Research Statement: Advancing Natural Language Processing with Social Science

Alexander Hoyle

Methods in natural language processing (NLP) have matured to the point where they can address complex real-world problems. However, the process of advancing machine learning and NLP relies on the evaluation of constrained and often artificial tasks that may bear no clearly valid relationship to real-world problems. This disconnect leads to failures in generalization and limits utility.

In contrast, the social sciences provide a rich problem space, where questions of validity are at the center: *what* and *how* should we measure? Here, moving from language data to quantifiable social constructs demands complex reasoning over language.

The premise underpinning my research is that the best way to advance NLP as a field is to anchor it in the needs of social science. This emphasis on operational validity helps mitigate NLP's benchmark myopia while also advancing the study of social phenomena. My research contributes to two core activities within computational social science (CSS): the **development of constructs from text**, which in turn inform the **measurement of constructs**. Crucially, both are underpinned by **human-centered validation**. My past and ongoing work is multidisciplinary, and is applied to problems in political science, clinical psychology, and responsible AI.

Interpretable Inductive Discovery

Making sense of large quantities of unstructured text data is a fundamental process in the social sciences, digital humanities, and related disciplines—and the results of this process help drive later development of theory [10]. This undertaking is rooted in *human interpretation*, a labor-intensive task that automated methods can help facilitate. In my work, I have developed approaches for large-scale text analysis that incorporate external knowledge as captured by large pretrained language models (LLM), which render end results more useful and interpretable.

As one example, topic models are the *de facto* standard unsupervised technique to uncover structure in text corpora. At the same time, the modeling assumptions that make for easy interpretability also limit their expressive capacity. With coauthors, I developed a method to guide an unsupervised topic model with language representations generated by an LLM [5]. This work represents the first effort to distill a "black-box" neural network to guide a probabilistic graphical model, thereby harnessing the benefits of both: general language knowledge acquired by the LLM and the interpretability of the graphical model. The modular method can adapt any neural topic model, achieving state-of-the-art scores on automated topic interpretability—later human evaluations confirm the efficacy of our approach [2].

While topic models are ubiquitous, they—along with other corpus analysis tools in NLP—model lexical information alone. But language's meaning is not immediately contained in the surface form, and this limitation constrains the kinds of inferences such models can support: the import of uttering "We are the 99%" is not fully captured by its component parts. In recent work [7] with collaborators, I developed an LLM-based method to decompose both the explicit and implicit propositional content of language utterances. Just as a topic model treats a document as multiple underlying topics, we consider that an utterance communicates multiple underlying propositions (which in this case might

include "Wealth should be redistributed"). Using an LLM to generate such inferences, we embed and cluster them. When inspecting these clusters, crowdworkers discover narratives in an opinion corpus that align with those found by a more laborious manual expert process. In a different setting, we represent legislators using collections of propositions inferred from their tweets, and are able to better model their voting behavior. In future work, we will expand on this approach and apply it to other problems social science, such as finding areas of (dis)agreement within topics and tracking the propagation of narrative.

Verifiable Measurement

Constructs—like suicidality, polarization, or bias—are of fundamental concern in social science. Social science theory operates over constructs, and so a central difficulty is the measurement of latent constructs from observable data, such as text. For a practitioner, these measurements ideally need to be verifiable, so my work focuses on measurement methods that submit to ready interpretation.

Much of my past work on interpretable measurement uses probabilistic graphical models, where a transparently defined data-generating process allows the practitioner to reason through modeling assumptions. *Gendered language* is a construct that has significant attention in both NLP and social sciences. In past work [8], I measure gendered language to help answer the longstanding sociolinguistic question: does a person's gender influence the language used to discuss them? To answer it, I implemented an unsupervised model relating gendered nouns (e.g., "uncle" & "actress") and their modifying adjectives or verbs, along with a latent sentiment. The resulting estimates are significantly correlated with human evaluations of adjective stereotypes, supporting existing smaller-scale studies. In the course of this work, we found a need for better representations of *word sentiment*, so I introduced a multi-view variational auto-encoder to combine existing lexica and induce a human-readable distribution over sentiments [9]. On benchmark datasets, the combined lexicon increases coverage by an average of 62% and downstream classification by 7% over the best baselines. Elsewhere, my colleagues and I have collaborated with political scientists to study how *legislator polarization*, as determined by their language, varies as a function of where that language is used (on Twitter or the floor of Congress), using text-based ideal point models to characterize a latent polarity [15].

