

I develop computational approaches to analyze nuanced rhetorical strategies in sociopolitical discussions, and highlight the broader societal implications of these linguistic choices. By synthesizing natural language processing (NLP), linguistics, political communication, and psychology, my deeply interdisciplinary research program has largely focused on computationally modeling political framing [1, 2, 3, 4], implicitly harmful language [4, 5, 6], and linguistic variation on social media [7, 8, 9]. I am particularly interested in understanding linguistic mechanisms used to uplift or ostracize members of marginalized communities on social media, and understanding the risks of such biases for modern language technology systems. Ultimately, I endeavor to use NLP to promote social justice and make the online world a safer and more welcoming place for everybody.

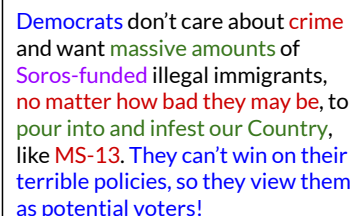
Computational linguistic approaches to political speech typically focus on broad characterizations of a text’s stance and sentiment. From that lens, the post in Figure 1 is straightforward: it expresses an anti-immigration stance with negative sentiment. Beneath the surface, however, the author uses subtle rhetorical strategies to communicate their perspective and influence their readers. They *frame* immigration by drawing connections to **Democrats’ political ambitions** and **criminal activity**. They *dehumanize* immigrants by likening them to water and vermin with phrases such as “**massive amounts**” and “**pour into and infest our Country**”. Lastly, they use the antisemitic *dogwhistle* “**Soros-funded**”, which covertly links immigration to conspiracy theories about Jewish plots for world domination. These linguistic mechanisms can be particularly insidious because they can shape how audiences fundamentally understand and think about political issues. By focusing on strategies such as framing, dehumanization, and dogwhistle communication, my research facilitates a deeper understanding of political language and its societal impact.

Modeling framing in political discourse

Framing, the process of selecting and emphasizing particular aspects of political issues, is a key mechanism by which a text influences its audience; it shapes how readers evaluate political problems and potential solutions, thus having major implications for public opinion and policy. However, surprisingly little is known about how social media users frame political issues, despite a majority of the U.S. population getting their news from social media.

I addressed this gap by introducing a computational methodology grounded in political communication to analyze how the public produces and responds to framing in tweets about immigration [1]. I created a novel codebook, annotated dataset, and neural text classifiers to detect frames from three typologies: issue-generic (topic-like categories that could generalize across issues), immigration-specific (centered around representations of immigrants as heroes, threats, and victims), and narrative frames (whether messages focus on specific episodes or the broader societal context). By automatically detecting frames for 2.6M tweets, I demonstrate how Twitter users’ identities, particularly political ideology and region, are associated with different framing choices. Conservatives tend to frame immigrants as threats to public safety and a burden on taxpayers, while liberals tend to frame immigrants as heroes or victims (Figure 2). Moreover, I assess how audiences engage with different frames, finding that human interest-related frames receive the most favorites, while safety and security frames are the most amplified through retweets.

This work establishes two key takeaways for computational framing research. First, this was the first computational study to consider narrative and immigration-specific frames. I show that these typologies reveal meaningful patterns that are otherwise obscured by issue-generic frames, which are predominant in NLP, thus highlighting the importance of deeply engaging with social science



Democrats don't care about crime and want massive amounts of Soros-funded illegal immigrants, no matter how bad they may be, to pour into and infest our Country, like MS-13. They can't win on their terrible policies, so they view them as potential voters!

Figure 1: Example social media post about immigration

scholarship to guide modeling decisions. In later work published in PNAS, my colleagues and I adapt this perspective of immigrants as heroes, threats, and victims to study how the framing of immigration has changed over 140 years of U.S. Congressional speeches [4].

