take into consideration ethical and legal issues. First, intellectual property and copyright laws apply to these websites, so authors should not republish the texts of news articles (at least without explicit consent from the news source). Note that this is tricky, because it could make replication harder. Some researchers suggest that publishing the stemmed document-term matrix may be ethical, but there are no clear standards in the field yet.

A related concern is reading and understanding the terms and conditions of a website before scraping, as some explicitly ban it. In theory, scholars should use the Application Programming Interface (API) when scraping; however, in practice, I am not aware of any news websites in Arabic with APIs. These may be developed by some of these websites in the future.

Finally, it is important to make requests at a reasonable pace. Many Arabic news websites, especially small independent ones, cannot handle a lot of traffic. So if research make too many repeated scraping requests, researchers could unintentionally slow down or even shut down the websites, and this could be interpreted as a Denial of Service attack by the scholar.

There are many tools that make scraping relatively easy in R and Python. In R, *rvest* is a powerful library that makes it easy to scrape many news websites. After downloading the data, scholars often need to manipulate it. Sometimes websites include JavaScript, making it difficult to scrape using *rvest*. The open-sourced tools *Selenium* in Python or *Rselenium* in R can be particularly helpful to deal with these websites, as they allow the researcher to write code that browses the Internet like a user. There are several packages in R that are helpful to process and clean text data, including *tidytext*, *broom*, and *stringr*/*stringi*. As change in trends over time is often important with news media, *lubridate* is a particularly powerful package to manipulate dates. With Arabic texts, Nielsen's *arabicStemR* is particularly helpful for removing stop words and stemming. *Farasa* is another tool that allows parsing of Arabic texts, which also includes identifying parts of speech in Arabic sentences.

Conclusion

Conducting political research on the Middle East can often be challenging. Many regimes do not maintain easily accessible archives, restrict survey work, and sometimes even put the safety of researchers at risk. Arabic media thus offers a valuable source of data to learn about many questions related to regime behavior, intervention by foreign powers, public diplomacy, and political opposition in Arab countries. Yet the role of Arabic media goes beyond providing a source to study these topics. Research has shown that Arabic news media itself plays an important role in influencing regional politics. Quantitative text analysis provides researchers with the tools to use vast amounts of texts in order to study some of these topics and, in particular, the role of news media in politics. While there are important challenges and limitations, ongoing research projects that apply quantitative text analysis to Arabic media demonstrate the potential of these tools.

IDEOLOGICAL SCALING IN A POST-ISLAMIST AGE

By Nate Grubman, Yale University

In recent years, a growing body of scholarship has focused on the interaction of cultural identity and class as potential bases for partisanship in the Arab world. A number of puzzles have emerged: Following the 2010 to 2011 uprisings driven in large part by economic grievances, why did party systems in Egypt and Tunisia revolve more tightly around competing notions of religious and national identity than competing economic orientations?[120] Why did many of the poor turn to Islamist parties rather than Marxist-Leninist or Arab nationalist ones?[121] How can the perceived dominance of a secularist-Islamist cleavage and the popularity of Islamist parties be reconciled with the observation that citizens of Arab countries are concerned with the mundane economic issues that preoccupy other people of the world?[122]

In this article, I argue that the application of text-as-data methods to the speech disseminated by politicians can contribute new insight to each of these questions. Text-as-data methods can help researchers identify the issues on which politicians focus and the main differences in the ways they discuss said issues. These methods can be particularly valuable in understanding unfamiliar political actors, learning where familiar political actors stand with regard to unfamiliar issues (such as Islamists talking about purchasing power), or some combination thereof. To ground the discussion of the methods, I discuss my use of ideological scaling methods in an ongoing project to understand the choices presented by post-uprising Tunisian politicians, especially on economic issues. I reflect on my unreasonably high initial hopes, the conclusions I have drawn after several years of research, and potential uses for these methods beyond Tunisia.