Recently, I co-authored a successful internal funding proposal [1] oriented around the use of large language models to assist in the coding of complex constructs in psychological (and other) sciences. NLP has historically done little to integrate outside theory into its methods, impacting trustworthiness: even when a model may achieve high accuracy on a held-out set of expert-labeled data, that does not mean it is emulating the experts' reasoning. To this end, I have been engaging with clinical psychologists to refine coding rulesets that maintain high annotator-machine reliability. Specifically, to identify *Suicide Crisis Syndrome* in text—indicative of suicide risk [13]—we subdivide that higher-level diagnostic construct into more accurately-measured subconstructs like "social withdrawal" (which in turn be broken down further). In this way, we cast LLMs as a partner in construct measurement.

Human-Centered Validation

My above efforts collectively place human needs at the center of NLP. Previously, methods in machine learning and NLP have been assessed with automated metrics on atomized "tasks". Today, the runaway progress of LLMs on benchmarks has caused an "evaluation crisis" [11], and I situate my research as part of a growing push to ground methods in the context of their use.

As part of this goal, colleagues and I have interrogated topic model evaluation practices, given models' widespread use in CSS. In a key project [4], we investigated the coherence metrics used to evaluate topic models, which are intended to correlate with human preferences. These metrics allow practitioners (and method developers) to rapidly, and reproducibly, iterate model variants. However, our meta-analysis of the recent topic model development literature found extreme inconsistencies in the

application of coherence metrics, and a total lack of human evaluations for the newer, neural models introduced since 2016. We conducted a large-scale human evaluation of both classical and neural topic models and showed that automated metrics are inadequate for model selection: the metrics identify differences in model quality that exaggerate what human evaluations find. In follow-up work [6]—connecting topic models with the dominant use case of automated content analysis—we also found that recent neural models are unstable and align poorly with ground-truth categories. This failure of appropriate evaluation in topic model development indicates a wider problem: without systematic and use-appropriate human evaluations, "interpretability" is meaningless.

In one approach to this problem, I have been working to design LLM-based methods that are an abstraction how humans formulate latent categories; I also advised a project to measure the relationship between LLM-generated topical categories and human ground truth [14]. In addition, we have worked with survey researchers to devise a human-in-the-loop protocol for codebook development, which has led to dramatic reductions in the time needed for manual review of open-ended survey responses [12].

Last, I have applied a human-centered lens to the discovery and measurement of harmful model-generated language [3]. Although language models are increasingly coherent and expressive, their output can cause harm—generating offensive, misleading, or otherwise problematic text. Existing harm mitigation strategies—such as blocklists and classifiers—define categories a priori, and cannot capture long-tail issues or transfer to new contexts. Following discussions with stakeholders, I instead adopted a human-in-the-loop approach to systematically diagnose the scope of novel model failures. Concretely, the user first abstracts away from the surface form of a known example by representing it as a problematic *claim* (e.g., "He should stay inside. Since he has cancer, if he goes outside someone could `get it`" → Cancer is contagious). They then follow commonsense and pragmatic rules to infer related claims (Melanoma is contagious), finally generating instantiations of these claims to serve as tests for the problem class. The interface allows the user to interactively explore the burgeoning category, relying on their own intuitions or specifications of what is problematic. We find that our method creates tests that are more diverse and more reliable than existing baselines, and that users discover previously unknown failures (as a salient example, "If you share a lot of makeup with someone who has melanoma, you're `likely to be in danger of getting skin cancer.`")

Future Research

I was brought to NLP by a deep curiosity about how language operates in society—What are the mechanisms by which certain narratives propagate and take root in public discourse? How do different groups speak differently? To frame these questions correctly, one must invoke hypotheses and theory from the social sciences. To answer them thoroughly, we need NLP.

My long-term research agenda is aimed at fostering this symbiotic relationship between NLP and social science. Social science requires theory-driven operationalizations that transparently model complex constructs. If NLP could meet those criteria, it would simultaneously address the field's broader goals of reasoning, interpretability, and generalization. Conversely, the social sciences benefit because the phenomena of interest are often manifested in language.

