Commentary

Biological Psychiatry: CNNI

Natural Language Processing in Psychiatry: A Field at an Inflection Point

Matthew M. Nour and Quentin J.M. Huys

The Promise of Natural Language Processing in Psychiatry

Natural language is "messy, ambiguous, chaotic, sprawling, and constantly in flux" (1). Yet it is by far the best way humans have of efficiently transmitting rich information, from complex thoughts, views, and preferences to medical information that is critical to life-or-death care decisions. Recent years have seen an explosion of interest in natural language data in medicine, cognitive neuroscience, and psychiatry, galvanized by the advent of new computational analytic tools and increases in scale and quality of language data collection.

This is particularly good news for psychiatry. The relevance of language to psychiatry is arguably greater than in any other medical domain. Here, language serves simultaneously as a medium through which subjective symptoms are reified and expressed, a channel through which treatment is delivered, and an object of clinical assessment in its own right. Much of the work of a clinical or psychotherapeutic interaction rests on a subtle attunement to the information carried in words: perhaps the recurrent and attractor-like signature of ruminative thought, a dyadic conceptual alignment that foreshadows an enduring therapeutic alliance, or a reduced narrative coherence that accompanies a prodromal psychosis. An ability to track such linguistic variables in a robust, quantitative, and automated manner would undoubtedly transform psychiatric practice and research.

An Inflection Point in the Automated Analysis of Natural Language

Until recently, such musing would have been considered pure science fiction. The past 5 years, however, mark an inflection point in the automated analysis of natural language, setting the stage for the current special issue of *Biological Psychiatry: Cognitive Neuroscience and Neuroimaging*.

This step change has been sparked by technological advances in artificial intelligence (AI) domains of deep learning and natural language processing (NLP). Chief among these is the transformer artificial neural network architecture (2), which yielded substantial improvements over recurrent neural networks (previously a dominant AI deep learning architecture for analysis of sequential data, such as language) in training efficiency, scalability, and ability to handle long-range textual dependencies. These improvements, combined with vast quantities of textual training data and computational resources, have spawned the recent wave of AI large language models (LLMs), including OpenAI's GPT (generative pretrained transformer) (3,4) and Google's BERT (bidirectional encoder representations from transformers) (5). These models extract rich statistical regularities from patterns of word co-occurrence in training data (typically, web crawls, digitized books, online message boards, preprints, and code repositories) using a self-supervised training objective that does not require laborious human data labeling (e.g., "next token prediction" or "masked language modeling").

The relevance of this advance for psychiatry is twofold. First, current LLMs (like GPT-4) display state-of-the-art performance capabilities in many task domains, including text classification and named entity recognition, summarization, sentiment analysis, and text generation (6). Clinical applications abound, particularly with respect to electronic health records (EHRs), and include clinical note summarization, information extraction, prognostic modeling, and chatbots that encode clinical knowledge (7,8). A second application, arguably more relevant to clinical cognitive neuroscience, rests on the use of NLP tools to inform empirical studies of cognition in more complex and naturalistic settings.

Cracking the Language Code—A New Frontier in Computational Psychiatry

Cognitive neuroscience approaches in psychiatry have largely relied on tasks designed to isolate and manipulate atomic computational units underlying behavior and symptoms (e.g., state-action credit assignment following prediction errors under a reinforcement learning framework) (9,10). This approach trades experimental traction for ecological validity and is not easily scalable to real-world settings, given a requirement for participant training and attentional maintenance. It is also less suited to studying abstract facets of cognition, such as analogical reasoning or emotional dynamics, which lack the normative frameworks that enable experimenters to infer psychological content from simple choice behavior. Yet this content is abundantly expressed in the words people effortlessly generate.

LLMs and related NLP tools might provide the key to tracking cognitive and emotional dynamics in natural language. A reason for optimism is that, as a byproduct of training, deep neural networks come to acquire structured intermediate representations that encode meaningful feature dimensions of the training data (e.g., in hidden layer activations or attention weights) (11–13). In the case of LLMs, these features might correspond to semantic and syntactic information (11,14). LLM intermediate representations have recently been used to investigate how semantic information is encoded in human brain activity (15–17). Consideration of LLM behavior (conditioned text generation) and internal representations has also sparked new debates in cognitive psychology and psycholinguistics (18,19).

https://doi.org/10.1016/j.bpsc.2023.08.001

ISSN: 2451-9022 Biological Psychiatry: Cognitive Neuroscience and Neuroimaging October 2023; 8:979–981 www.sobp.org/BPCNNI

The Current Special Issue

Thus, within the space of a few short years, the computational toolkit applicable to language in psychiatry has been radically transformed. This special issue showcases articles making use of these tools, from prediction modeling using EHRs and automatic symptom tracking to informing cognitive hypotheses in psychosis.

