TEXT AS DATA

APPLICATIONS OF AUTOMATED TEXT ANALYSIS IN THE MIDDLE EAST AND NORTH AFRICA

How do political parties in Tunisia present their economic platforms? How do Saudi political activists and their online followers change their social media behavior after arrest? How do Syrian state-owned media promote the political agenda of the state? These are just some of the types of questions that researchers are answering today using new text-as-data approaches.[55] Text-as-data applications have experienced a notable increase over the last decade as digitization of documents and the Internet make large corpora of texts more accessible and as greater computing power makes the processing of such texts more feasible. The study of Middle East politics is no exception.

We organized this symposium to highlight the new work being done using Arabic language and text-as-data methods, to address some of the risks and rewards of adopting these methods, and to familiarize the Arabic-language research community with what remains a relatively new methodological approach in comparative political science.[56]

The contributions included in this newsletter illustrate the wide variety of texts that can be analyzed using new computational methods: (1) **religious texts**[57] (2) **party platforms**[58] (3) **social media**[59] and (4) **traditional news media**.[60] These articles highlight how bodies of Arabic text can be analyzed to uncover new puzzles, measure key patterns of language usage, or make inferences about political behavior by key actors in the region. For example, Richard Nielsen discusses how male and female Salafi preachers appeal to different types of authority in their religious discourse.[61]

In addition to exploring various kinds of corpora for text analysis, each of the contributions details their methodological approaches and the challenges they faced. Broadly, these challenges include decisions about preprocessing the Arabic text data (e.g. dealing with distinguishing proper nouns from other words in the absence of capitalization and what types of stopwords to remove[62]) and decisions about how to analyze the data (e.g. using supervised or unsupervised methods).[63] Nathan Grubman's project on party ideology is an example of an unsupervised approach to ideological scaling, while Alexandra Siegel and Jennifer Pan's article on social media in Saudi Arabia uses supervised classification methods.[64]

Importantly, the authors each illustrate the way that a deep understanding of the region and the texts is pivotal to using text as data methods. Text analysis points to the importance of language learning, deep thinking about the meaning of words, and recognizing the limits of automated methods. Once a researcher has developed her own understanding of the purposeful and delicate ways language is used, she can make decisions about how to process the text, what type of approach to use, or how to work with research assistants to classify texts. The authors included in this symposium reflect on engaging in that process in their own research.

Close knowledge of the region also helps researchers to think carefully about the ethical issues associated with using text-as-data methods. Ala' Alrababa'h discusses intellectual property issues and the ethical concerns around ensuring that online access to the newspapers is not disrupted for other readers when researchers are scraping the websites. And Alex Siegel addresses the debates around highlighting an online individual's social media activity in repressive settings. Looking ahead these ethical issues should remain at the forefront of researchers' discussions.

Looking Ahead

If the recent uptick in social science research using text-as-data methods with Arabic language is any indication, these approaches will continue to develop and grow in political science. Work using computational text analysis methods in other languages offers some ideas about possible new avenues for research using these tools with Arabic, including: (1) **literature, political theory texts, and textbooks**[65] (2) **candidate platforms and manifestos**[66] (3) **open-ended survey questions, interview transcripts, or personal narratives**[67] (4) **political speeches and press releases**[68] and (5) **diplomatic records, judicial decisions, and other government documents**.[69]

Furthermore, from a methodological perspective, it would be interesting to see more work exploring technical aspects of Arabic text analysis. For instance, should analysis of Arabic traditional and social media incorporate French texts, particularly in former French colonies, and how consequential is that decision?[70] Do the results of automated text analysis of Arabic differ after stemming versus lemmatization? Because Arabic relies on a strong root system, lemmatization could also be a powerful way to analyze the language.