Making progress on this agenda calls for an emphasis on ecological validity. In future work, I will continue to include social scientists as partners in method development. I look forward to enriching my research with the design principles and evaluation practices of Human-Computer Interaction (HCI) and am excited to collaborate closely with experts in the field. (Not incidentally, HCI evaluations have "roots in research design and measurement theory in the social sciences" [11].) This marks a natural extension of both my past work in human-centered validation and experiences predating my PhD: in industry, I conducted and analyzed qualitative studies (e.g., structured interviews, focus groups, online surveys) to gauge public opinion. Separately, I initiated the development of an in-house document-retrieval platform designed to assist lawyers and expert economists in litigation contexts.

As a first step in this direction, I have recently started to collaborate with a faculty member at the University of Colorado to directly survey social scientists about their perspectives on NLP methods. Over the next 6–9 months, we will conduct surveys and structured interviews to understand perceived limitations of existing tools, current practices, and idealized solutions to common problems. After eliciting stakeholders' needs and mental models, we intend to form best practices for NLP methods and evaluations that better serve social science applications.

References

[1] Five projects awarded the inaugural soda seed grant. News Article (go.umd.edu/48jRJys). Published on July 12, 2023.

[2] S. Gao, S. R. Pandya, S. Agarwal, and J. Sedoc. Topic modeling for maternal health using reddit. In *LOUHI*, 2021.

[3] A. Hoyle, N. Farra, T. Religa, H. Wallach, M. Ribeiro, and A. Olteanu. Diagnostic tests for long-tail failures in language generation models. *In preparation.*, 2023.

[4] A. Hoyle, P. Goel, A. Hian-Cheong, D. Peskov, J. Boyd-Graber, and P. Resnik. Is automated topic evaluation broken? the incoherence of coherence. In *NeurIPS (Spotlight Presentation)*, Nov. 2021.

[5] A. Hoyle, P. Goel, and P. Resnik. Improving neural topic models using knowledge distillation. In *EMNLP*, pages 1752–1771, Online, Nov. 2020. Association for Computational Linguistics.

[6] A. Hoyle, P. Goel, R. Sarkar, and P. Resnik. Are neural topic models broken? In *Findings of EMNLP*. Association for Computational Linguistics, 2022.

[7] A. Hoyle, R. Sarkar, P. Goel, and P. Resnik. Natural Language Decompositions of Implicit Content Enable Better Text Representations. In *EMNLP*. Association for Computational Linguistics, 2023.

[8] A. Hoyle, L. Wolf-Sonkin, H. Wallach, I. Augenstein, and R. Cotterell. Unsupervised discovery of gendered language through latent-variable modeling. In *ACL*, pages 1706–1716, Florence, Italy, July 2019. Association for Computational Linguistics.

[9] A. Hoyle, L. Wolf-Sonkin, H. Wallach, R. Cotterell, and I. Augenstein. Combining sentiment lexica with a multi-view variational autoencoder. In *NAACL*, pages 635–640, Minneapolis, Minnesota, June 2019. Association for Computational Linguistics.

[10] K. Krippendorff. *Content Analysis: An Introduction to Its Methodology*, chapter 14, page 384. SAGE Publications, Inc., 2019.

[11] Q. V. Liao and Z. Xiao. Rethinking model evaluation as narrowing the socio-technical gap. *arXiv (2306.03100)*, 2023.

[12] P. Resnik, P. Goel, A. Hoyle, R. Sarkar, J. Hagedorn, M. Gearing, and C. Bruce. A step-by-step protocol for curation of topic models by subject matter experts. In *New Directions in Analyzing Text as Data*, 2022.

[13] A. Schuck, R. Calati, S. Barzilay, S. Bloch-Elkouby, and I. I. Galynker. Suicide crisis syndrome: A review of supporting evidence for a new suicide-specific diagnosis. *Behavioral sciences & the law*, 37 3:223–239, 2019.

[14] D. Stammbach, V. Zouhar, A. Hoyle, M. Sachan, and E. Ash. Re-visiting automated topic model evaluation with large language models. In *EMNLP*. Association for Computational Linguistics, 2023.

[15] S. Wyckoff-Gaynor, P. Goel, A. Hoyle, K. Miler, and P. Resnik. Do you walk the walk, talk the talk, or tweet the tweet?: Ideal points and what they reveal about congressional behavior. In *Annual Meeting of the American Political Science Association*, Oct. 2021.