Natural language processing in psychiatry: a field at an inflection point.

Matthew M Nour[1,2]
Quentin J.M. Huys[3]

1. Department of Psychiatry, University of Oxford, Oxford, OX3 7JX, UK
2. Max Planck University College London Centre for Computational Psychiatry and Ageing Research, London, WC1B 5EH, UK
3. 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, UCL, London, WC1B 5EH, UK

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 critical to life-or-death care decision. Abundant repositories of language data contain exceedingly rich information are everywhere, prompting an exploding interest in 'natural language' data in recent years. Tantalizingly, the complexities of language data are rapidly being addressed, bringing formal, quantitative analytic methods to this rich source of information.

This it is particularly good news for psychiatry. The relevance of language to psychiatry is arguably greater than 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.

This step change has been sparked by technological advances in 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 (RNNs, previously a dominant NLP tool) in both training efficiency 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 the training data (typically, web crawls, books, message boards, preprints, code repositories) using a self-supervised training objective that does not necessarily require laborious data labelling (though see .

The relevance of this advance for psychiatry is twofold. Firstly, current LLMs (like GPT-4) display state-of-the-art performance capabilities in many natural language task domains, including text classification and named entity recognition, summarization, sentiment analysis, and text generation [6]. Potential clinical applications abound, particularly in the domain of EHR, and include clinical note summarisation, information extraction, prognostic modelling, and chatbots that encode clinical knowledge [7,8]. A second application, arguably more relevant to clinical cognitive neuroscience, rests on the use of AI NLP tools to inform empirical studies of cognition in more complex and naturalistic experimental settings.

Cracking the language code – a new frontier in computational psychiatry

Traditional cognitive neuroscience approaches in psychiatry have relied on carefully curated tasks, designed to isolate, and manipulate 'atomic units' underlying the behaviour of interest, such as state-action credit assignment following prediction errors under a reinforcement learning framework [9,10]. While this curated approach has yielded valuable insights into the building blocks of cognition and their disruption in pathology, it trades experimental traction for ecological validity, and is not easily scalable to real-world clinical settings given a requirement for behavioural training and attentional maintenance. It is also less suited to studying more abstract areas of cognition, such as analogical reasoning or emotional dynamics, which lack well-established normative frameworks, yet are abundantly expressed in the words people effortlessly generate.

LLMs and related AI NLP tools might provide the missing key to tracking cognitive and emotional dynamics in both clinical and non-clinical settings. A reason for optimism is that, as a by-product of training (e.g., on a 'next token prediction' objective), deep neural networks like LLMs come to acquire structured intermediate representations of linguistic data that appear to encode semantic and syntactic information (e.g., in hidden layer activations or attention weights) [11,12]. These intermediate representations have recently been used to study how the brain encodes semantic and predictive information [13–15]. More broadly, consideration of LLM behaviour (conditioned text generation) and internal representations has also sparked new debates within cognitive psychology and psycholinguistics [16,17].

The current Special Issue

Thus, within the space of a few short years the computational toolkit available for analysis of language in psychiatry has been radically transformed. The current Special Issue showcases articles that demonstrate how this expanded toolkit is being applied, spanning a spectrum including prediction modelling using EHR, automatic symptom tracking, and informing cognitive hypotheses in psychosis.

NLP and data-driven precision psychiatry.
Patel and colleagues outline the opportunities afforded by NLP approaches to large EHR, pertaining to the first broad application of NLP in psychiatry, described above [18]. They focus on transdiagnostic classification and personalised treatment selection ('precision psychiatry'), outlining how automated NLP approaches can address inherent challenges, including data harmonisation, imputation of missing data, and extraction of clinical information from unstructured free text. They describe a modular sequential NLP pipeline that transforms an unstructured free-text EHR to structured data, amenable to standard machine learning methods (e.g., binary classification). Such pipelines are currently highly domain- and data-set specific, and often still require human-labelled training datasets. Recent incarnations increasingly employ outputs from LLMs, including contextualised word embeddings. Patel et al. highlight the importance of model validation using data from multiple clinical settings, a critical pre-requisite to adoption of NLP in clinical settings.

NLP tools for characterising speech structure and thought content.
Two articles focus on the application of NLP tools to characterise the dynamics and symptom content of clinical speech data. Srivastava and colleagues present an empirical study using NLP tools to detect subjective symptoms and thought content from open-ended interviews conducted with patients with early psychosis (n = 89), or at high risk of developing psychosis (n = 167) [19]. They focus on anomalous self-experiences (*ipseity* disturbances, including altered sense of first-person subjectivity, diminished self-presence, and diminished ownership of experience), and use a sentence-level LLM (S-BERT) to quantify the semantic similarity (cosine distance) between participants' self-referential utterances and self-report items on a validated anomalous experience questionnaire. This authors thus provide a proof-of-concept for use of NLP tools to track the most abstract of subjective experiences, and also present fascinating new data on how the expression of anomalous self-experience varies across a psychosis spectrum.

Approaching the analysis of speech from a complementary direction, Mota and colleagues present a comprehensive overview of the use of a non-semantic speech graphs to characterise word use patterns in psychosis [20]. 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 organisation. As applied to psychosis, graph properties such as connectedness have been found to relate to diagnostic status, cognitive variables, and social functioning.

Generative LLMs in simulation-based studies of thought disorder.
Finally, both Palaniyappan and colleagues [21] and Fradkin and colleagues [22] consider the potential of natural language generation (NLG) AI models (i.e., autoregressive LLMs like GPT) to inform cognitive-linguistic hypothesis and validate NLP metrics, respectively, in the case of schizophreniform formal thought disorder. Palaniyappan et al., taking inspiration from language-evolution theories of psychosis, propose using NLG 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., instead emphasise the capacity of NLG 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, and size of conditioning context window) can be parametrically controlled, thus providing an opportunity to test how well different NLP summary metrics (such as word- and sentence-level semantic similarity) track 'ground truth' generative parameters.

