Research Statement—Social Data Science with Text

Katherine A. Keith
University of Massachusetts Amherst

My research is in the domain of *social data science*, answering questions about human behavior through quantitative analysis of large-scale data. I focus on methods and applications with *text* data because *language* is one of the richest and most salient expressions of human thought and behavior. Yet, drawing empirical insights from language is technically challenging because of its:

- **High-dimensionality.** Suppose a sample of documents are k words long and these words are drawn from a vocabulary of w possible words. Then there are w^k unique representations of these documents. Consider a short document with only 10 words and a vocabulary of 1,000; there are then $1000^{10} = 10^{30}$ unique representation possibilities. Thus, analyzing text data typically requires some sort of dimensionality reduction.
- **Linguistic complexity.** Language is more than just a *bag-of-words*. *Semantics*, the meaning of language, is built from *syntax*, the structure of language. For instance, "Mary likes John" takes a different semantic meaning than "John likes Mary" even though both examples contain the same words. Language can also be ambiguous; in the example "She saw the man with the telescope" the prepositional phrase "with the telescope" could attach to either "saw" or "man," giving the sentence two distinct but plausible meanings.
- **Large-scale.** The digital age has led to collections of increasingly large textual datasets. My research involves datasets such as 1.2 million news articles [8]; 12,000 transcripts of financial earnings calls that contain over 0.5 million question-answer pairs between company executives and analysts [11]; 4.1 million online user reviews [10]; and 53.1 million comments from a top social network. Increased size can lead to increased statistical power, but at a cost—manual analysis does not scale and the computational infrastructure for storing and analyzing this data is increasingly complex.

To address these technical challenges, my research expands methods in **machine learning** and **natural language processing** to *social data science* goals including: obtaining quantifiable **measurements** from text data (§3), **aggregating** said measurements in a statistically rigorous manner, and improving **causal** estimations from text (§4). I then apply these methods to applications such as extracting macro-social measures from newspapers (§1) and studying the language of economic decision making (§2). These applications require implementation of the entire data science pipeline: data collection, data cleaning, crowd-sourcing, applied machine learning, statistical and causal analysis, and visualization. My work complements the research goals of a liberal arts college because it is interdisciplinary, accessible to undergraduate researchers (§5), and has the potential for social impact.

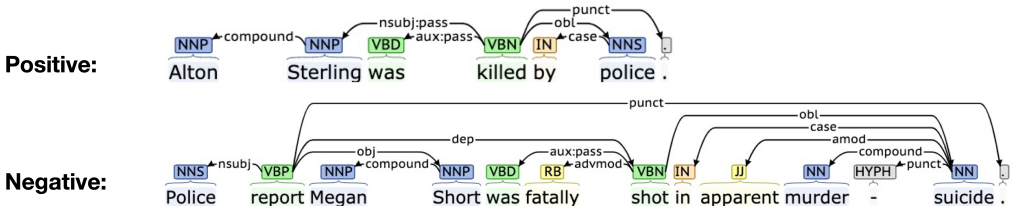

Figure 1: Instances of positive (top) and negative (bottom) classes of civilians killed by police [8]. Sentences are automatically labeled with *dependency parses*.

1 *Application Area:* News-Based Macro-Social Measures

News reports are a large-scale, historical record of happenings in society. My research uses news to measure *counterdata*, data under-collected or not collected by central governments, and extracts economic or political signals from news articles. In one project, my collaborators and I built a probabilistic machine learning model that extracts the names of civilians killed by police in the United States from news reports [8]. As of 2015 when our project started, the United States government severely undercounted the number of civilians killed by police [5]. As *counterdata*, non-profits and news organizations have manually read news articles to build databases of U.S. police killings, but only with enormous costs of human time and emotional strain. Our work augmented this manual effort by building an automated data science pipeline that continuously scrapes news articles from Google News, cleans and de-duplicates the article text, and uses our machine learning model to extract the names of victims (see Figure 1). This is not the only source of *counterdata* currently being collected by manually reading news reports; for instance, a single volunteer citizen has read news articles to log more than five thousand cases of femicide in Mexico since 2016.[1] Future research could engage undergraduates majoring in Communications, Spanish, and/or Gender Studies to extend my previous work to document femicide or collect other *counterdata* from news reports.