Despite advances in Google Translate, automated language translation from Arabic performs more poorly than other major languages (and does not translate colloquial Arabic), and Arabic language remains underrepresented in artificial intelligence applications more broadly.[71] Based on my own experience working in Arabic, for these reasons, automated translation from Arabic often fails to capture the meaning of a phrase or the correct translation of specific words within a given phrase. Thus, in analyzing computer-translated Arabic texts, the bag-of-words assumption could be violated.[72] The works highlighted in this newsletter do not use automated language translation, and, in the short to medium term, that is likely to remain the gold standard for Arabic automated text analysis.

Some Arabic Text Analysis Resources

Rich Nielsen, one of the contributors to this newsletter, has developed stemmers for both Arabic (Arabic Stemmer) and Persian (Persian Stemmer). Additionally, there are many resources that have been developed outside of political science with applications to Arabic text-as-data analysis. For instance, the Computational Approaches to Modeling Language (CAMeL) Lab is a research lab at New York University Abu Dhabi focused on Arabic language analysis, and the Stanford University Natural Language Processing Group has developed software that can also process Arabic language texts (Stanford CoreNLP).

As highlighted by this collection of essays on recent research that employs Arabic text as data methods, there are many research questions - both new and old - to which these methods can contribute. We hope that by featuring this work, we provoke further discussion around the promise and pitfalls of these methods, particularly as it relates to Arabic language texts, and encourage scholars to familiarize themselves with these methods, if not add these tools to their repertoire.

– **Alexandra Blackman, New York University - Abu Dhabi**

[31] There are two other known biases in the Algerian Facebook population, but these are correctable. The first is gender: men represent 50.6% of the population, but 64% of Facebook users.[31] Second, Facebook users tend to be younger than average: 64% of the overall population are less than 35, but 76% of Algerian Facebook users are less than 35. We corrected for age and gender biases by creating separate Facebook advertisements for each age-gender demographic (i.e., women aged 25-34). We then increased the number of ads shown to demographic groups under-represented on Facebook, such as older women, in order to create a more balanced sample.

[32] For a list of those arrested, see: https://www.tsa-algerie.com/les-personnalites-mises-en-detention-depuis-le-depart-de-bouteflika/

[33] See Bruce Riedel, "Unveiling Algeria's Dark Side: The Fall of the Butcher of Algiers," Brookings, May 8, 2019. https://www.brookings.edu/blog/order-from-chaos/2019/05/08/unveiling-algerias-dark-side/

[34] See https://www.france24.com/en/20190802-algeria-protest-civil-disobedience.

Buehler notes:

[35] Kamrava, Mehran. 2014. *The Nuclear Question in the Middle East*. London: Hurst. 2.

[36] Nacir, B. 2010. "Moroccan Triga Mark II Research Reactor Utilization." Centre Nationale de l'Energie des Sciences et des Techniques Nucléaires.

[37] Rosatom. 2016. "Tunisia and Russia signed an Inter-governmental Agreement on Peaceful Uses of Atomic Energy," September 26, 2016. https://www.rosatom.ru/en/press-centre/news/tunisia-and-russia-signed-an-intergovernmental-agreement-on-peaceful-uses-of-atomic-energy/

[38] Ebinger, Charles, John Banks, Kevin Masssy, and Govinda Avasarala. 2011. "Models for Aspirant Civil Nuclear Energy Nations in the Middle East." *Brookings Institution*. 55-57.

[39] For an excellent survey of Morocco's contemporary civilian nuclear program, see: Adamson, Matthew. 2017. "Peut-on faire une histoire nucléaire du Maroc? Le Maroc, l'Afrique et l'énergie nucléaire," *Afrique contemporaine* n. 261-262. 94-97. Also, for a fascinating study of Morocco's nuclear politics in the 1950s, see: Adamson, Matthew. 2017. "The Secret Search for Uranium in Cold War Morocco," *Physics Today*. 55-60.

[40] Rosa, Eugene A. and Riley E. Dunlap. 1994. "Nuclear Power: Three Decades of Public Opinion," *Public Opinion Quarterly*v58: 295-325; de Groot, Judith I.M. 2013. "Values, Perceived Risks and Benefits, and Acceptability of Nuclear Energy," *Risk Analysis* 33(2)..

