Forthcoming in the APSA-MENA Section Newsletter

What Counting Words Can Teach Us about Middle East Politics
Richard A. Nielsen

I stared at the word "God" – *allah* in Arabic – at the top of the list on my computer screen. I was puzzled. I rechecked the computer code. It seemed correct. But how could it be that the single word that most distinguished male and female preachers on the Salafi missionary website www.saaid.net was the word "God?" From reading articles posted there, I knew that it wasn't because female preachers were any less fervent than their male counterparts about orienting their followers towards the divine. But the word count was correct. Male preachers used the word "God" incredibly frequently, once every thirty-three words; almost every other sentence. Female preachers used it only half as often.

My next discovery deepened the puzzle. Female preachers were using the word "God" less than men because they use fewer citations to the Quran and the *hadith* tradition (the sayings of Muhammad and his companions). Following Islamic custom, these citations involve bound phrases that almost always include the word "God." There isn't a gendered piety gap in Salafi Islam, but rather a gendered citation-use gap. But why? Citations to these authoritative texts are the defining feature of the "Salafi method" (*al-manhaj al-salafiyya*) of establishing legal-religious authority with readers. Ethnographers observing these same women preaching in person have concluded that their "knowledge of Quran and Sunnah [is] exhaustive" (Le Renard 2012, 116). So why do Salafi women cite the *hadith* and Quran only half as often as men when they write online?

The answer is that women in the Salafi movement construct their authority differently from men. Rather than relying as heavily on citations for authority, they invoke *identity authority* as women to deliver religious messages that men can't. For example, female preachers are uniquely able to oppose the UN women's rights laws (a common Salafi target) by saying "As a woman, I don't want the West's so-called 'rights.'" Although the Salafi movement's norms are unfriendly to the theoretical idea of women's religious authority, male movement leaders nevertheless promote these female authorities because their messages defend patriarchal practices and attract new online audiences of both women and men. These insights challenge previous conclusions about these female Salafi preachers (Le Renard 2012, 2014; Al-Rasheed 2013): that they use the Salafi method of *hadith* citation just as much as men, that they write exclusively on so-called "women's issues," and that men are uninterested in their preaching (in fact, 70% of Twitter reactions to women's preaching are from men). My findings form the basis of my article "Women's Authority in Patriarchal Social Movements: The Case of Female Salafi Preachers," forthcoming in the *American Journal of Political Science.*

I share this behind-the-scenes story about my research process to illustrate the power of quantifying text. When we read, our brains fill in the gaps with our prior beliefs about what we ought to find (Goodman 2014). Despite years of familiarity with the texts on this Salafi website, it wasn't until I started counting words that I was able to see the stark gender differences. Of course, these quantitative differences are merely a numerical summary of a qualitative difference, but one that I was blind to until I started counting.

Statistical text analysis is sometimes dressed up with terms like "artificial intelligence" and "machine learning." These descriptors aren't wrong, at least in their technical usage, but they obscure the fundamental simplicity of text analysis: it is largely based on counting words. The difference between the simple methods and the complex ones is the complexity of the word count. Scholars of the Middle East have occasionally turned to computational text analysis in the past (Bulliet 1979), but recent resurgence in interest and new tools are drawing a new wave of young scholars to forge ahead (Mitts 2019, Karell and Freedman 2019, Siegel and Tucker 2018)

Counting words is no substitute for reading them, but by the same token, reading words is not always a substitute for counting them. Our brains understand narrative and insinuation in a way computers cannot, and they bring a wealth of prior knowledge to every reading. But they are stunningly bad at probability and prone to a variety of cognitive biases. And we get bored. The promise of applying statistical text analysis methods to Arabic text is that we can harness both modes of investigation for greater insight about the politics of the Arab world.

Counting to discover

Discovering new things with statistical text analysis begins with a text, or set of texts, that are puzzling in some way. Most guides to statistical text analysis assume that the analyst already has texts in hand (Grimmer and Stewart 2013; Lucas et al 2015). This obscures the reality that in my experience, I spend upwards of 80% of my time on any given project selecting, collecting, and curating the texts I will analyze. Selecting texts for discovery is both a science and an art. As usual, "the cases you choose affect the answers you get" (Geddes 1990) but when I pursue discovery via text analysis, I rarely have a fully developed question. Instead, I am usually intrigued and puzzled by a collection of texts.