Second, I move beyond frame detection as the primary objective. Rather, **I treat framing as a process that is closely intertwined with speakers’ social identities, their audiences, and their political environment.** For example, in studying information manipulation in Russian media [2], I uncovered frame variation between posts on Twitter and VKontakte (a Russian social media platform), suggesting that media outlets tailor their framing based on whether their audience is primarily Russian or international. I expanded upon these two key takeaways in a recent study of online social movement framing [3]. Grounded in sociological theories of movement mobilization, I focus on how framing accomplishes core functions of identifying a problem, proposing solutions, and motivating audience members to participate in collective action, and show that activists attend to these functions differently across sociocultural and conversational contexts. Through applications to these domains, my research spotlights the role of language in shaping political processes and public perception.

Computational analysis of implicitly harmful language

Dehumanization, “the act of perceiving or treating people as less than human” [10], is a pernicious psychological process that leads to extreme intergroup bias, hate speech, and violence against marginalized groups. Dehumanization is difficult to detect because it is typically communicated implicitly through subtle linguistic cues (e.g. rather than directly comparing people to animals, describing people with adjectives and verbs that are associated with animals). Social psychologists have identified varied dimensions of dehumanization, such as moral disgust, denial of agency, and likening members of a target group to animals or machines. Drawing upon this literature, **I proposed a framework for analyzing dehumanization by identifying linguistic correlates for each dimension that we can measure with computational techniques** [5]. For example, associations with vermin could be measured by comparing of target group word embeddings with a vermin concept vector representation. More recently, my colleagues and I developed updated methods for quantifying dehumanization using masked language model predictions [4].

I applied this framework to explore evolving representations of LGBTQ groups in the *New York Times* over thirty years (1986-2015). Overall, I found increasingly humanizing descriptions of LGBTQ people over time. Moreover, my methodology captured changes in social meaning on a large scale, revealing that the label *homosexual* remained more strongly associated with dehumanization than other semantically-similar labels such as *gay*. My work on dehumanization, published in the Frontiers Special Issue on Computational Sociolinguistics, has been taught in linguistics, NLP, and computational ethics courses at the University of Washington, Georgia Tech, Northwestern, and Carnegie Mellon. It has had a significant impact in social psychology [11] and NLP: dehumanization has come to be recognized as a key challenge for hate speech detection [12, 13] and an important consideration in evaluating potential harms of large language models [14, 15].

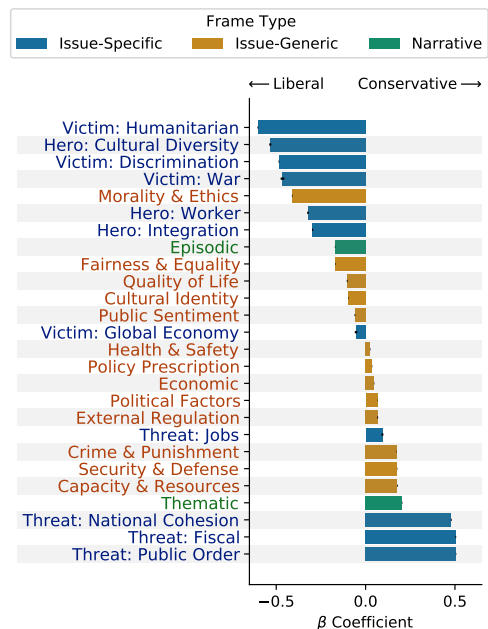


Figure 2: Associations between ideology and framing [1]. Positive (negative) values correspond to more conservative (liberal) ideology.

Another linguistic strategy to implicitly communicate hateful attitudes is through **dogwhistles**, coded expressions that simultaneously convey one meaning to a general audience and a second covert meaning that is only recognizable to a smaller in-group audience [16]. For example, in the sentence “we need to end the cosmopolitan experiment”, the word “cosmopolitan” likely means “worldly” to many, but secretly means “Jewish” to a select few. Dogwhistles are powerful mechanisms of political influence and are often deployed online to evade automatic content moderation. Dogwhistle research is thus essential across disciplines, but remains a challenge. Unless they are part of an in-group, researchers may be completely unaware of a dogwhistle’s existence. Moreover, unlike overtly hateful or toxic language, dogwhistles’ meanings cannot be determined by form alone, but rather their interpretation relies on a complex interplay of factors such as speaker and audience identities and conversational contexts.

