Political Communication Report
Spring 2023 - Issue 27
"New Methodological Diversity in PolComm"

http://dx.doi.org/10.17169/refubium-39046

Political Language and the Computational Turn

Dr. Josephine Lukito, University of Texas at Austin

When I was asked to write this essay, my first step was to do something increasingly common for natural language processing scholars: see how Chat-GPT would write it. My prompt: "Write a 2000-word academic essay on the benefits and disadvantages of using natural language processing to study political language. The paper should discuss how NLP has been used, why it helps research, and its limitations." You can find the essay, written in less than a minute (43.73 seconds, to be precise), here.

By comparison, the essay you will be reading took a few weeks to plan and several days to write. It is inspired by a decade's worth of conversations I have had with colleagues and friends about the "ease" of using computational methods to study political language. It was substantively more labor intensive, and, I admit, some of the points I make are similar to those made by the chat-GPT article.

Was it worth the effort? Of course. For one, trying to pass off a chat-GPT article as your own constitutes plagiarism. And one would hope my writing is more entertaining and novel than output from a generative AI. But, most importantly, language models (even very large ones) lack the ability to fully understand or recreate the unpredictability of natural language. This is partly because these models cannot understand social context without human intervention. However, natural language is also flawed and inconsistent (like its creators), and even unexpected mistakes (like "covfefe") can enter a language's lexicon.

We should see language's unpredictability as, generally, a good thing. Language is an ever-evolving social system and people's constant re-shaping of language is necessary for societal development. This is especially true for political language, which is used by citizens, activists, journalists, and public figures to deliberate, argue, and persuade. In doing so, political language is constantly shaping, and shaped by, the people who use it.

And yet, the growing production of political language, especially online, creates new opportunities for both political activism and harassment. To study this "at scale" (at the size required to understand the scope of the problem), computational methods can be helpful to identify potentially meaningful patterns. These approaches, known as natural language processing (NLP) or text-as-data (TAD) tools, have been used to study a wide range of spoken and written political language, from social media content to broadcasted debates.

While helpful, it is important that political communication scholars employing NLP or TAD methods be mindful of both the limitations of these methods, and the consequences surrounding how these methods are applied. The problem with computational tools has never been the tools themselves. Rather, it is human trust in human-constructed tools. In other words: computational tools are useful for studying political language insofar as we do not become reliant on these tools for interpretation or decision-making.

This is most clearly noticeable with the "AI hype," or the exaggerated perception of artificial intelligence and machine learning as either the savior or the downfall of societies. These claims often evoke a sense of technological determinism, and sometimes remove the human agency from the process. Many have asked, "is AI good or bad for society?" But even the way this question is posed obfuscates the human by placing the noun phrase "artificial intelligence" in the subject position and not mentioning people at all. This hype, regardless of whether one sees computational tools as good or bad, is flawed in two ways. First, the success of tools such as text classifiers and language generation will never be perfect. As with any human-built system, there will be mistakes. Second, there is the belief that the tool can be structured to prevent harm (as suggested by the advocacy letter, "Pause Giant AI Experiments: An Open Letter"). The truth is that any human tool can be used for societally harmful *and* beneficial purposes.

So, knowing this, how should political communication scholars use NLP tools? I argue that a cyborgian approach is necessary—one that leverages computational pattern recognition with human interpretation, such that the sum of its cumulative labor is greater than the individual parts. In particular, I make three recommendations for political communication researchers seeking to use NLP in their work: incorporate linguistic theory, validate your approaches, and acknowledge the normative underpinnings of your scholarship.

First, research should combine our field's rich tradition of studying political languages (e.g., Edelman 2013) not only with computational methods, but also with the linguistics literature, which is far more specific regarding its assessment of political language. With a stronger understanding of language structure, political communication researchers would be able to better leverage NLP tools such as dependency parsing (e.g., Borah et al., 2013), which require some knowledge about syntax to effectively use. Similarly, while text-as-data remains popular, there is a growing interest in multi-modal communication, and NLP tools for spoken language,

like Parselmouth (Jadoul et al., 2018), can create new ways to study spoken rhetoric and political discourse at scale.

One way to do this is to consider a layered approach to pre-processing language data. There is already precedence to this in linguistics, which includes the following subfields: phonology, or the study of how humans combine sounds ("phonemes") in language; morphology, which studies how words are constructed (using morphemes like prefixes and suffixes); syntax, which studies how words are combined into phrases or sentences, and semantics. In political communication, it is particularly common to study the semantics of individual words or phrases, (i.e., the lexicon). (Other subfields, such as sociolinguistics and cognitive linguistics, are also relevant, but the aforementioned four address the forms within a language system [see Kastovsky, 1977].) When pre-processing language data for computational analyses, researchers can add information to, or reduce, these layers (phonological, morphological, lexical, and syntactic), as noted in Table 1.

Table 1: Linguistic Layers of Computational Political Language Processing

	Unit of Analysis	**Reductive**	**Additive**
Phonological	Phoneme	Text / Transcript	Pitch, Tone, Prosody Notation
Morphological	Morpheme	Lemmatization	POS Tagging
Lexical (Semantic)	Word/Lemma	Stop Words	Tokenizing, Word Lists
Syntactic	Sentence	Bag of Words	Dependency, Clausal Analysis, Word Embeddings*

* Word embeddings are not a full annotation of syntax, but it does retain critical word-order information.