How should one sample texts for discovery? When I explore a collection of texts, I generally try to explore the entire collection as demarcated by some kind of natural boundary: a website, an author, a movement, or an era. For example, my puzzlement about female Salafi preachers began with two lists of 172 male and 43 female preachers on www.saaid.net, so I collected every document available on the website by all of these authors. Collecting only the easiest-to-get texts or the most famous preachers might have biased my results. On the other hand, I did not collect texts from the many other Salafi websites on the Internet. I believe my findings hold true there as well, but Internet-wide data collection is extremely time-consuming (Internet-wide data collection for my book *Deadly Clerics* took several years). The general principle here is that context is just as crucial for statistical text analysis as it is for close reading. The religious texts I analyze in my research are in dialogue with each other, and artificially subdividing the documents risks missing these connections. Natural boundaries that authors and readers draw around corpora can often serve as useful analytic boundaries for scholars as well.

Using text analysis for discovery is often called *unsupervised learning*, a name from computer science referencing the goal of having a computer "learn" what a corpus of texts means without human input. The most widely used unsupervised learning method in political science is a topic model, which is one of the ways I explored the gendered differences in Salafi preaching. The insight of a topic model is that rather considering differences between texts on a word-by-word basis, we can group those words into topics and consider differences topic-by-topic. This is still word counting, but through a far more complex calculation. Several papers give treatments of the technical details (Blei, Ng, and Jordan 2003, Roberts, Stewart, and Airoldi 2016); my goal here is instead to give the intuition.

The goal of a topic model is to summarize the words in a corpus with a small number of dimensions, colloquially called "topics" because in practice, they often correspond to what humans think of as topical. The model proceeds with a set of unrealistic assumptions about how documents are written. The imagined author has a fixed list of topics, each with words that they are more likely to use when writing on that topic. When they sit down to write each document, they sample proportions for how much of each topic will be in this particular document, and then for each word, they sample first a topic, and then word conditional on that topic. The model takes word counts we observe in a corpus and estimates the parameters for this imagined model that would have been most likely to result in those word counts, if the model were an accurate summary of how the texts were written. Practically speaking, analysts look at the resulting lists of correlated words and interpret them as the main topics of the corpus.

The topic model is clearly unrealistic; no one writes in this way. In fact, if I did, the next sentence in this paragraph might have been the agrammatical: "One scholar model interpret algorithm."[1] But despite these unrealistic assumptions, topic models have proven very useful

[1] I created this short sentence by estimating a topic model on the words in this article, treating each paragraph as a separate document. The model estimated that this paragraph was 98% devoted to a topic I interpret to be about topic models (keywords: topic, model, human, corpus, algorithm, goal, interpret). I sampled the five words in this sentence using the word probabilities estimated by the model for this topic. Code to reproduce this process is available on my website at http://www.mit.edu/~rnielsen/research.htm.

for a wide range of researchers seeking to make discoveries. Why? Because models do not have to be realistic in all the particulars to be useful. Clark and Primo (2012) argue that we should view models as maps, and that maps routinely employ unrealistic distortions to be useful while remaining parsimonious. Topic models have proven to be useful maps for exploring a wide variety of corpora.

Topic models are useful because the latent dimensions returned by the algorithm often help researchers interpret the contents of their texts. Although the model is statistical, the goal is generally *interpretive* as defined by Pachirat (2006): "Humans making meaning out of the meaning making of other humans." Scholars of Middle East politics are at the forefront of interpretive approaches to social science (Wedeen 2002, Sarah Parkinson 2013, Jones 2015). As quantitatively oriented scholars increasingly use topic models that put interpretation front and center, perhaps this will create space for connection between these two traditions.