My work also automates and extends efforts by social scientists to extract economic or political signals from news. In economics, Baker et al. derive an impactful measure of *economic policy uncertainty* by matching keywords in news articles [1].

[1] `https://feminicidiosmx.crowdmap.com`

I replicated and extended their data science pipeline and found that some disagreements in their human annotations can be attributed to inherent ambiguity in language, and that swapping measurement methods from keyword-matching to supervised machine learning classifiers results in low correlation, a concerning implication for the validity of their original measure [12]. In another ongoing project, I am collaborating with a political scientist at Massachusetts Institute of Technology to build an *entity-anchored* data science pipeline that detects and extracts the actions of political entity types over time from news articles. Building from the manual efforts of political scientists who study state-level actors in communal violence [13], our work extracts all actions of *police* from the *Times of India* across 20 years. With funds from a *Kaggle Open Data Research Grant*, we have currently employed 12 political science undergraduates to manually annotate six months of documents in order to validate our machine learning model's predictions. In this and other large-scale data science projects with news, data collection and annotation are an accessible first research experience for undergraduates.

2 *Application Area:* Language of Economic Decision Making

As a *Bloomberg Data Science PhD Fellow*, I have had the privilege to work in close collaboration with economists, financial experts, and financial data scientists at Bloomberg over the last several years. This experience has revealed an interesting line of research that applies *social data science with text* methods to study the decision-making of economic and financial actors. In particular, a collaborator and I examined analysts' decision making behavior as it pertains to the language content of earnings calls, quarterly conference calls between company executives and financial analysts [11]. We extracted a set of 20 pragmatic and discourse features from the questions of earnings calls and correlated these with analysts' pre-call judgments. We found bullish analysts tend to be called on earlier in calls, and ask questions that are more positive, more concrete, and less about the past. We also applied natural language processing methods to predict changes in analysts' post-call price targets. Our null hypothesis was that earnings calls are not predictive of forecast changes since analysts' have access to private information; however, our best model reduced relative accuracy error by 25% over a majority class baseline, suggesting there is signal in the noise. In ongoing and future research, I have several economist collaborators who are interested in examining the pragmatic and semantic features of economic bargaining in mobile apps and understanding how adding natural language communication to classical economic games influences how participants cooperate or defect.

3 *Methods Theme:* Measurement from Text

A major methodological theme that runs through many of these previous application areas is *measurement*, quantifying theoretical concepts in observable data. My work has addressed measuring social signals from text by expanding natural language processing and machine learning methods of dependency parsing, distant supervision, quantifying uncertainty, and prevalence estimation.

In previous work, I used rules over *dependency parses* to measure entities and events in text [7], and dependency parse paths as features in machine learning models [8]. *Dependency parsing* is a natural language processing task that formalizes relationships between words in a sentence as a directed graph. For example, in the top sentence in Figure 1, there is a directed edge from "killed" to "Sterling" with edge label "nsubj:pass," which stands for "passive nominal subject" and indicates Alton Sterling is the recipient of the "kill" action. The bottom sentence in Figure 1 would be correctly classified as a negative via a rule over dependency paths that requires a word like "police" to be attached to a word like "shot." My collaborators and I demonstrated that we can improve downstream performance on these sorts of entity or event measurement tasks by replacing a standard, greedy dependency parse algorithm that infers a single dependency tree with an algorithm that uses *Monte Carlo sampling* to sample from the full joint distribution of parse trees [7]. The latter algorithm is able to better propagate parse uncertainty to downstream tasks.