[41] Schwedler, Jillian. 2014. "Jordan's Nuclear Project is Bound to Fail." *Middle East Report*; *Nicholas Seeley,* "The Battle Over Nuclear Jordan," *Middle East Report* (2014)

[42] Inglehart, Ronald. *Modernization and Postmodernization: Cultural Economic, and Political Change in 43 Societies* (Princeton: Princeton University Press, 1997).

[43] Cohen, Ibid. 34-55.

[44] Rost Rublee, Maria. 2009. *Nonproliferation Norms: Why States Choose Nuclear Restraint*. Athens: Georgia University Press. 109-118

[45] Ebinger et al. Ibid., 38

[46] Samina Ahmed. 1999. "Pakistan's Nuclear Weapons Program: Turning Points and Nuclear Choices" *International Security* 23(4).

[47] Parsi, Trita. 2017. *Losing an Enemy: Obama, Iran, and the Triumph of Diplomacy*. New Haven: Yale University Press.

[48] Barzegar, Kayhan. "Iran's Nuclear Program" in *The Nuclear Question in the Middle East*, ed. by Mehran Kamrava, 225-264. London: Hurst 2014.

[49] Landau, Ibid. 102.

[50] Ahmad, Ali & M.V. Ramana. 2014. "Too Costly to Matter: Economics of Nuclear Power for Saudi Arabia." *Energy*; Luomi, Ibid. 125-158.

[51] Fuhrmann, Matthew. 2009. "Spreading Temptation: Proliferation and Peaceful Nuclear Cooperation Agreements." *International Security* 34(1): 8-10. Fuhrmann, Matthew and Benjamin Tkach. 2015. "Almost Nuclear: Introducing the Nuclear Latency Dataset." *Conflict Management and Peace Science* 32(4): 443-461..

[52] Tucker, Jonathan. 1993. "Monitoring and Verification in a Non-cooperative Environment: Lesson from the U.N. Experience in Iraq." *The Nonproliferation Review.*

[53] World Nuclear News, "South Korea and UAE seek cooperation beyond Barakah." February 27, 2019. https://www.world-nuclear-news.org/Articles/South-Korea-and-UAE-seek-cooperation-beyond-Baraka

[54] Ahmad, Ali. 2015. "Economic Risks of Jordan's Nuclear Program," *Energy for Sustainable Development* (29): 34.

Blackman notes:

[55] Ala' Alrababa'h and Lisa Blaydes. "Authoritarian Media and Diversionary Threats: Lessons from Thirty Years of Syrian State Discourse." *Working Paper* (2019); Nathan Grubman, "Ideological Scaling in a Neoliberal, Post-Islamist Age," *APSA Middle East Politics Newsletter* (2019); Jennifer Pan and Alexandra Siegel, "How Saudi Crackdowns Fail to Silence Online Dissent," *Working Paper* (2019).

[56] Christopher Lucas, 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.

[57] Richard Nielsen, *Deadly Clerics: Blocked Ambition and the Paths to Jihad* (Cambridge: Cambridge University Press, 2017).

[58] Grubman 2019.

[59] Pan and Siegel 2019; Alexandra Siegel, "Using Social Media Data to Study Arab Politics," *APSA Middle East Politics Newsletter* (2019); Amaney A. Jamal, Robert O. Keohane, David Romney, and Dustin Tingley, "Anti-Americanism and Anti-Interventionism in Arabic Twitter Discourses," *Perspectives on Politics* 13, no. 1 (2015): 55-73.

[60] Ala' Alrababa'h, "Quantitative text analysis of Arabic news media," *APSA Middle East Politics Newsletter* (2019); Alrababa'h and Blaydes 2019. For a relevant application using U.S. news media, see: Rochelle Terman, "Islamophobia and Media Portrayals of Muslim Women: A Computational Text Analysis of US News Coverage," *International Studies Quarterly* 61, no. 3 (2017): 489-502.