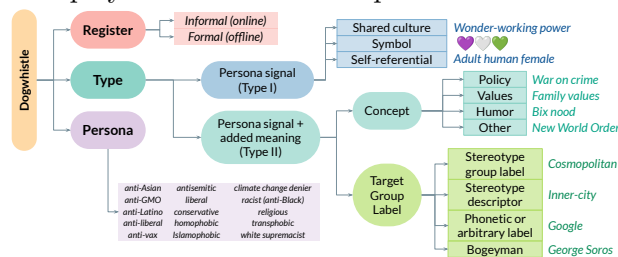


Figure 4: Dogwhistle typology developed in [6]

whistles and identify their covert meanings. Although GPT-3’s performance varied widely across types of dogwhistles and targeted groups, results demonstrated that LLMs offer a unique opportunity to assist dogwhistle research and political content analysis more broadly. Finally, I showed that a popular toxicity detection system widely used for content moderation (Perspective API) consistently fails to identify hateful speech as toxic when standard group labels are substituted with dogwhistles. This not only highlights the risks of such coded language online, but also points to one way in which NLP systems perpetuate harms against marginalized groups.

Research Agenda

My past work has led me to identify four overarching questions to guide my future research program:

Q1: How can we model political language when language and politics keep changing?

The predominant approach for computational content analysis requires time-consuming manual data labeling for training machine learning models. While my framing research has highlighted this approach’s theoretical strengths [1, 4, 3], I have also demonstrated its severe limitations for analyzing framing in emerging crises [2]. It is essential to understand how narratives unfold day-by-day, or even hour-by-hour, during high-stakes events such as natural disasters, political protest, and war. In such situations, there is no time to construct large annotated datasets. While LLMs could aid analysis with few labeled examples, they are also of limited use due to outdated pretraining data. Dogwhistle research faces a similar challenge; dogwhistles can rapidly evolve to avoid out-group recognition and we are not well-equipped to handle such moving targets.

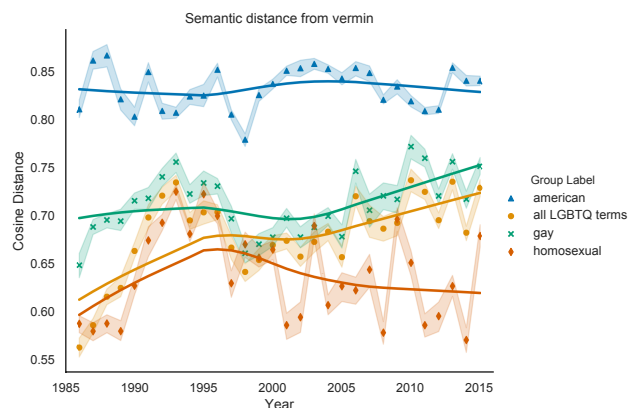


Figure 3: Changing associations between LGBTQ labels and the dehumanizing *Vermin* concept in the NY Times, measured via embedding representation distances [5].

My **ACL 2023** paper establishes the foundations for large-scale computational investigation of dogwhistles [6]. I developed a typology to better characterize dogwhistles (Figure 4) and curated the largest-to-date glossary of 340 expressions with rich contextual information and real-world examples. I assessed the extent to which GPT-3, a large language model (LLM), could surface examples of dog-

Achieving such flexibility requires modeling not just text, but also the underlying sociocultural context and cognitive processes that give rise to the linguistic patterns observed in text. In prior work, my colleagues and I have uncovered variation in words' meanings and social norms across online communities [9, 8], and have shown that modeling community and conversational contexts can improve NLP systems [8, 17]. My current research project extends these ideas to the study of implicit hate; I am incorporating cultural context into antisemitism detection and analysis by mapping incoming data to longstanding historical antisemitic tropes. I am particularly excited to develop systems that bridge NLP and media psychology. For example, framing operates via the cognitive mechanism of applicability; it affects the associations that people draw to existing beliefs and values when interpreting incoming information [18]. Future work could incorporate networks of such conceptual associations into computational models of framing, thus improving adaptability to new settings and yielding richer insights.