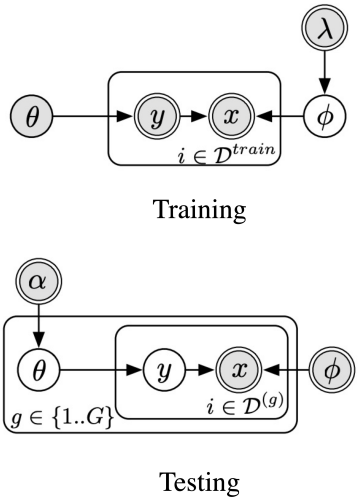

Figure 2: Generative model for *prevalence estimation* [10]. *Top:* Class-conditional language models (ϕ) are learned at training time. *Bottom:* Test-time interference for multiple groups' (g) latent prevalences (θ).

With unlimited labeled examples, a *supervised machine learning* approach to most *measurement* tasks would be straightforward. However, requiring humans to manually label thousands if not millions of examples is extremely costly. Instead, my collaborators and I expanded *distant supervision* for better measurement of social variables from text. *Distant supervision* heuristically aligns structured data in a knowledgebase to text and imputes positive labels for supervised learning [6]. For inferring civilians killed by police, we align names from an external database, *Fatal Encounters*, with mentions of those names in text. Yet, this distant supervision heuristic is often wrong because examples labeled positives by the distant supervision heuristic are often true negatives; for example, "Alton Sterling's mother spoke at his funeral" is an example that does not express the event in question, that a person was killed by police, but would be labeled as positive by the heuristic. Thus, my collaborators and I developed a model that treats sentence labels from the distant supervision heuristic as latent variables and assumes that at least one of the sentences for a given name asserts a fatality event, but leaves

uncertainty as to which one is a positive. We used an *expectation maximization algorithm* which infers a posterior over the latent sentence-label during the *E-step* and then calculates the expected log-likelihood in the *M-step* [8].

Although dependency parsing and distant supervision can be used to measure social signals from individual documents, social scientists are often more interested in an *aggregate* statistic over a group of documents. For example, one may want to measure the aggregate daily sentiment on Twitter about a U.S. president, but predicting the sentiment of an individual Tweet is not important. This is a task called *prevalence estimation*, inferring the relative frequency of classes of unlabeled examples in a group. My collaborators and I developed a *generative* probabilistic model to prevalence estimation, and constructed and evaluated prevalence confidence intervals (see Figure 2) [10]. Empirically, we demonstrated our approach provides better confidence interval coverage than more widely used alternative prevalence estimation methods, and our approach is dramatically more robust to shifts in the class prior between training and testing. Future work could focus on *cardinality estimation* of events from corpora and explicitly model these events using a similar generative framework.

4 *Methods Theme:* Causal Inference with Text

Although measuring and aggregating social phenomena from text is extremely useful to understand human behavior, one of the ultimate goals of science is *causal* understanding, identifying relationships that remain invariant when external conditions change. Many applications aim to infer causal conclusions from observational (non-experimental) data, which often contains *confounders*, variables that influence both potential causes and potential effects. After reviewing existing literature that adjusts for confounding using text data, my collaborators and I provided a guide to data-processing and evaluation decisions in this space (see Figure 3) [9]. In the future, I plan to focus on some of the open research problems we presented in our paper including: How does one mitigate error that arises from approximating confounding variables with imperfect natural language processing methods? For causal inference matching methods, how can human judgement experiments be improved and standardized? How sensitive are causal effects to hyperparameter and network architecture choices and what should researchers do in these settings? How do we develop text-based causal inference datasets with ground truth? In an ongoing project, I am developing *text-conditional causal methods* for content moderation on online platforms. So far, we have collected the text of over 0.5 million comments that have been removed by moderators on *Reddit*, and we clustering semantically near-equivalent comments in order to estimate the moderation policies of different communities on the platform.