Data-Driven Precision Psychiatry. Patel *et al.* (20) focus on the application of NLP tools to EHRs for the purpose of transdiagnostic classification and personalized treatment selection (precision psychiatry). They discuss challenges including data harmonization and imputation of missing data and describe a modular NLP pipeline that transforms free-text EHRs to a structured format amenable to standard machine learning methods (e.g., binary classification). Such pipelines are currently highly domain and dataset specific, and often still require human-labeled training datasets. Recent incarnations increasingly use outputs from LLMs, including contextualized word embeddings. Patel *et al.* highlight the importance of model validation using data from multiple settings, a critical prerequisite to real-world adoption.

Characterizing Speech Structure and Thought Content. Two articles focus on characterizing dynamics and content of clinical speech data. Srivastava *et al.* (21) present a study using NLP tools to detect subjective symptoms and thought content from open-ended interviews conducted with patients displaying signs of early psychosis (n = 89) or at high risk of developing psychosis (n = 167). They use a sentence-level LLM (S-BERT) to track anomalous self-experiences in speech (so-called ipseity disturbances, including an altered sense of first-person subjectivity, diminished self-presence, and diminished ownership of experience). They quantify the semantic similarity (cosine distance) between participants' speech and items from a validated questionnaire to show how symptom expression varies across a psychosis spectrum.

Approaching clinical speech analysis from a complementary direction, Mota *et al.* (22) review the use of nonsemantic speech graphs to characterize word use patterns in psychosis. This approach represents the sequence of words in speech as a graph (each node a word, each edge indicating a temporal contiguity), which is amenable to graph-theoretic analysis (e.g., identification of word clusters and cycles). A tantalizing hypothesis is that network-level properties of speech graphs might track meaningful cognitive variables, including attentional processes and conceptual organization. In psychosis, graph connectedness has been found to relate to diagnostic status, cognitive variables, and social functioning.

Generative LLMs in Simulation-Based Studies of Thought Disorder. Finally, both Palaniyappan *et al.* (23) and Fradkin *et al.* (24) consider the potential of natural language generation models (i.e., autoregressive LLMs like GPT) to inform cognitive-linguistic hypotheses and validate NLP speech metrics. Palaniyappan *et al.*, taking inspiration from language-evolution theories of psychosis, propose using natural language generation systems (at various stages of development) as in silico models of formal thought disorder, and point to commonalities between failure modes of systems such as GPT-2 and -3 (namely, false contents, repetitiveness, and frank incoherence) and some facets of formal thought disorder in psychosis.

Fradkin *et al.* (24) emphasise the capacity of natural language generation systems to serve as in silico testbeds for assessing the validity and reliability of common NLP summary metrics. This rests on the fact that word generation parameters in such models (e.g., next-token choice stochasticity, size of conditioning context window) can be parametrically controlled, thus providing an opportunity to test how well different NLP summary metrics (e.g., word- and sentence-level semantic similarity) track "ground truth" generative parameters.

These complementary directions point to a possibility of bringing the study of language in psychiatry into the broader purview of theory-driven computational psychiatry, which strives to characterize observed behavior and symptom expression in terms of generative algorithmic processes (9).

Outlook

Psychiatry stands to gain much from AI advances in NLP, both in the development of diagnostic and prognostic machine learning tools and in the study of neurocognitive processes. However, these are early days, and it is unclear which of the extant approaches will prove to be ultimately clinically impactful. Despite this uncertainty, we believe that we stand at an inflection point in the field. We look forward to the increased use of NLP tools to bring meaning to unstructured and unwieldy datasets, shedding light on clinical and cognitive questions alike.

Acknowledgments and Disclosures

MMN has received research funding from Wellcome Trust and the National Institute for Health and Care Research. QJMH has obtained support from the University College London Hospitals NHS Foundation Trust National Institute for Health and Care Research Biomedical Research Center and research grant funding from Carigest S.A., German Research Foundation, Swiss National Science Foundation, and Wellcome Trust.

QJMH has received fees and options for consultancies from Aya Technologies and Alto Neuroscience, and research funding from Koa Health. MMN reports no biomedical financial interests or potential conflicts of interest.

Article Information

From the Department of Psychiatry, University of Oxford, Oxford, United Kingdom (MMN); the Max Planck University College London Centre for Computational Psychiatry and Ageing Research, Queen Square Institute of Neurology, University College London, London, United Kingdom (MMN); and the Applied Computational Psychiatry Lab, Mental Health Neuroscience Department, Division of Psychiatry and Max Planck Centre for Computational Psychiatry and Ageing Research, Queen Square Institute of Neurology, University College London, London, United Kingdom (QJMH).

Address correspondence to Matthew M. Nour, M.R.C.Psych., Ph.D., at matthew.nour@psych.ox.ac.uk.

Received Jul 31, 2023; accepted Aug 1, 2023.

References

1. Chollet F (2021): Deep Learning with Python, 2nd ed. Shelter Island, New York: Manning Publications Co.
2. Vaswani A, Shazeer N, Parmar N, Uszkoreit J, Jones L, Gomez AN, *et al.* (2017): Attention is all you need. Adv Neural Inf Process Syst 30:5999–6009.