Counting to scale up

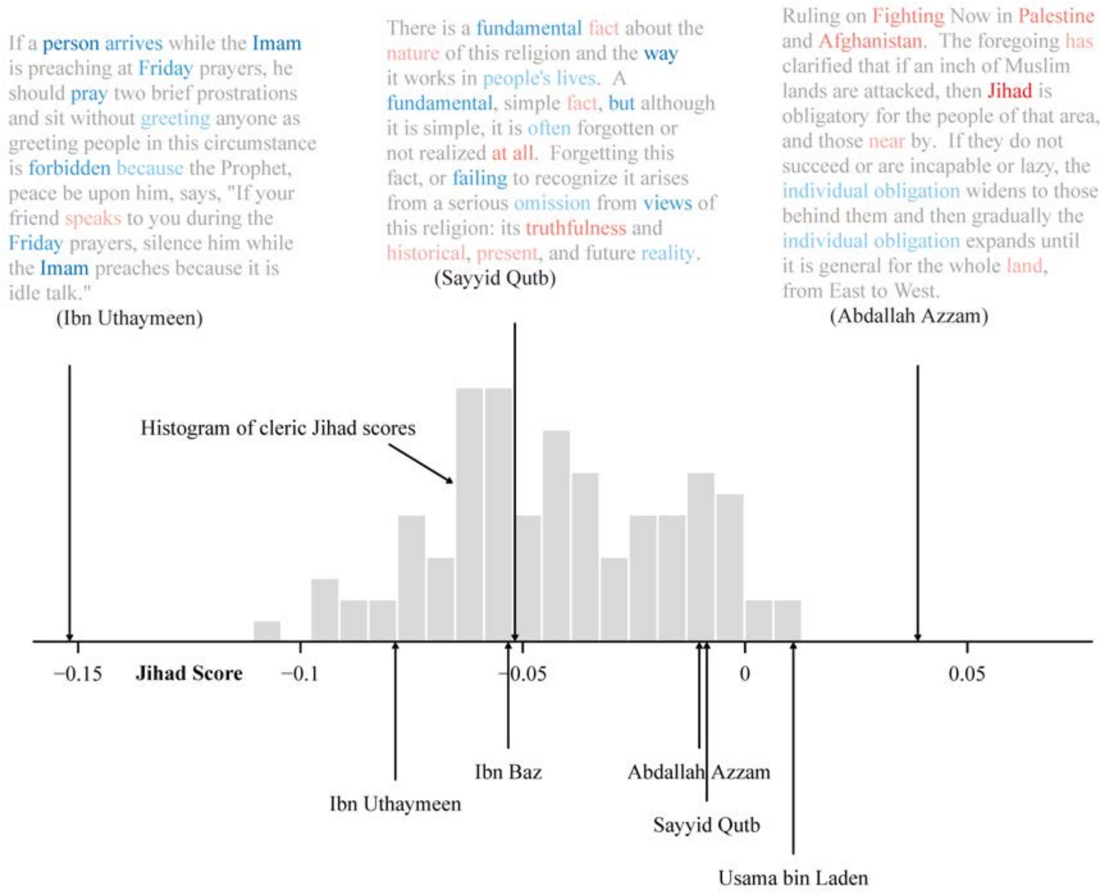
When I get questions about whether text analysis is appropriate for some project, I typically return this question with one of my own: if you had infinite time to read all of the text yourself, what would you do? Often, researchers respond with a relatively simple reading task for "coding" each text, but reading and coding every text in a large corpus would take months or years. Statistical text analysis offers a way to "scale up" reading and coding. This approach is called *supervised learning*, a name that refers to the notion that the computer is learning to reproduce the task that a human would do using supervision from human inputs.

There are many, many supervised learning algorithms, but the essence of these algorithms is similar. The analyst begins with labeled data, a subsample of the texts where the desired coding has already been done. The analyst then uses this labeled data to "train" one or more algorithms, selecting parameters for each algorithm that give good performance when attempting to relabel the already labeled data. The analyst then applies the trained algorithm to the unlabeled data to generate labels in minutes, rather than years.

I used supervised learning to classify the writings of Muslim clerics as jihadist or not in my book *Deadly Clerics* (2017). This classification was part of a larger analysis testing whether weak academic networks make clerics more likely to preach jihad, but for now, I focus only on the classification task. I was working with approximately 150,000 documents by 200 Arabic-speaking clerics, with lengths ranging from a few sentences to multivolume tomes. If I had been able to skim each document in five minutes on average to render a rough coding, classifying each of these documents would have taken me approximately 12,000 continuous hours. Instead, I used *The Jihadist's Bookbag*, a set of jihadist documents circulating on the web, to train an algorithm called a naïve Bayes classifier to detect other jihadist documents.

Heuristically, the model compares the word counts in a new document to word counts in *The Jihadist's Bookbag* and classifies the new document as jihadist if they are similar. This approach is faster than human coding and allowed me to use the expertise of jihadists themselves to determine which documents count as jihadist.