Second, as many scholars have argued, there is a need to compare and validate different computational approaches to language analyses (Van Atteveldt et al., 2021; Muddiman et al., 2019). One potential method for comparing classifiers would be to use benchmark datasets, a common strategy for validating text classifiers in computer science and engineering (e.g., Su et al., 2020) alongside novel datasets. This can be especially useful for content that is otherwise difficult to access, such as mis/disinformation content. Additionally, mixed-methods work with a closer, qualitative examination of language features can help inform a researcher's NLP approach (Lukito & Pruden, 2023).

An important part of this validation process is the need for political communication scholars to develop a humanistic, ethical approach to using natural language processing. This includes advocating for both data ethics (Lazer et al., 2020) and data access for research (EDMO, 2022). More tangible tasks include encouraging ethical statements in research papers, creating norms

for anonymization when sharing data, and advocating for policies that support data transparency and independent research. We can already see the start of these efforts through the Coalition for Independent Technology Research (CITR), the Media and Democracy Data Cooperative (MDDC), and the Social Media Archive at ICPSR (SOMAR).

Another key consideration for NLP validation is acknowledging the limits of one's study. For example, text classifiers may be good at aggregating trends, but it has a non-inconsequential chance of making an error for an individual case. Similarly, language generation outputs can help scholars avoid the dreaded blank sheet problem (Evans, 2013), but these tools are much more suited to scripted, systematic conversations (like those between customer service and customer) and must be tested, modified, and validated when using them to study political communication and the human experience.

And finally, political communication scholars should acknowledge the normative underpinnings of any research studying political language. Computational work has been described as more "objective" (Singh & Glińska-Neweś, 2021), but each step of the NLP analysis process—from the pre-processing of stopwords to the interpretation of a semantic network or a text classifier—is subjective. Different decisions can change the results of an analysis; changing and potentially improving on how machines interpret natural language (Haddi, Liu & Shi, 2013). And decisions made by a researcher are not made in a vacuum. In conducting their work, researchers bring their own experiences with political communication (academically and interpersonally) to their work. Rather than shying away from this, researchers should embrace normative commitments and be upfront about the goals of their work.

Though some forms of natural language processing have existed since the 1950's (Kumar, 2013), their use in political communication remains relatively nascent. These newly developed methods can help researchers advance more democratic and inclusive societies that empower citizens and shape governance to benefit the many rather than the few. But in order to do so, researchers must consider the limitations of these methods, avoid the AI hype, and play an active role in the interpretation of the data. Because, at the end of the day, it is not about how novel or sophisticated your language model is. What matters is what you plan to achieve with it.

References

Borah, P., Ghosh, S., Hwang, J., Shah, D. V., & Brauer, M. (2023). Red Media vs. Blue Media: Social Distancing and Partisan News Media Use during the COVID-19 Pandemic. *Health Communication*, 1-11.

Edelman, M. (2013). *Political language: Words that succeed and policies that fail*. Elsevier.

European Digital Media Observatory (EDMO). (2022, May 31) Report of the European Digital Media Observatory's Working Group on Platform-to-Researcher Data Access. https://edmoprod.wpengine.com/wp-content/uploads/2022/02/Report-of-the-European-Digital-Media-Observatorys-Working-Group-on-Platform-to-Researcher-Data-Access-2022.pdf

Evans, K. (2013). *Pathways through writing blocks in the academic environment*. Springer Science & Business Media.

Haddi, E., Liu, X., & Shi, Y. (2013). The role of text pre-processing in sentiment analysis. *Procedia computer science*, *17*, 26-32.

Kastovsky, D. (1977). Word-formation, or: At the crossroads of morphology, syntax, semantics, and the lexicon.

Kumar, E. (2013). *Natural language processing*. IK International Pvt Ltd.

Jadoul, Y., Thompson, B., & De Boer, B. (2018). Introducing parselmouth: A python interface to praat. *Journal of Phonetics*, *71*, 1-15.

Lazer, D. M., Pentland, A., Watts, D. J., Aral, S., Athey, S., Contractor, N., ... & Wagner, C. (2020). Computational social science: Obstacles and opportunities. *Science*, *369*(6507), 1060-1062.

Lukito, J., & Pruden, M. L. (2023). Critical computation: mixed-methods approaches to big language data analysis. *Review of Communication*, *23*(1), 62-78.

Muddiman, Ashley, Shannon C. McGregor, and Natalie Jomini Stroud. "(Re) claiming our expertise: Parsing large text corpora with manually validated and organic dictionaries." *Political Communication* 36, no. 2 (2019): 214-226.

Singh, A., & Glińska-Neweś, A. (2022). Modeling the public attitude towards organic foods: A big data and text mining approach. *Journal of big Data*, *9*(1), 1-21.

Su, Q., Wan, M., Liu, X., & Huang, C. R. (2020). Motivations, methods and metrics of misinformation detection: an NLP perspective. *Natural Language Processing Research*, *1*(1-2), 1-13.

Tolochko, P., & Boomgaarden, H. G. (2019). Determining political text complexity: Conceptualizations, measurements, and application. *International Journal of Communication*, *13*, 21.

Van Atteveldt, W., Van der Velden, M. A., & Boukes, M. (2021). The validity of sentiment analysis: Comparing manual annotation, crowd-coding, dictionary approaches, and machine learning algorithms. *Communication Methods and Measures*, *15*(2), 121-140.

Josephine ("Jo") Lukito is an Assistant Professor at the University of Texas at Austin's School of Journalism and Media, Director of the Media & Democracy Data Cooperative, and a Senior Faculty Research Affiliate for the Center for Media Engagement. She uses computational linguistics and mixed methods to study multi-platform flows of political discourse.