How and why to scale campaign materials

Survey research has raised a set of puzzles regarding the place of economic policy differences in partisanship in the Arab world. Researchers have noted that those usually expected to support the left—the poor and other advocates of redistribution—have generally not in the Arab world.[123] Furthermore, shortly after Egypt and Tunisia's party-system liberalizations, efforts to task survey respondents with mapping the parties onto an economic spectrum produced inconclusive or puzzling results.[124]

These findings raise questions about the choices parties have presented to voters, especially with regard to economic problems. Did political parties in Egypt and Tunisia diverge in their approaches to economic issues? If so, which parties positioned themselves on the "economic left" traditionally identified by researchers and which ones set up on the right? Alternatively, did they diverge in ways not captured by capitalist-socialist, equality-growth seeking, or statist-individualist dichotomies? How did these distinctions regarding economic issues map onto other ideological dimensions such as the one pitting Islamists against more secularist rivals?

Motivated in part by these questions, my dissertation focuses on why the post-uprising Tunisian party system seemed to offer starker choices regarding identity issues and muted those related to economic policy, despite many of the ingredients for the type of default left-right politics often assumed by political scientists. Part of the project uses text-as-data methods to understand and describe distinctions in how Tunisian politicians talk about economic problems. Using unsupervised methods of scaling political party platforms, as well as conducting interviews with their authors and other politicians, I show that, rather than staking out positions varying in their orientation to capitalism or socialism, Tunisian parties have instead generally competed through valence-based claims to competence.

I was initially drawn to text analysis because of several virtues over alternative methods of ideological scaling, such as surveys. Text analysis is cheaper, easier to apply retroactively, and perhaps less subject to reactivity—all major concerns for a young scholar from the United States studying events by then several years past. When I initially set out to understand the differences in the ways Tunisian politicians talked about economic issues, I imagined that I might be able to throw a hodgepodge of different types of text into R, sit back, and bask in the insight. I gathered Arabic and French Facebook posts, televised campaign statements, election posters, newspaper articles from party newspapers, and party platforms.

In practice, I have learned that the fundamental assumption of using text analysis for ideological scaling—that unobservable political differences can be discerned from observable patterns in word usage—requires more careful selection of texts. The problem is that word choice is not just a function of political attitudes or positions. Differences in word usage in one party's Facebook posts and another party's manifesto may have more to do with the differences in the purposes and constraints of these media than underlying differences in political attitudes. So if a researcher wants to use words to uncover

differences in political attitudes, it is helpful to select texts likely to express these differences but do not differ for other reasons. And if a researcher wants to understand differences regarding a particular dimension, such as transitional justice or public-welfare provision, then it helps to find texts focused on these topics.

> ***If a researcher wants to use words to uncover differences in political attitudes, it is helpful to select texts likely to express these differences but do not differ for other reasons.***

I began with party platforms for the same reasons that have drawn other scholars to them: They are long statements of political priorities and so they constitute large samples of words presumably drawn from some underlying political message. Party platforms enjoy somewhat of a tradition in Tunisia. Opposition parties published platforms at least as far back as the 1981 elections, and Ben Ali's RCD routinely published lengthy electoral platforms. After the revolution, most of the major parties published platforms and disseminated the messages in them through their Facebook pages, newspapers, campaign speeches, and other events.

Of course, platforms have well-known limits. They paper over intraparty differences and, because citizens generally do not read them, there are questions regarding the degree to which they reflect the messages citizens actually see and use to make choices. In my research, I have engaged with these limits in two ways. To nuance the role of platforms within Tunisian parties, I have conducted dozens of interviews with those who worked on platform-committees for Tunisia's largest parties. To assess the degree to which platform messages reflect those communicated to the public through other media, I am currently expanding the project to incorporate the analysis of nationally televised, subnational campaign videos.[125] These are just a few ways in which scholars can validate the importance and reception of the texts that they study.