[61] Richard Nielsen, "What Counting Words Can Teach Us About Middle East Politics," *APSA Middle East Politics Newsletter* (2019); Richard Nielsen, "Women's Authority in Patriarchal Social Movements: The Case of Female Salafi Preachers," *American Journal of Political Science* (2019).

[62] For more details on preprocessing steps, see: Matthew J. Denny and Arthur Spirling. "Text Preprocessing For Unsupervised Learning: Why It Matters, When It Misleads, And What To Do About It." *Political Analysis* 26, no. 2 (2018): 168-89. Arabic texts can present a challenge because the same sequence of characters can

contain different parts of speech and have an entirely different meaning as a result of how words are connected in Arabic. For example, the sequence وجهته can mean "and his/its side" or "I/She sent him/it [in the direction of]" and can be segmented into entirely different parts of speech.

[63] For an overview of text as data methods, see: Justin Grimmer and Brandon M. Stewart, "Text as Data: The Promise and Pitfalls of Automatic Content Analysis Methods for Political Texts," *Political Analysis* 21, no. 3 (2013): 267-97. Supervised methods include the use of human coders and thus rely on the researcher to know *ex-ante* what to code for in the data. Unsupervised methods generate the topics from the data but require researcher decisions about issues such as the number of topics to generate.

[64] Siegel and Pan's research is discussed in Alexandra Siegel's contribution to this newsletter.

[65] Lisa Blaydes, Justin Grimmer, and Alison McQueen. "Mirrors for Princes and Sultans: Advice on the Art of Governance in the Medieval Christian and Islamic Worlds," *Journal of Politics* 80, no. 4 (2018): 1150-1167; Jennifer A. London, "Re-imagining the Cambridge School in the Age of Digital Humanities," *Annual Review of Political Science* 19, no. 1 (2016): 351-373; Tamar Mitts, "Terrorism and the Rise of Right-Wing Content in Israeli Books," *International Organization* 73, no. 1 (2019): 203-24.

[66] Amy Catalinac, "From Pork to Policy: The Rise of Programmatic Campaigning in Japanese Elections," *The Journal of Politics* 78, no. 1 (2016): 1-18.

[67] Mathilde Emeriau, "Learning to be Unbiased: Evidence from the French Asylum Office," *Working Paper* (2019); Margaret E. Roberts, Brandon M. Stewart, Dustin Tingley, Christopher Lucas, Jetson Leder-Luis, Shana Kushner Gadarian, Bethany Albertson, David G. Rand, "Structural Topic Models for Open-Ended Survey Responses," *American Journal of Political Science* 58, no. 4 (2014): 1064-1082.

[68] Justin Grimmer, *Representational Style in Congress: What Legislators Say and Why It Matters* (Cambridge: Cambridge University Press, 2013).

[69] Azusa Katagiri and Eric Min, "The Credibility of Public and Private Signals: A Document-Based Approach," *American Political Science Review* 113, no. 1 (2019): 156-72; Benjamin Liebman, Margaret Roberts, Rachel Stern, and Alice Wang, "Mass Digitization of Chinese Court Decisions: How to Use Text as Data in the Field of Chinese Law," *21st Century China Center Research Paper* no. 2017-01 (2017).

[70] Another interesting question is how the choice of language in social media corresponds to social class and other individual characteristics? Fred Schaffer's ethnographic study of conceptions of democracy in Senegal suggests that these language choices are closely related to class and have important implications for how people understand and engage with important political concepts like democracy. See: Frederic C. Schaffer, *Democracy in Translation: Understanding Politics in an Unfamiliar Culture* (Ithaca: Cornell University Press, 1998).

[71] Freya Pratty, "Arabic and AI: Why voice-activated tech struggles in the Middle East," *Middle East Eye*, September 10, 2019. Examining Danish, German, Spanish, French and Polish, de Vries, Schoonvelde, and Schumacher make the case for Google Translate. However, I have not found a comparable analysis of the performance of Google Translate for Arabic. See: Erik de Vries, Martijn Schoonvelde, and Gijs Schumacher, "No Longer Lost in Translation: Evidence That Google Translate Works for Comparative Bag-of-Words Text Applications," *Political Analysis* 26, no. 4 (2018): 417-30.

