeled) topic models in that it grounds out as multinomials over the vocabulary, but with linear parameterization of the multinomials, using additive effects instead of mixture memberships to select word probabilities.

3 Considerations

Which method to use completely depends on the goals and needs of the analysis: all three can be used for descriptive analysis and prediction. One consideration is the usual tradeoff between **simplicity** and **expressiveness**. Frequencies and correlations are easily computed and replicable; regressions require more computation, though often have unique solutions and off-the-shelf solvers; while topic models use fitting procedures that are more expensive (MCMC), or less flexible (variational inference), and may be less stable in that different runs can produce different results. This is part of the tradeoff of their greater expressive power. Regressions have the same level of expressiveness as word frequencies, but control for covariation through additive effects, where a word's coefficient explains the specific effect of that word when controlling for other words and other covariates. ([37] illustrates how this can make a difference for analysis.)

We should note that all the methods described in the previous section assign vectors of weights across the vocabulary, giving words associations to non-textual document-level variables, and are therefore fundamentally **interpretable**, because the researcher can inspect words' numeric weights. Word correlations and regressions associate words to observed document label variables, while topic models associate words to hidden topic variables. (Per-word association weights are individual correlations, regression coefficients, or conditional topic probabilities, respectively.) In all these cases, a way to summarize a particular document-level variable, then, is to look at the top-weighted words for that vector – e.g., the top 10 words with highest probability under a topic, or highest coefficient for a label class, or highest correlation/frequency. An analyst can then view the corpus through the lens of these top-words lists and their associated variables. This level of interpretability is a major advantage over black-box non-linear methods like kernel methods (e.g. kernelized SVMs) or neural networks, especially given that linear methods often have similar predictive performance.

A third consideration is what sort of the relationship between text and **observed variables** the researcher is interested in. If there are few observed variables, then topic models can still be used for purely exploratory analysis. However, since many of the substantive questions researchers are interested in typically involve conditioning on observed variables to make comparisons (whether the observed variables are natural or hand-labeled), it is useful to allow the model to tie relevant textual features to the variables in question.

For some problems, like analyzing Congressional floor speeches, topics correspond quite well to the substantive issues under consideration [27]. But for other problems, they can work less well. As one example, we have observed several cases where SLDA (an LDA variant that models a document-level variable through a GLM regression on topic proportions [50]) has similar [59] or worse [39, 57] predictive performance than regularized text regression, For the problem of predicting U.S. users' locations from their microblog text [39], we observed that Lasso regression selected a small number of words to have non-zero coefficients (e.g., "taco" to indicate the West Coast). We believe that in SLDA, the impact of these sparse cues was blunted from their incorporation into broad topics, since the model had to explain not just the response, but also the entirety of all the text. Sometimes the relationship between text and the document variable is better explained by individual words alone.

The extremes of individual word frequencies versus broad topic proportions are only two points in the space of possible text representations; it remains an interesting open question how to design models that can reliably abstract beyond individual words in service of social science analysis.

Acknowledgments

This research was supported by the NSF through grants IIS-0915187 and IIS-1054319 and by Google through the Worldly Knowledge Project at CMU. The authors thank the anonymous reviewers for helpful comments.

References

[1] David Lazer, Alex Pentland, Lada Adamic, Sinan Aral, Albert-Lszl Barabsi, Devon Brewer, Nicholas Christakis, Noshir Contractor, James Fowler, Myron Gutmann, Tony Jebara, Gary King, Michael Macy, Deb Roy, and Marshall Van Alstyne. Computational social science. *Science*, 323(5915):721 –723, February 2009.

[2] Paul C. Tetlock. Giving content to investor sentiment: The role of media in the stock market. *The Journal of Finance*, 62(3):11391168, 2007.

[3] Victor Lavrenko, Matt Schmill, Dawn Lawrie, Paul Ogilvie, David Jensen, and James Allan. Mining of concurrent text and time series. In *Proceedings of KDD Workshop on Text Mining*, pages 37—44, 2000.

[4] Eric Gilbert and Karrie Karahalios. Widespread worry and the stock market. In *Proceedings of the International Conference on Weblogs and Social Media*, 2010.

[5] Sanjiv R. Das and Mike Y. Chen. Yahoo! for amazon: Sentiment extraction from small talk on the web. *Management Science*, 53(9):1375–1388, September 2007.

[6] Johan Bollen, Huina Mao, and Xiao-Jun Zeng. Twitter mood predicts the stock market. *1010.3003*, October 2010.

[7] Shimon Kogan, Dimitry Levin, Bryan R. Routledge, Jacob S. Sagi, and Noah A. Smith. Predicting risk from financial reports with regression. In *Proceedings of Human Language Technologies: The 2009 Annual Conference of the North American Chapter of the Association for Computational Linguistics*, page 272280, 2009.

[8] Tim Loughran and Bill McDonald. When is a liability not a liability? Textual analysis, dictionaries, and 10-Ks. *Journal of Finance (forthcoming)*, 2011.