There are multiple methods for using texts to ideologically scale actors, and the appropriate approach depends on the nature of the research problem. Broadly speaking, researchers can use dictionary-based, supervised, or unsupervised methods. A researcher wanting to place unfamiliar political actors in a debate with well-known poles and well-known terminology associated with them might create a dictionary, perhaps coding documents as falling on the left based on the number of times they include words such as distribution, state, or socialist. Alternatively, a researcher with a good idea of the sides in a debate but doubt about the lexicons associated with them might use a supervised method such as *wordscores*,[126] identifying reference texts to represent each pole, and then coding the remaining documents based on whether their word usage approximates that of one pole or the other. This method depends on the researcher to have a good understanding of the main dimension of difference. For example, in Tunisia, speeches by communists and Islamists might make good reference texts in a debate about inheritance-law reform but poor choices in a debate about transitional justice, where they arguably share interests as the formerly oppressed.

As my research is interested in not only who set up on the left and right but also the ideological content of this spectrum, I chose to use an unsupervised-scaling model called *wordfish*, developed by Slapin and Proksch.[127] *Wordfish* has been used to scale German party platforms,[128] Turkish party platforms,[129] speeches in Irish parliamentary debate,[130] press releases issued by U.S. Senators,[131] and other sets of text. It is based on item-response theory. The method makes a strong assumption that words are drawn stochastically from an underlying ideological message according to a Poisson distribution. The model includes fixed-effects to account for the fact that some documents are longer than others and some words are generally more common than others. It then gives the research two outputs with which to try to make sense of the underlying differences in the documents. Each word is assigned a score according to which it is used with different frequency across the spectrum; each document is assigned a score according to which

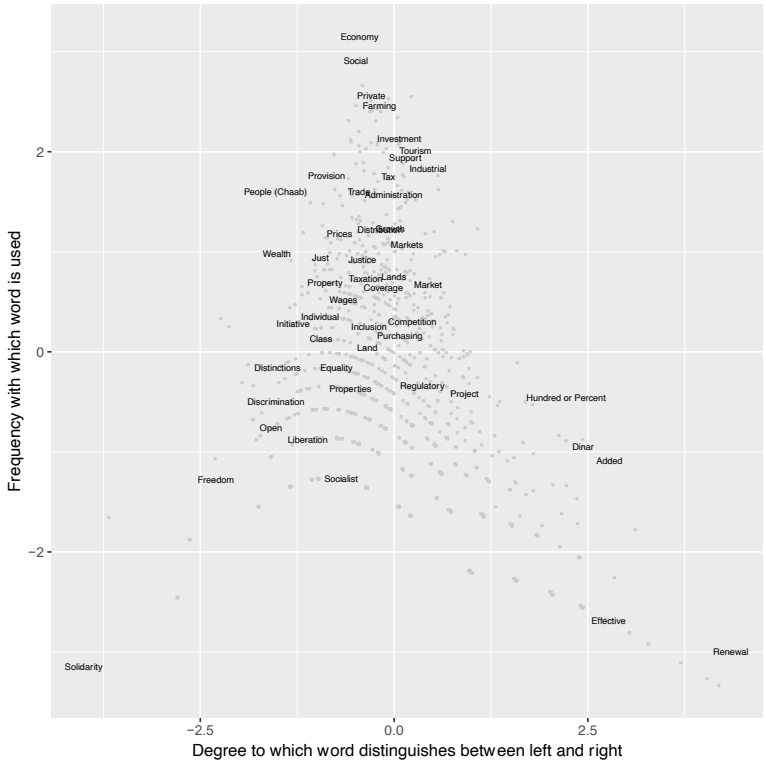
it uses words associated with one end or the other of that spectrum. This is where the hard work starts.

To give a concrete example, I applied the method to seven of the nine highest performing parties in the 2011 Tunisian Constituent Assembly elections.[132] According to the results, the Progressive Democratic Party (PDP) and the Democratic Modernist Pole (PDM) appeared on opposite ends of the spectrum, which we can call right and left. Ennahdha, the main Islamist party, and, to a lesser extent, Ettakatol and Congress for the Republic (CPR), appeared closer to the PDP on the right and the Popular Petition and the Communist Workers' Party (POCT) appeared closer to the PDM on the left. It is easy to construct a story explaining these results, with two of the parties on the left sharing Marxist-Leninist roots and most of the parties on the right having banded together in the prerevolutionary October 18 Collective.

