Computational Text Analysis for Social Science: Model Assumptions and Complexity

Brendan O'Connor  
*Carnegie Mellon University*

David Bamman  
*Carnegie Mellon University*

Noah A. Smith  
*Carnegie Mellon University, nasmith@cs.cmu.edu*

Follow this and additional works at: [http://repository.cmu.edu/lti](http://repository.cmu.edu/lti)

Published In  
Proceedings of the NIPS Workshop on Computational Social Science and the Wisdom of Crowds.
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 sentiment in the media affect the stock market [2, 3]? 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]? 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. 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. 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 \( p(x | y) \) 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


