Research Statement:
Social Science as a Problem Space for Natural Language Processing

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 such 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 an ideal way to advance NLP as a field is by anchoring it in the needs of social science. Social scientists care about valid measurements, transparent operationalizations, and human usability—the same properties that make for better NLP. Conversely, the appropriate application of NLP methods to these fields can help answer substantive open questions.

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 *within* text. Crucially, both are underpinned by human-centered validation. My work is multidisciplinary, and has been applied to problems from political science, sociolinguistics, and law.[1]

Conceptualizing from Text Data

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 [11]. 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 (LLMs)**, which render end results more useful and interpretable [5, 7, 15, 1].

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

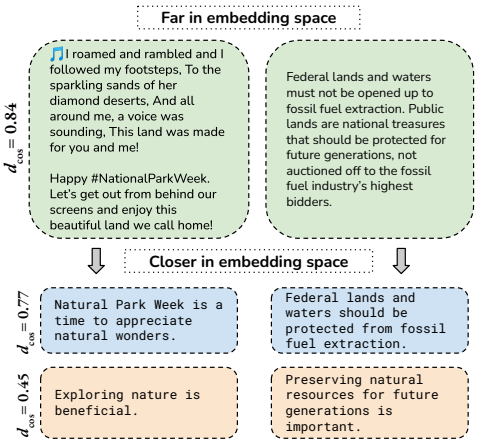

Figure 1: Pair of tweets from U.S. legislators along with inferentially-related propositions. Sentence embeddings over observed tweets have a high cosine distance, while embeddings over propositions place them closer to each other.

[1] A note on AI use: I take pride in my writing and refrain from using AI assistants for generating prose; I'd estimate fewer than 1% of these words were sourced from an AI assistant, where I use it as a glorified thesaurus for finding context-appropriate *mots justes*. I also had it minimally reformulate 2-3 stubborn sentences. The ideation, argumentation, and structure are also my own. Last, all em dashes are hand-crafted from three -.

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. This 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 two recent papers [7, 15], **our novel LLM-based methods help practitioners make sense of text data**. In the first effort, a language model decomposes the explicit and implicit propositional content of language utterances (which in this case might include "Wealth should be redistributed", fig. 1). Crowdworkers reviewed clusters of these propositions drawn from commentary about COVID vaccines, identifying narratives that align with those found in a manual content analysis by experts in prior work. Then, taking a problem from political science, we represent legislators using collections of propositions inferred from their tweets and are able to better model their voting behavior. While this first work abstracts text items as propositions, the second, *TopicGPT*, [15] generalizes them further as high-level topics; I managed the evaluation protocol, establishing that these topics aligned better with human-curated labels than existing methods.

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.**

Female		Male	
Positive	Negative	Positive	Negative
beautiful	battered	just	unsuitable
lovely	untreated	sound	unreliable
chaste	barren	righteous	lawless
gorgeous	shrewish	rational	inseparable
fertile	sheltered	peaceable	brutish
beauteous	heartbroken	prodigious	idle
sexy	unmarried	brave	unarmed
classy	undernourished	paramount	wounded

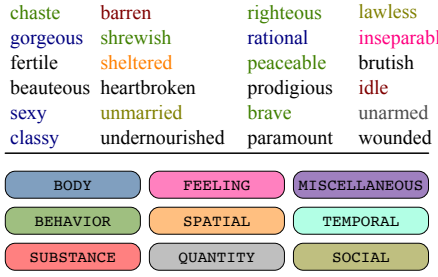

Figure 2: Adjectives, with sentiment, used to describe male and female people, as represented by our model.

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 one project [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 (fig. 2). 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.

My work on measurement also extends to expressions of ideology, which has involved collaborations with political scientists. In one project, we demonstrate that *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 [3]. More recently, I co–supervised a follow-up to the LLM-facilitated text analysis project I mentioned above [17]; here, our method places social media text on a scalar "support-oppose" spectrum with respect to controversial topics like gun control. We can then track how fine-grained perspectives change within political communities over time.

Last, I have studied the **use of LLMs for the measurement of constructs in social science** [13, 20]. LLMs have become ubiquitous for measurement in the social sciences, but their opacity makes them difficult to use effectively and reliably. I helped lead a systematic investigation into their use for scalar measurement, which led to actionable insights both for social scientists and for LLM alignment—findings I applied when aligning the public LLM trained at ETH and EPFL [19]. And yes: although LLM usage for this task is nominally not "verifiable", I have contributed to other work that helps measure the influence of a prompt on model behavior [14] and that assists in the development of codebooks, which enable reliable human annotation [16, 20].

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" [12], and **I situate my research as part of a growing push to ground methods in the context of their use.**

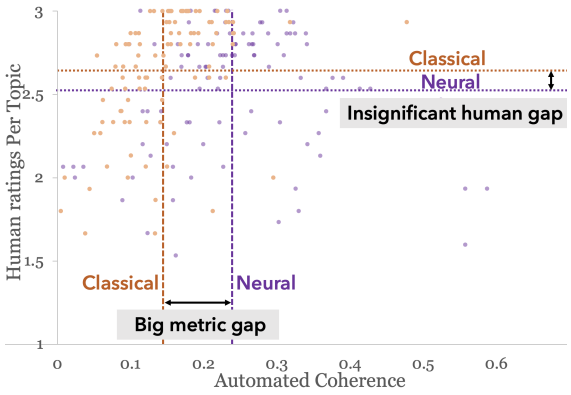

Figure 3: Despite a significant gap in topic coherence metrics between two model types (x-axis), human evaluations of topic quality are not meaningfully different (y-axis).