These complementary directions point to an exciting possibility of bringing the study of language in psychiatry into the broader purview of theory-driven Computational Psychiatry, which strives to characterise observed behaviour and symptom expression in terms of generative algorithmic processes.

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 neuro-cognitive processes. 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.

Disclosures

QJMH has obtained support by the UCLH NIHR BRC; fees and options for consultancies for Aya Technologies and Alto Neuroscience; and research grant funding from Carigest S.A., German Research Foundation, Koa Health, Swiss National Science Foundation and Wellcome Trust.

References

1. Francois Chollet. *Deep Learning with Python*. 2nd ed.; 2021.

2. Vaswani A, Shazeer N, Parmar N, et al. Attention is all you need. *Advances in Neural Information Processing Systems*. 2017;2017-Decem(Nips):5999-6009.

3. Brown TB, Mann B, Ryder N, et al. Language Models are Few-Shot Learners. Published online July 22, 2020. Accessed July 17, 2023. http://arxiv.org/abs/2005.14165

4. Radford A, Narasimhan K, Salimans T, Sutskever I. Improving Language Understanding by Generative Pre-Training. Published online 2018:12.

5. Devlin J, Chang MW, Lee K, Toutanova K. BERT: Pre-training of deep bidirectional transformers for language understanding. *NAACL HLT 2019 - 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies - Proceedings of the Conference*. 2019;1(Mlm):4171-4186.

6. Bubeck S, Chandrasekaran V, Eldan R, et al. Sparks of Artificial General Intelligence: Early experiments with GPT-4. Published online March 22, 2023. Accessed March 24, 2023. http://arxiv.org/abs/2303.12712

7. Singhal K, Azizi S, Tu T, et al. Large language models encode clinical knowledge. *Nature*. Published online July 12, 2023. doi:10.1038/s41586-023-06291-2

8. Kraljevic Z, Bean D, Shek A, et al. Foresight - Deep Generative Modelling of Patient Timelines using Electronic Health Records.

9. Huys QJM, Maia TV, Frank MJ. Computational psychiatry as a bridge between neuroscience and clinical applications. *Nature Neuroscience*. 2016;(February):1-21. doi:10.1038/nn.4238

10. Nour MM, Liu Y, Dolan RJ. Functional neuroimaging in psychiatry and the case for failing better. *Neuron*. 2022;110(16):2524-2544. doi:10.1016/j.neuron.2022.07.005

11. Piantadosi ST, Hill F. Meaning without reference in large language models. Published online August 12, 2022. Accessed December 17, 2022. http://arxiv.org/abs/2208.02957

12. Manning CD, Clark K, Hewitt J, Khandelwal U, Levy O. Emergent linguistic structure in artificial neural networks trained by self-supervision. *Proc Natl Acad Sci USA*. 2020;117(48):30046-30054. doi:10.1073/pnas.1907367117

13. Caucheteux C, Gramfort A, King JR. Evidence of a predictive coding hierarchy in the human brain listening to speech. *Nat Hum Behav*. Published online March 2, 2023. doi:10.1038/s41562-022-01516-2

14. Goldstein A, Zada Z, Buchnik E, et al. Shared computational principles for language processing in humans and deep language models. *Nature Neuroscience*. 2022;25(3):369-380. doi:10.1038/s41593-022-01026-4

15. Schrimpf M, Blank IA, Tuckute G, et al. The neural architecture of language: Integrative modeling converges on predictive processing. *Proc Natl Acad Sci USA*. 2021;118(45):e2105646118. doi:10.1073/pnas.2105646118

16. Binz M, Schulz E. Using cognitive psychology to understand GPT-3. *Proc Natl Acad Sci USA*. 2023;120(6):e2218523120. doi:10.1073/pnas.2218523120

17. Piantadosi ST. Modern language models refute Chomsky's approach to language. Published online 2023.

18. Patel R, Wickersham M, Cardinal RN, Fusar-Poli P, Correll CU. Natural Language Processing: Unlocking the Potential of Electronic Health Record Data to Support Transdiagnostic Psychiatric Research. *Biological Psychiatry: Cognitive Neuroscience and Neuroimaging*. Published online September 2022:S2451902222002166. doi:10.1016/j.bpsc.2022.09.002

19. Agrima Srivastava, Alexandra Selloni, Zarina Bilgrami, et al. Differential Expression of Anomalous Self-Experiences in Spontaneous Speech in Clinical High Risk and Early Course Psychosis Quantified by Natural Language Processing. *Biological Psychiatry: Cognitive Neuroscience and Neuroimaging*. Published online 2023.

20. Mota NB, Weissheimer J, Finger I, Ribeiro M, Malcorra B, Hübner L. Speech as a Graph: Developmental Perspectives on the Organization of Spoken Language. *Biological Psychiatry: Cognitive Neuroscience and Neuroimaging*. Published online April 2023:S2451902223000988. doi:10.1016/j.bpsc.2023.04.004

21. Palaniyappan L, Benrimoh D, Voppel A, Rocca R. Studying psychosis using Natural Language Generation: A review of emerging opportunities. *Biological Psychiatry: Cognitive Neuroscience and Neuroimaging*. Published online April 2023:S2451902223001040. doi:10.1016/j.bpsc.2023.04.009

22. Fradkin I, Nour MM, Dolan R J. Theory Driven Analysis of Natural Language Processing Measures of Thought Disorder using Generative Language Modeling. *Biological Psychiatry: Cognitive Neuroscience and Neuroimaging*. Published online May 2023:S2451902223001258. doi:10.1016/j.bpsc.2023.05.005