Q2: How can we build trustworthy LLM pipelines for social science research? Since OpenAI's release of ChatGPT last year, there has been increasing interest in using LLMs to annotate text for political content analysis and computational social science more broadly [19, 20]. My own work has shown that LLMs can be useful in surfacing and explaining hard-to-observe phenomena like dogwhistles [6], but they are less accurate for some social groups. Relying on LLMs for text analysis thus risks perpetuating biases against these groups. **It is thus critical to develop frameworks that guide fair and trustworthy uses of LLMs in social science research.**

There are multiple avenues for achieving this goal. First, we need to understand implicit biases in LLMs and the downstream harms of such biases when LLMs are used in practice. Within this space, I am especially interested in understanding how LLMs encode language ideologies: does LLM-generated text evoke stereotypes against non-prestige dialects or accents, or have worse performance responding to inputs from those language varieties? Second, we can explore strategies for integrating LLMs into research pipelines while ensuring the integrity of our findings. This could involve creative uses of LLMs beyond text annotation, such as in content analysis codebook creation or developing human-and-LLM-in-the-loop methods for interpreting results from unsupervised models.

Q3: How does exposure to nuanced political rhetoric impact people? Linguistic processes such as framing, dehumanization, and dogwhistles are effective because they affect how audiences understand political issues. **Moving beyond *how people talk about politics*, future research thus ought to investigate *how political talk affects people*.** Traditional experimental designs rely heavily on single stimuli and limited types of outcomes that can only be measured in surveys. Computational methods can complement such efforts by improving generalizability and facilitating analyses of how these linguistic strategies impact a wide range of audience behaviors. I began to explore these questions by measuring associations between framing and user engagement metrics [1, 2]. In other work, I combined NLP with causal inference methods to quantify how social media users' linguistic performances affect their friends' information sharing behavior [7]. Currently, I am mentoring a student who is extending this approach to understand how content producers' framing strategies influence how their social network connections frame political issues.

Q4: How can we improve online safety and civic health? Finally, I aim to establish collaborations that can put my research into practice and take action to make the online world a safer place. As part of my current project on antisemitism, I established a collaboration with Cyberwell, a nonprofit organization that works with social media platforms to identify and remove explicitly dangerous antisemitic content. I hope to work with human-computer interaction researchers to design systems that respond to online hate through empathy, de-escalation, and education, ultimately empowering individuals to cultivate a safer and more compassionate space for everybody.