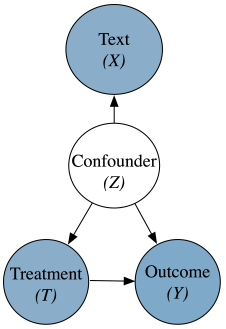

Figure 3: Causal graph for settings in which text is used to infer latent confounders [9]

5 Future Research Collaborations

Undergraduates in research. I view my work as an accessible entry into research for undergraduates. I have found undergraduates are successful when they are recruited in *cohorts*, and I plan to *scaffold* future undergraduate research experiences such that younger students with less technical backgrounds could be introduced to research by collecting new textual data sources, cleaning data, and manually annotating data while more experienced students with backgrounds in statistics or machine learning could work on targeted methodological improvements. During the summer of 2019, I recruited and supervised three undergraduates from the University of Massachusetts Amherst who researched extensions of our generative prevalence estimation method [10]. One student extended this method from binary to multi-class, another student substituted different classification methods into the model and measured results, and the final student applied our prevalence estimation model to applications in the online legal help domain. One of these these students expanded his summer work into a successful senior honors thesis, which I supervised, that bootstrapped training data to account for uncertainty in the learned parameters of the model and evaluated the model's calibration. In order to set students up for success, I provide thorough organization and preparation before students begin research projects; for example, I created a multi-page "onboarding" document[2] for my REU students that detailed project goals, communication protocol, research expectations, and general research advice.

Interdisciplinary research. *Social data science with text* is a research domain that is inherently interdisciplinary with demonstrated interest from political scientists [4], sociologists [2], and economists [3]. In the future, I could extent past projects [8, 11] to build automatic natural language processing tools for computational journalism, or extend other projects [11, 12] to dig deeper into the substantive economic or financial results. The interdisciplinary nature of my research makes working at a liberal arts college ideal, and in the future, I plan to collaborate with other social science and humanities colleagues who have large-scale, text-based research questions.

References

[1] Scott R Baker, Nicholas Bloom, and Steven J Davis. Measuring Economic Policy Uncertainty. *The Quarterly Journal of Economics*, 131(4):1593–1636, 2016.

[2] James A Evans and Pedro Aceves. Machine translation: Mining text for social theory. *Annual Review of Sociology*, 42:21–50, 2016.

[2] `https://github.com/kakeith/reu-resources/blob/master/onboarding.pdf`

[3] Matthew Gentzkow, Bryan Kelly, and Matt Taddy. Text as data. *Journal of Economic Literature*, 57(3):535–74, 2019.

[4] 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, 2013.

[5] Kristian Lum and Patrick Ball. Estimating undocumented homicides with two lists and list dependence. *Human Rights Data Analysis Group*, 2015.

[6] Mike Mintz, Steven Bills, Rion Snow, and Dan Jurafsky. Distant supervision for relation extraction without labeled data. In *Proceedings of the Joint Conference of the 47th Annual Meeting of the ACL and the 4th International Joint Conference on Natural Language Processing of the AFNLP*, pages 1003–1011, 2009.

[7] **Keith, Katherine A.**, Su Lin Blodgett, and Brendan O'Connor. Monte Carlo Syntax Marginals for Exploring and Using Dependency Parses. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (NAACL-HLT)*, 2018.

[8] **Keith, Katherine A.**, Abram Handler, Michael Pinkham, Cara Magliozzi, Joshua McDuffie, and Brendan O'Connor. Identifying civilians killed by police with distantly supervised entity-event extraction. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing (EMNLP)*, 2017.

[9] **Keith, Katherine A.**, David Jensen, and Brendan O'Connor. Text and Causal Inference: A Review of Using Text to Remove Confounding from Causal Estimates. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics (ACL)*, 2020.

[10] **Keith, Katherine A.** and Brendan O'Connor. Uncertainty-aware generative models for inferring document class prevalence. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing (EMNLP)*, 2018.

[11] **Keith, Katherine A.** and Amanda Stent. Modeling financial analysts' decision making via the pragmatics and semantics of earnings calls. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics (ACL)*, 2019.

[12] **Keith, Katherine A.**, Christoph Teichmann, Brendan O'Connor, and Edgar Meij. Uncertainty over Uncertainty: Investigating the Assumptions, Annotations, and Text Measurements of Economic Policy Uncertainty. In *Proceedings of the Fourth Workshop on Natural Language Processing and Computational Social Science (NLP+CSS) at EMNLP*, 2020.

[13] Steven I Wilkinson. *Votes and violence: Electoral competition and ethnic riots in India.* Cambridge University Press, 2006.