[9] Nikolay Archak, Anindya Ghose, and Panagiotis Ipeirotis. Deriving the pricing power of product features by mining consumer reviews. *Management Science*, page mnsc1110, 2011.

[10] Nikolaos Askitas and Klaus F. Zimmermann. Google econometrics and unemployment forecasting. *Applied Economics Quarterly*, 55(2):107–120, April 2009.

[11] Matthew E. Kahn and Matthew J. Kotchen. Environmental concern and the business cycle: The chilling effect of recession. http://www.nber.org/papers/w16241, July 2010.

[12] Brendan O'Connor, Ramnath Balasubramanyan, Bryan R. Routledge, and Noah A Smith. From tweets to polls: Linking text sentiment to public opinion time series. In *International AAAI Conference on Weblogs and Social Media, Washington, DC*, 2010.

[13] Sitaram Asur and Bernardo A. Huberman. Predicting the future with social media. *1003.5699*, March 2010.

[14] Mahesh Joshi, Dipanjan Das, Kevin Gimpel, and Noah A. Smith. Movie reviews and revenues: An experiment in text regression. In *Human Language Technologies: The 2010 Annual Conference of the North American Chapter of the Association for Computational Linguistics*, page 293296, 2010.

[15] Yla R. Tausczik and James W. Pennebaker. The psychological meaning of words: LIWC and computerized text analysis methods. *Journal of Language and Social Psychology*, 2009.

[16] Scott A. Golder and Michael W. Macy. Diurnal and seasonal mood vary with work, sleep, and daylength across diverse cultures. *Science*, 333:1878–1881, September 2011.

[17] Dan Cohen, Frederick Gibbs, Tim Hitchcock, Geoffrey Rockwell, Jorg Sander, Robert Shoemaker, Stefan Sinclair, Sean Takats, William J. Turkel, Cyril Briquet, Jamie McLaughlin, Milena Radzikowska, John Simpson, and Kirsten C. Uszkalo. Data mining with criminal intent. Final white paper, 2011.

[18] David Bamman and Gregory Crane. Measuring historical word sense variation. In *Proceeding of the 11th annual international ACM/IEEE joint conference on Digital libraries*, page 110, 2011.

[19] Russell Horton, Robert Morrissey, Mark Olsen, Glenn Roe, and Robert Voyer. Mining Eighteenth Century Ontologies: Machine Learning and Knowledge Classification in the Encyclopédie. *Digital Humanities Quarterly*, 3(2), 2009.

[20] Frederick Mosteller and David Wallace. *Inference and Disputed Authorship: The Federalist*. Addison-Wesley, Reading, 1964.

[21] David Bamman and Gregory Crane. The logic and discovery of textual allusion. In *Proceedings of the 2008 LREC Workshop on Language Technology for Cultural Heritage Data*, 2008.

[22] John Lee. A computational model of text reuse in ancient literary texts. In *Proceedings of the 45th Annual Meeting of the Association of Computational Linguistics*, pages 472–479, Prague, Czech Republic, June 2007. Association for Computational Linguistics.

[23] Shlomo Argamon, Charles Cooney, Russell Horton, Mark Olsen, Sterling Stein, and Robert Voyer. Gender, race, and nationality in black drama, 1950-2006: Mining differences in language use in authors and their characters. *Digital Humanities Quarterly*, 3(2), 2009.

[24] David I. Holmes. The evolution of stylometry in humanities scholarship. *Literary and Linguistic Computing*, 13(3):111–117, 1998.

[25] Roberto Busa. The annals of humanities computing: The index thomisticus. *Language Resources and Evaluation*, 14:83–90, 1980.

[26] Roberto Busa. *Index Thomisticus: sancti Thomae Aquinatis operum omnium indices et concordantiae, in quibus verborum omnium et singulorum formae et lemmata cum suis frequentiis et contextibus variis modis referuntur quaeque / consociata plurium opera atque electronico IBM automato usus digessit Robertus Busa SI*. Frommann-Holzboog, Stuttgart-Bad Cannstatt, 1974–1980.

[27] 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):209228, 2010.

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

[29] Ryan C. Black, Sarah A. Treul, Timothy R. Johnson, and Jerry Goldman. Emotions, oral arguments, and Supreme Court decision making. *The Journal of Politics*, 73(2):572–581, April 2011.

[30] Panagiotis T. Metaxas, Eni Mustafaraj, and Daniel Gayo-Avello. How (Not) to predict elections. Boston, MA, 2011.

[31] Philip A. Schrodt, Shannon G. Davis, and Judith L. Weddle. KEDS – a program for the machine coding of event data. *Social Science Computer Review*, 12(4):561 –587, December 1994.

[32] Gary King and Will Lowe. An automated information extraction tool for international conflict data with performance as good as human coders: A rare events evaluation design. *International Organization*, 57(3):617–642, July 2003.

[33] Stephen M. Shellman. Coding disaggregated intrastate conflict: machine processing the behavior of substate actors over time and space. *Political Analysis*, 16(4):464, 2008.