But the really interesting findings usually lie not in the scale but the word scores according to which it is constructed. Here, it appears that what groups the parties on the left, the side of the spectrum with the two parties of Marxist-Leninist origin, is their focus on civil and political rights and issues related to the constitution—belonging, husband/wife, sex, discrimination, equality, and constitution, for example. Words associated with specific economic problems—gross, product, dinar, value-added, industry, etc.—distinguish the right.

The results of the analysis of the full platforms highlight a shortcoming with using *wordfish:* it assumes that the spectrum from which words are drawn is one-dimensional, but in the case of the 2011 elections, there were at least two main dimensions, one regarding the new constitution and the other regarding economic problems. To focus on distinctions with regard to economic problems, one must limit the analysis to the parts of the text focused on it. Although party platforms usually include section labels to facilitate this, slicing up the text in this way can raise questions if for example one party treats the inheritance law in its "cultural" section and another discusses it in its economic section.

In any case, applying the method to the economic sections of each platform yields a similar scaling, with the PDP on the right, the parties that later formed the Troika government in the center, and the PDM and POCT on the left. Below, I include a word pyramid showing the degree to which a set of words distinguish between the economic sections of the various parties' platforms.



What is striking in the word scores is not only that some parties seem to devote more attention to specific programs (numbers, percentage, dinars, and value-added are associated with the right) but also that many of the words one might expect to fall on the left or right do not. Words such as growth, distribution, support, and justice, for example, tell us little about whether a document will fall on the left or right. In my interview research, I have found that this absence of a clear distinction between statist and market-oriented economic approaches is reflective of a politics whereby leaders of the largest parties see their economic orientation as shared but their endowments of valence attributes—mainly competence and integrity—as the key distinguishing factor. For example, many of those who crafted the economic policies for Nidaa Tounes and Ennahdha in 2014 saw the two parties as sharing a general orientation toward economic problems but differing in the experience, knowledge, and integrity needed to address them.

Moving forward

The importance and competitiveness of electoral politics in Tunisia make it somewhat of a singularity in the region. But this does not mean that ideological scaling through study of text as data should be confined to the study of Tunisian politics. Where the main dimension of competition is unclear, where the lexicons associated with different ideological poles are unknown, or new political groups are entering the fray, ideological scaling through text analysis may be useful.

I argue that these will likely be widespread conditions in the coming years, particularly as many parts of the Arab world—including Algeria, Egypt, Iraq, Jordan, Lebanon, Morocco, and Sudan—grapple with pressures for contentious economic reform. In many world regions, the politics of welfare reform no longer resemble a conflict over expansion and retrenchment; if this is the case in these countries, as it was the case early in Tunisia's transition, then text-as-data methods could provide useful tools for understanding the contents of the conflicts over economic reform and how they map onto other types of political divisions.

Grubman notes:

[120] Ellen Lust and David Waldner, "Parties in Transitional Settings" in Nancy Bermeo and Deborah J. Yashar, *Parties, Movements, and Democracy in the Developing World* (Cambridge: Cambridge University Press, 2016).
[121] Tarek Masoud, *Counting Islam: Religion, Class, And Elections In Egypt* (Cambridge: Cambridge University Press, 2014) and Sharan Grewal, Amaney A. Jamal, Tarek Masoud, and Elizabeth R. Nugent, "Poverty and Divine Rewards: The Electoral Advantage of Islamist Political Parties," *American Journal of Political Science* (2019). This observation is puzzling given that prior to democratization these movements enjoyed a base primarily among the professional middle class. See Masoud (2014); Carrie Rosefsky Wickham, *Mobilizing Islam: Religion, Activism, and Political Change in Egypt* (New York: Columbia University Press, 2002); Janine A. Clark, *Islam, Charity, and Activism: Middle-Class Networks and Social Welfare in Egypt, Jordan, and Yemen* (Bloomington: Indiana University Press, 2004).
[122] This is a point of emphasis for Masoud 2014. In coining the term post-Islamism, Bayat similarly noted that the predominance of Islamist actors may coincide with a politics dominated by secularist concerns. See Asef Bayat, "The Coming of a Post-Islamist Society," *Critique* (Fall 1996): 43–52.
[123] Grewal et al 2019 and Eva Wegner and Francesco Cavatorta, "Revisiting the Islamist–Secular Divide: Parties and Voters in the Arab World," *International Political Science Review* 40, no. 4 (September 2019): 558–75.
[124] With regard to Tunisia, see Lindsay J. Benstead, Ellen Lust, and Dhafer Malouche, "Tunisian Post-Election Survey: Presentation of Initial Results." Transitional Governance Project, 2012; with regard to Egypt, see Masoud (2014).
[125] For me this turn to subnational platforms was inspired by Amy Catalinac, *Electoral Reform and National Security in Japan: From Pork to Foreign Policy* (Cambridge: Cambridge University Press, 2016).
[126] Michael Laver, Kenneth Benoit, and John Garry, "Extracting Policy Positions from Political Texts Using Words as Data," *American Political Science Review* 97 (2, 2003): 311–32.
[127] Jonathan Slapin and Sven-Oliver Proksch, "A Scaling Model for Estimating Time-Series Party Positions From Text," *American Journal of Political Science* 52 (3, 2008): 705–22.
[128] Slapin and Proksch 2008.
[129] Abdullah Aydogan and Jonathan B. Slapin. "Left–Right Reversed: Parties and Ideology in Modern Turkey." *Party Politics* 21 (4, 2015): 615-25.
[130] Will Lowe and Kenneth Benoit, "Validating Estimates of Latent Traits from Textual Data Using Human Judgment as a Benchmark," *Political Analysis* 21 (2013): 298–313.
[131] Justin Grimmer and Brandon M. Stewart, "Text as Data: The Promise and Pitfalls of Automatic Content Analysis Methods for Political Texts," *Political Analysis* (2013): 1–31.
[132] I omitted the platforms of the National Initiative of Kamel Morjane and Afek Tounes. The former seemingly did not publish a platform in 2011; the latter published a pithy Arabic document on the constitution and a lengthy French one focused on economic issues. Beforehand, I stemmed the texts using Richard Nielsen's R package (Richard A. Nielsen, *Deadly Clerics: Blocked Ambition and the Paths to Jihad* (Cambridge: Cambridge University Press, 2017). To fit the *wordfish* model, I used the Quanteda R package: Kenneth Benoit, Kohei Watanabe, Haiyan Wang, Paul Nulty, Adam Obeng, Stefan Müller, and Akitaka Matsuo, "Quanteda: An R Package for the Quantitative Analysis of Textual Data," *Journal of Open Source Software* 3 (30, 2018).

Nielsen notes:

[133] 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 percent 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
[134] An updated version of this figure appears in my book *Deadly Clerics* (2017) on page 122, along with more explanation of the method.
[135] Ibn Uthaymeen (d. 2001) was a prominent Salafi cleric from Saudi Arabia who did not write in support of jihadist ideology. This excerpt is from a short fatwa on prayer. Sayyid Qutb (d. 1966) was a prominent jihadist thinker from Egypt. This excerpt, from his famous work *Social Justice in Islam*, is not jihadist, so it does not get scored as jihadist by the classification model. Abdallah Azzam (d. 1989) was a prominent jihadist thinker who mentored Usama Bin Laden. This excerpt is from a treatise on jihad titled *In Defense of Muslim Lands*.
[136] http://www.mit.edu/~rnielsen/arabicTextWorkshop.zip

Nielsen 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.