References

- [1] **Julia Mendelsohn**, Ceren Budak, and David Jurgens. Modeling framing in immigration discourse on social media. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 2219–2263, 2021.
- [2] **Julia Mendelsohn**, Chan Young Park, Anjalie Field, and Yulia Tsvetkov. Challenges and opportunities in information manipulation detection: An examination of wartime russian media. *Findings of the Association for Computational Linguistics: EMNLP 2022*, pages 5209–5235, 2022.
- [3] **Julia Mendelsohn**, Maya Vijan, Dallas Card, and Ceren Budak. Framing social movements on social media: Unpacking diagnostic, prognostic, and motivational strategies. *Journal of Quantitative Description*, [under review], 2024.
- [4] Dallas Card, Serina Chang, Chris Becker, **Julia Mendelsohn**, Rob Voigt, Leah Boustan, Ran Abramitzky, and Dan Jurafsky. Computational analysis of 140 years of us political speeches reveals more positive but increasingly polarized framing of immigration. *Proceedings of the National Academy of Sciences*, 119(31):e2120510119, 2022.
- [5] **Julia Mendelsohn**, Yulia Tsvetkov, and Dan Jurafsky. A framework for the computational linguistic analysis of dehumanization. *Frontiers in artificial intelligence*, 3:55, 2020.
- [6] **Julia Mendelsohn**, Ronan Le Bras, Yejin Choi, and Maarten Sap. From dogwhistles to bullhorns: Unveiling coded rhetoric with language models. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 15162–15180, Toronto, Canada, July 2023. Association for Computational Linguistics.
- [7] **Julia Mendelsohn**, Sayan Ghosh, David Jurgens, and Ceren Budak. Bridging nations: Quantifying the role of multilinguals in communication on social media. In *Proceedings of the International AAAI Conference on Web and Social Media*, volume 17, pages 626–637, 2023.
- [8] Chan Young Park, **Julia Mendelsohn**, Karthik Radhakrishnan, Kinjal Jain, Tushar Kanakagiri, David Jurgens, and Yulia Tsvetkov. Detecting community sensitive norm violations in online conversations. In *Findings of the Association for Computational Linguistics: EMNLP 2021*, pages 3386–3397, 2021.
- [9] Li Lucy and **Julia Mendelsohn**. Using sentiment induction to understand variation in gendered online communities. In *Proceedings of the Society for Computation in Linguistics (SCiL) 2019*, pages 156–166, 2019.
- [10] Nick Haslam and Michelle Stratemeyer. Recent research on dehumanization. *Current Opinion in Psychology*, 11:25–29, 2016.
- [11] Nour S Kteily and Alexander P Landry. Dehumanization: Trends, insights, and challenges. *Trends in cognitive sciences*, 2022.
- [12] Paul Röttger, Bertie Vidgen, Dong Nguyen, Zeerak Waseem, Helen Margetts, and Janet Pierrehumbert. Hatecheck: Functional tests for hate speech detection models. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 41–58, 2021.

- [13] Michael Wiegand, Josef Ruppenhofer, and Elisabeth Eder. Implicitly abusive language—what does it actually look like and why are we not getting there? In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 576–587, 2021.
- [14] Emily M Bender, Timnit Gebru, Angelina McMillan-Major, and Shmargaret Shmitchell. On the dangers of stochastic parrots: Can language models be too big? In *Proceedings of the 2021 ACM conference on fairness, accountability, and transparency*, pages 610–623, 2021.
- [15] Percy Liang, Rishi Bommasani, Tony Lee, Dimitris Tsipras, Dilara Soylu, Michihiro Yasunaga, Yian Zhang, Deepak Narayanan, Yuhuai Wu, Ananya Kumar, et al. Holistic evaluation of language models. *arXiv preprint arXiv:2211.09110*, 2022.
- [16] Elin McCready. How dogwhistles work. In *New Frontiers in Artificial Intelligence: JSAI-isAI Workshops, JURISIN, SKL, AI-Biz, LENLS, AAA, SCIDOCA, kNeXI, Tsukuba, Tokyo, November 13-15, 2017, Revised Selected Papers 9*, pages 231–240. Springer, 2018.
- [17] Ethan Fast, Binbin Chen, **Julia Mendelsohn**, Jonathan Bassen, and Michael S Bernstein. Iris: A conversational agent for complex tasks. In *Proceedings of the 2018 CHI conference on human factors in computing systems*, pages 1–12, 2018.
- [18] Dietram A Scheufele. Framing as a theory of media effects. *Journal of communication*, 49(1):103–122, 1999.
- [19] Fabrizio Gilardi, Meysam Alizadeh, and Maël Kubli. Chatgpt outperforms crowd-workers for text-annotation tasks. *arXiv preprint arXiv:2303.15056*, 2023.
- [20] Caleb Ziems, William Held, Omar Shaikh, Jiaao Chen, Zhehao Zhang, and Diyi Yang. Can large language models transform computational social science? *arXiv preprint arXiv:2305.03514*, 2023.