[34] Seth Stephens-Davidowitz. The effects of racial animus on voting: Evidence using Google search data. Job market paper, downloaded from http://www.people.fas.harvard.edu/~sstephen/papers/RacialAnimusAndVotingSethStephensDavidowitz.pdf, Novermber 2011.

[35] Sean M. Gerrish and David M. Blei. A language-based approach to measuring scholarly impact. In *Proceedings of ICML Workshop on Computational Social Science*, 2010.

[36] Daniel Ramage, Christopher D. Manning, and Susan Dumais. Partially labeled topic models for interpretable text mining. In *Proceedings of the 17th ACM SIGKDD international conference on Knowledge discovery and data mining*, page 457465, 2011.

[37] Dani Yogatama, Michael Heilman, Brendan O'Connor, Chris Dyer, Bryan R. Routledge, and Noah A. Smith. Predicting a scientific community's response to an article. In *Proceedings of the 2011 Conference on Empirical Methods in Natural Language Processing*, 2011.

[38] Steven Bethard and Dan Jurafsky. Who should I cite: learning literature search models from citation behavior. In *Proceedings of the 19th ACM international conference on Information and knowledge management*, page 609618, 2010.

[39] Jacob Eisenstein, Brendan O'Connor, Noah A. Smith, and Eric P. Xing. A latent variable model for geographic lexical variation. In *Proceedings of the 2010 Conference on Empirical Methods in Natural Language Processing*, page 12771287, 2010.

[40] Jacob Eisenstein, Noah A. Smith, and Eric P. Xing. Discovering sociolinguistic associations with structured sparsity. In *Proceedings of ACL*, 2011.

[41] Jeremy Ginsberg, Matthew H. Mohebbi, Rajan S. Patel, Lynnette Brammer, Mark S. Smolinski, and Larry Brilliant. Detecting influenza epidemics using search engine query data. *Nature*, 457(7232):1012–1014, February 2009.

[42] Aron Culotta. Towards detecting influenza epidemics by analyzing twitter messages. 2010.

[43] Michael J. Paul and Mark Dredze. You are what you tweet: Analyzing twitter for public health. In *Proceedings of ICWSM*, 2011.

[44] Peter S. Dodds and Christopher M. Danforth. Measuring the happiness of Large-Scale written expression: Songs, blogs, and presidents. *Journal of Happiness Studies*, page 116, 2009.

[45] Jean-Baptiste Michel, Yuan Kui Shen, Aviva Presser Aiden, Adrian Veres, Matthew K. Gray, The Google Books Team, Joseph P. Pickett, Dale Hoiberg, Dan Clancy, Peter Norvig, Jon Orwant, Steven Pinker, Martin A. Nowak, and Erez Lieberman Aiden. Quantitative analysis of culture using millions of digitized books. *Science*, 331(6014):176 –182, January 2011.

[46] Klaus Krippendorff. *Content analysis: an introduction to its methodology*. Sage Publications, Inc, 2004.

[47] Kurt Bollacker, Colin Evans, Praveen Paritosh, Tim Sturge, and Jamie Taylor. Freebase: a collaboratively created graph database for structuring human knowledge. In *Proceedings of the 2008 ACM SIGMOD international conference on Management of data*, pages 1247–1250, Vancouver, Canada, 2008. ACM.

[48] Justin Grimmer and Brandon M. Stewart. Text as data: The promise and pitfalls of automatic content analysis methods for political texts. `http://www.stanford.edu/~jgrimmer/tad2.pdf`, 2011.

[49] David M. Blei, Andrew Y. Ng, and Michael I. Jordan. Latent dirichlet allocation. *The Journal of Machine Learning Research*, 3:9931022, 2003.

[50] David M. Blei and Jon D. McAuliffe. Supervised topic models. *arXiv:1003.0783*, March 2010.

[51] David Mimno and Andrew McCallum. Topic models conditioned on arbitrary features with dirichlet-multinomial regression. In *Uncertainty in Artificial Intelligence*, page 411418, 2008.

[52] Trevor Hastie, Robert Tibshirani, and Jerome H. Friedman. *The elements of statistical learning: data mining, inference, and prediction*. Springer, June 2009.

[53] Christopher D. Manning, Prabhakar Raghavan, and Hinrich Schtze. *Introduction to Information Retrieval*. Cambridge University Press, 1st edition, July 2008.

[54] Sanford Weisberg. *Applied linear regression*. John Wiley and Sons, 2005.

[55] Alan Agresti. *Categorical data analysis*. John Wiley and Sons, 2002.

[56] Andrew Gelman and Jennifer Hill. *Data Analysis Using Regression and Multilevel/Hierarchical Models*. Cambridge University Press, 1 edition, December 2006.

[57] Matthew A. Taddy. Inverse regression for analysis of sentiment in text. *arXiv:1012.2098*, December 2010.

[58] Jacob Eisenstein, Ahmed Ahmed, and Eric P. Xing. Sparse additive generative models of text. *Proceedings of ICML*, 2011.

[59] Sean M. Gerrish and David M. Blei. Predicting legislative roll calls from text. In *Proceedings of ICML*, 2011.