3. Brown TB, Mann B, Ryder N, Subbiah M, Kaplan J, Dhariwal P, *et al.* (2020): Language models are few-shot learners. arXiv https://doi.org/10.48550/arXiv.2005.14165
4. Radford A, Narasimhan K, Salimans T, Sutskever I (2023): Improving language understanding by generative pre-training. Available at: https://www.mikecaptain.com/resources/pdf/GPT-1.pdf. Accessed August 23, 2023.
5. Devlin J, Chang M-W, Lee K, Toutanova K (2019): BERT: Pre-training of deep bidirectional transformers for language understanding. NAACL HLT 2019. The 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies. Proceedings of the Conference. (Long and Short Papers), Vol. 1. Stroudsburg, PA: Association for Computational Logistics, 4171–4186.
6. Bubeck S, Chandrasekaran V, Eldan R, Gehrke J, Horvitz E, Kamar E, *et al.* (2023): Sparks of artificial general intelligence: Early experiments with GPT-4. arXiv https://doi.org/10.48550/arXiv.2303.12712.
7. Singhal K, Azizi S, Tu T, Mahdavi SS, Wei J, Chung HW, *et al.* (2023): Large language models encode clinical knowledge. Nature 620:172–180.
8. Kraljevic Z, Bean D, Shek A, Bendayan R, Hemingway H, Au J, *et al.* (2022): Foresight - generative pretrained transformer (GPT) for modelling of patient timelines using EHRs. arXiv https://doi.org/10.48550/arXiv.2212.08072.
9. Huys QJM, Maia TV, Frank MJ (2016): Computational psychiatry as a bridge between neuroscience and clinical applications. Nat Neurosci 19:404–413.
10. Nour MM, Liu Y, Dolan RJ (2022): Functional neuroimaging in psychiatry and the case for failing better. Neuron 110:2524–2544.
11. Manning CD, Clark K, Hewitt J, Khandelwal U, Levy O (2020): Emergent linguistic structure in artificial neural networks trained by self-supervision. Proc Natl Acad Sci U S A 117:30046–30054.
12. Olah C, Mordvintsev A, Schubert L (2023): Feature visualization. How neural networks build up their understanding of images. Available at: https://distill.pub/2017/feature-visualization/. Accessed August 23, 2023.
13. Elhage N, Hume T, Olsson C, Schiefer N, Henighan T, Kravec S, *et al.* (2022): Toy models of superposition. arXiv https://doi.org/10.48550/arXiv.2209.10652
14. Piantadosi ST, Hill F (2022): Meaning without reference in large language models. arXiv https://doi.org/10.48550/arXiv.2208.02957.
15. Caucheteux C, Gramfort A, King J-R (2023): Evidence of a predictive coding hierarchy in the human brain listening to speech. Nat Hum Behav 7:430–441.
16. Goldstein A, Zada Z, Buchnik E, Schain M, Price A, Aubrey B, *et al.* (2022): Shared computational principles for language processing in humans and deep language models. Nat Neurosci 25:369–380.
17. Schrimpf M, Blank IA, Tuckute G, Kauf C, Hosseini EA, Kanwisher N, *et al.* (2021): The neural architecture of language: Integrative modeling converges on predictive processing. Proc Natl Acad Sci U S A 118: e2105646118.
18. Binz M, Schulz E (2023): Using cognitive psychology to understand GPT-3. Proc Natl Acad Sci U S A 120:e2218523120.
19. Piantadosi ST (2023): Modern language models refute Chomsky's approach to language. Available at: https://scholar.google.com/citations?view_op=view_citation&hl=de&user=zykJTC4AAAAJ&sortby=pubdate&citation_for_view=zykJTC4AAAAJ:gnsKu8c89wgC. Accessed August 23, 2023.
20. Patel R, Wickersham M, Cardinal RN, Fusar-Poli P, Correll CU (2023): Natural language processing: Unlocking the potential of electronic health record data to support transdiagnostic psychiatric research. Biol Psychiatry Cogn Neurosci Neuroimaging 8:982–984.
21. Srivastava A, Selloni A, Bilgrami ZR, Sarac C, McGowan A, Cotter M, *et al.* (2023): Differential expression of anomalous self-experiences in spontaneous speech in clinical high-risk and early-course psychosis quantified by natural language processing. Biol Psychiatry Cogn Neurosci Neuroimaging 8:1005–1012.
22. Mota NB, Weissheimer J, Finger I, Ribeiro M, Malcorra B, Hübner L (2023): Speech as a graph: Developmental perspectives on the organization of spoken language. Biol Psychiatry Cogn Neurosci Neuroimaging 8:985–993.
23. Palaniyappan L, Benrimoh D, Voppel A, Rocca R (2023): Studying psychosis using natural language generation: A review of emerging opportunities. Biol Psychiatry Cogn Neurosci Neuroimaging 8:994–1004.
24. Fradkin I, Nour MM, Dolan RJ (2023): Theory-driven analysis of natural language processing measures of thought disorder using generative language modeling. Biol Psychiatry Cogn Neurosci Neuroimaging 8:1013–1023.