[72] It would be interesting to see applications and evaluations of recent advances in language models, such as word embeddings, on Arabic texts. For an application in political science, see: Yaoyao Dai, "Measuring Populism in Context: A Supervised Approach with Word Embedding Models," *Working Paper* (2019).

Siegel notes:

[73] Salem, Fadi. "The Arab social media report 2017: Social media and the internet of things: Towards data-driven policymaking in the Arab World (Vol. 7)." *Dubai: MBR School of Government* (2017).

[74] Zeitzoff, Thomas. "Using social media to measure conflict dynamics: An application to the 2008–2009 Gaza conflict." Journal of Conflict Resolution 55, no. 6 (2011): 938-969.

[75] Kubinec, Robert, and John Owen. "When Groups Fall Apart: Measuring Transnational Polarization with Twitter from the Arab Uprisings." Unpublished Manuscript (2018).

[76] Weber, Ingmar, Venkata R. Kiran Garimella, and Alaa Batayneh. "Secular vs. islamist polarization in egypt on twitter." In Proceedings of the 2013 IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining, pp. 290-297. ACM, 2013; Lynch, Marc, Deen Freelon, and Sean Aday. "Online clustering, fear and uncertainty in Egypt's transition." Democratization 24, no. 6 (2017): 1159-1177; Siegel, Alexandra, Joshua Tucker, Jonathan Nagler, and Richard Bonneau. "Tweeting beyond Tahrir: Ideological diversity and political tolerance in Egyptian twitter networks." Unpublished Manuscript (2019).

[77] Siegel, Alexandra. *Sectarian Twitter Wars: Sunni-Shia Conflict and Cooperation in the Digital Age*. Vol. 20. Carnegie Endowment for International Peace, 2015.; Siegel, Alexandra, Joshua Tucker, Jonathan Nagler, and Richard Bonneau. "Socially Mediated Sectarianism." *Unpublished Manuscript.* (2018).

[78] Pan, Jennifer and Siegel, Alexandra A. "How Saudi Crackdowns Fail to Silence Online Dissent." Unpublished Manuscript. 2019.

[79] For example: Howard, Philip N., and Muzammil M. Hussain. "The upheavals in Egypt and Tunisia: The role of digital media." *Journal of democracy* 22, no. 3 (2011): 35-48; Howard, Philip N., Aiden Duffy, Deen Freelon, Muzammil M. Hussain, Will Mari, and Marwa Maziad. "Opening closed regimes: what was the role of social media during the Arab Spring?." *Available at SSRN 2595096* (2011).

[80] See Smidi, Adam, and Saif Shahin. "Social Media and Social Mobilisation in the Middle East: A Survey of Research on the Arab Spring." India Quarterly 73, no. 2 (2017): 196-209. for an overview and Aday, Sean, Henry Farrell, Marc Lynch, John Sides, and Deen Freelon. "New media and conflict after the Arab Spring." *United States Institute of Peace* 80 (2012): 1-24. for an example of empirical evidence using Twitter data to make this argument.

[81] For an overview of this debate and the empirical evidence on both sides, see: Tucker, Joshua A., Jonathan Nagler, Megan MacDuffee, Pablo Barbera Metzger, Duncan Penfold-Brown, and Richard Bonneau. "Big data, social media, and protest." *Computational Social Science* 199 (2016).

[82] Lynch, Marc ; Freelon, Deen and Aday, Sean. 2014. *Syria's Socially Mediated Civil War*. United States Institute of Peace.

[83] Starbird, Kate, Ahmer Arif, Tom Wilson, Katherine Van Koevering, Katya Yefimova, and Daniel Scarnecchia. "Ecosystem or echo-system? Exploring content sharing across alternative media