Figure 1



If a person arrives while the Imam is preaching at Friday prayers, he should pray two brief prostrations and sit without greeting anyone as greeting people in this circumstance is forbidden because the Prophet, peace be upon him, says, "If your friend speaks to you during the Friday prayers, silence him while the Imam preaches because it is idle talk."
(Ibn Uthaymeen)

There is a fundamental fact about the nature of this religion and the way it works in people's lives. A fundamental, simple fact, but although it is simple, it is often forgotten or not realized at all. Forgetting this fact, or failing to recognize it arises from a serious omission from views of this religion: its truthfulness and historical, present, and future reality.
(Sayyid Qutb)

Ruling on Fighting Now in Palestine and Afghanistan. The foregoing has clarified that if an inch of Muslim lands are attacked, then Jihad is obligatory for the people of that area, and those near by. If they do not succeed or are incapable or lazy, the individual obligation widens to those behind them and then gradually the individual obligation expands until it is general for the whole land, from East to West.
(Abdallah Azzam)

The payoff, shown in Figure 1, is a ranking of Muslim clerics from least to most jihadist, based on the similarity of their writing to *The Jihadist's Bookbag*.[2] The numeric scale is arbitrary; what matters is that non-jihadists fall to the left of the histogram and jihadists fall to the right. For comparison, I plot scores for excerpts from the writings of Ibn Uthaymeen, Sayyid Qutb, and Abdallah Azzam. For each excerpt, the words that actually enter the model are colored (the classifier omits the most and least frequent terms), with words that predict Jihadism in darker red and words that predict non-Jihadism in darker blue. A careful reader could have made the same judgment about these texts, but coding would have taken years instead of hours.

[2] An updated version of this figure appears in my book *Deadly Clerics* (2017) on page 122, along with more explanation of the method.

Getting Started with Statistical Text Analysis in Arabic

I've said relatively little about the particulars of statistical text analysis in Arabic because my view is that the principles of text analysis are fundamentally similar across languages (Lucas et al. 2015). I think of an analysis of Arabic-language texts no differently than an analysis of English-language texts; there is merely an added technical challenge of representing Arabic in a computer program, and of replacing English-specific preprocessing steps with an appropriate Arabic-language equivalent. But these technical challenges can be frustrating for scholars making their first foray into statistical text analysis, especially because not all of these challenges have well-established solutions.

As a language, Arabic presents a number of challenges that methodologists working with English-language texts have rarely considered. There is substantial dialect variation across the Arab world; enough so that different dialects appear to be different languages to a computer algorithm. Existing approaches to multi-language text analysis rely on translating to a single "pivot language" (Lucas et al. 2015), but automated translation systems for most Arabic dialects do not exist. Couple this with occasional script variation, and the frequent use of Latin characters to represent Arabic letters (called "Arabizi") in online writing, and the challenges can become overwhelming. I have largely side-stepped these problems because the clerics I study tend to write in regularized, formal Arabic. But several of the other essays in this symposium deal with these challenges head-on.

Text analysis also involves language specific preprocessing steps. Often, these preprocessing decisions are assumed to be innocuous, but recent research shows they are not (Denny and Spirling, 2018). For Arabic, the step that is most different is stemming: the process of combining words with similar "stems" into a single term. For example, in English, we might combine the words "teacher," "teaching," and "teachable" into a single stem "teach." This reduces the complexity of the text by helping the computer "learn" that all of these words relate to a single concept. English is relatively easy to stem because it uses suffixes and prefixes to create new words from older concepts. But Arabic morphology relies on infixing as well, which presents a serious challenge. When I started, resources for working with Arabic-language text in modern statistical languages were underdeveloped or non-existent. I coded and released an Arabic-language stemmer for the R programming language (`arabicStemR`, Nielsen 2017) because this was the most crucial tool that was missing. Scholars interested in learning the details of my Arabic text analysis workflow should check out online materials I developed for a workshop on the topic at Cairo University this year, available on my website.[3]

[3] http://www.mit.edu/~rnielsen/arabicTextWorkshop.zip

Despite these challenges, Arabic text analysis is going to become a mainstay method for Middle East scholars. The technical challenges will be met with technical solutions in fairly short order. The new wave of research, described in this symposium, will make a splash and inspire even more research. But Arabic text analysis will also gain traction for a more somber reason: access to field sites in the Middle East and North Africa is closing, especially to researchers asking political questions. A resurgence of authoritarianism in the wake of the Arab uprisings means that almost any political inquiry crosses regime red lines in much of the region and field research can look a lot like spycraft to paranoid autocrats (Driscoll and Schuster 2018). Local activists are responding to this repression by moving online; their conversations create the social media data that Alex Siegel is analyzing in this symposium. As face-to-face fieldwork becomes more difficult, and even life-threatening, senior scholars must sometimes make the ethical choice to *not* encourage students to place their bodies into harm's way as they carry out dissertation projects (Lynch 2018). As the physical field closes, and the online field opens, statistical analysis of Arabic texts offers one way forward.