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 modeling 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 an extensive 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** (fig. 3).

This failure of appropriate evaluation of topic models indicates a wider problem: without use-appropriate human evaluations, how are can we be sure our methods matter? In follow-up work—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 [6], then developed both unsupervised LLM-as-judge metrics more grounded in that use-case that *do* agree with human judgments [18, 10].

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. I believe that making progress on this agenda will involve work falling under three broad categories:

Tailoring methods to answer substantive questions in social science. In the social sciences, there is a growing appetite for using computational text analysis to answer larger, important questions. Yet the specifics of a given research problem often require bespoke methods, creating opportunities to apply technical expertise. I want to develop methods to answer pressing questions in social science; my most recent interests relate to the *the epistemics of online communities.* Today, much of the "real world" is influenced by what happens online, in particular, within emergent communities of like-minded individuals. Anti-vaccine groups, once considered fringe in the U.S., have metastasized into a major political constituency: how did that happen? Building on past work in online polarization [17], I am supervising a junior PhD student in a project to understand the potential causes and dynamics of online misogyny: what brings men to hateful communities, and how does it affect their beliefs and how they are expressed?

Use-appropriate evaluations. Building effective methods within a domain calls for an emphasis on ecological validity: how are they used in practice? What makes a tool "good"? As a standard approach to both devising methods and constructing their evaluations, I have found it useful to decompose and approximate steps from a general use case that can then be assessed independently. For example, in a recent project, I separate qualitative content analysis into document labeling, classification, and ranking tasks [10]. This process leads to greater experimental flexibility and statistical power, while deferring costly expert evaluation until a later stage of development. I also am interested in expanding beyond individual system evaluation to larger-scale, naturalistic benchmarks where making progress leads to beneficial outcomes; for instance, I recently initiated a long-term project to improve to social science replicability by compiling a large benchmark from published code.

Social scientists as partners in method development. One-sided method development on the part of computer scientists tends to create "state-of-the-art" techniques that go unused. Social scientists are demanding users, and if we want uptake, we need to take their needs seriously—making methods that are both accessible and interpretable. Hence, my work also falls under the umbrella of human-AI collaboration, and I expect my research to increasingly incorporate the design principles and evaluation practices of HCI, and I hope to collaborate with faculty in that area. 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 and led the development of an in-house document-retrieval platform designed to assist lawyers and expert economists in litigation contexts.

References

[1] Y. Fan, Y. Tian, S. Ravfogel, M. Sachan, E. Ash, and A. Hoyle. The medium is not the message: Deconfounding document embeddings via linear concept erasure. In *EMNLP*, 2025.

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

[3] S. Gaynor, K. Miler, P. Goel, A. M. Hoyle, and P. Resnik. Express yourself (ideologically): Legislators' ideal points across audiences. *The Journal of Politics*, 2025.

[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*. Assoc. for Comp. Linguistics, 2019.

[9] A. Hoyle, L. Wolf-Sonkin, H. Wallach, R. Cotterell, and I. Augenstein. Combining sentiment lexica with a multi-view variational autoencoder. In *NAACL*. Assoc. for Comp. Linguistics, 2019.

[10] A. M. Hoyle, L. Calvo-Bartolomé, J. L. Boyd-Graber, and P. Resnik. ProxAnn: Use-oriented evaluations of topic models and document clustering. In W. Che, J. Nabende, E. Shutova, and M. T. Pilehvar, editors, *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 15872–15897, Vienna, Austria, July 2025. Association for Computational Linguistics.

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

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

[13] H. Licht, R. Sarkar, P. Y. Wu, P. Goel, N. Stoehr, E. Ash, and A. Hoyle. Measuring scalar constructs in social science with llms. In *EMNLP*, 2025.

[14] T. Nguyen, K. Du, A. Hoyle, and R. Cotterell. How persuasive is my context? In *EMNLP*, 2025.

[15] C. Pham, A. Hoyle, S. Sun, P. Resnik, and M. Iyyer. TopicGPT: A prompt-based topic modeling framework. In K. Duh, H. Gomez, and S. Bethard, editors, *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pages 2956–2984, Mexico City, Mexico, June 2024. Association for Computational Linguistics.

[16] 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.

[17] R. Sarkar, P. Wu, K. Miler, A. Hoyle, and P. Resnik. Pairscale: Analyzing attitude change with pairwise comparisons. In *Findings of NAACL*. Association for Computational Linguistics, 2025.

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

[19] A. Team. Apertus: Democratizing open and compliant llms for global language environments. Technical report, Swiss AI Initiative, 2025.

[20] C. Xiong, J. Ni, Y. Fan, V. Zouhar, D. Rooein, L. Calvo-Bartolomé, A. Hoyle, Z. Jin, M. Sachan, M. Leippold, D. Hovy, M. El-Assady, and E. Ash. Co-detect: Collaborative discovery of edge cases in text classification. In *EMNLP: System Demonstrations*, 2025.