References

Al-Rasheed, Madawi. 2013. *A Most Masculine State: Gender, Politics, and Religion in Saudi Arabia.* New York: Cambridge University Press.

Blei, David M., Andrew Y. Ng, and Michael I. Jordan. "Latent dirichlet allocation." Journal of machine Learning research 3.Jan (2003): 993-1022.

Bulliet, Richard W. "Conversion to Islam in the medieval period: an essay in quantitative history." (1979).

Clarke, Kevin A., and David M. Primo. A model discipline: Political science and the logic of representations. Oxford University Press, 2012.

Denny, Matthew J., and Arthur Spirling. "Text preprocessing for unsupervised learning: Why it matters, when it misleads, and what to do about it." Political Analysis 26.2 (2018): 168-189.

Driscoll, Jesse, and Caroline Schuster. "Spies like us." Ethnography 19.3 (2018): 411-430.

Geddes, Barbara. "How the cases you choose affect the answers you get: Selection bias in comparative politics." Political analysis 2 (1990): 131-150.

Goodman, Kenneth S. "Reading: A psycholinguistic guessing game." Making Sense of Learners Making Sense of Written Language. Routledge, 2014. 115-124.

Grimmer, Justin, and Brandon M. Stewart. "Text as data: The promise and pitfalls of automatic content analysis methods for political texts." Political analysis 21.3 (2013): 267-297.

Jones, Calvert W. "Seeing like an autocrat: Liberal social engineering in an illiberal state." Perspectives on Politics 13.1 (2015): 24-41.

Karell, Daniel, and Michael Freedman. "Rhetorics of Radicalism." American Sociological Review 84.4 (2019): 726-753.

Le Renard, Amelie. 2012. "From Qur'anic Circles to the Internet: ´Gender Segregation and the Rise of Female Preachers in Saudi Arabia." In Women, Leadership, and Mosques: Changes in Contemporary Islamic Authority, ed. Masooda Bano and Hilary Kalmback. Leiden, The Netherlands: Brill, 105–26.

Le Renard, Amélie. A society of young women: opportunities of place, power, and reform in Saudi Arabia. Stanford University Press, 2014.

Lucas, Christopher, Richard A. Nielsen, Margaret E. Roberts, Brandon M. Stewart, Alex Storer, and Dustin Tingley. "Computer-assisted text analysis for comparative politics." Political Analysis 23, no. 2 (2015): 254-277.

Lynch, Marc. 2018. "What the UAE's arrest of Matthew Hedges means for political science research in the Middle East" *The Washington Post, The Monkey Cage,* 29 October 2018.

Mitts, Tamar. "From isolation to radicalization: anti-Muslim hostility and support for ISIS in the West." American Political Science Review 113.1 (2019): 173-194.

Nielsen, Richard A. 2017. *Deadly Clerics: Blocked Ambition and the Paths to Jihad*. New York: Cambridge University Press.

Nielsen, Richard A. Forthcoming. "Women's Authority in Patriarchal Social Movements: The Case of Female Salafi Preachers," *American Journal of Political Science.*

Pachirat, Timothy. 2006. "We Call It a Grain of Sand: The Interpretive Orientation and a Human Social Science," in Interpretation and Method: Empirical Research Methods and the Interpretive Turn, Dvora Yanow and Peregrine Schwartz-Shea, eds., pages 426-432.

Parkinson, Sarah Elizabeth. "Organizing rebellion: Rethinking high-risk mobilization and social networks in war." American Political Science Review 107.3 (2013): 418-432.

Roberts, Margaret E., Brandon M. Stewart, and Edoardo M. Airoldi. "A model of text for experimentation in the social sciences." Journal of the American Statistical Association 111.515 (2016): 988-1003.

Siegel, Alexandra A., and Joshua A. Tucker. "The Islamic State’s information warfare." Journal of Language and Politics 17.2 (2018): 258-280.

Wedeen, Lisa. "Conceptualizing culture: Possibilities for political science." American Political Science Review 96.4 (2002): 713-728.