Mental Health Evaluation through Text Analysis (META)

Gabriel Bonnin

Ruhr University Bochum, Germany

Author Note

Correspondence concerning this article should be addressed to Gabriel Bonnin, Email: gabriel.bonnin@ruhr-uni-bochum.de

Mental Health Evaluation through Text Analysis (META)

Abstract

Psychotherapy is one of the most effective treatments for mental health problems, but its success depends on accurate diagnostic assessments. Current diagnostic tools often use standardized closed-ended scales that, while reliable, may fail to capture the complexity and individuality of mental states. In collaboration with Dr. Oscar Kjell at Lund University, this project leverages advances in artificial intelligence (AI) and natural language processing (NLP) to transform the way mental health is assessed.

This project is based on a unique longitudinal dataset collected over 10 years at the Mental Health Research and Treatment Center (FBZ) at Ruhr University Bochum. It consists of written self-reports in which patients describe their mental health problems, functional impairment, and therapy goals in their own words. These texts are linked to key clinical outcomes such as diagnoses, symptom severity, functional impairment, and treatment success, providing a rich, ecologically valid resource for understanding patient progress in psychotherapy.

While previous NLP-based mental health studies have focused on social media posts, limiting their clinical applicability, this project applies state-of-the-art large language models (LLMs) to real-world clinical data. By analyzing patients' open-ended responses, we aim to uncover patterns in how they articulate mental health, emotions, and treatment trajectories. These insights will inform the development of AI-powered tools that offer more personalized and clinically relevant assessments, surpassing traditional methods in accuracy and depth. Ultimately, this research aims to support clinicians in making more informed treatment decisions, enhance personalized care, and contribute to the modernization of mental health assessment.

Introduction

Mental health problems pose a significant global challenge, accounting for a considerable proportion of deaths and disability-adjusted life years (World Health Organization, 2017). Psychotherapy is an effective and sustainable intervention for reducing symptoms and improving quality of life (Chorpita et al., 2011; Wampold & Imel, 2015), but it's success critically depends on accurate assessments (Jensen-Doss & Weisz, 2008; Lutz et al., 2022).

Current assessment practices typically combine subjective self-reports with clinical observations (Bonnin et al., 2024). Standardized closed-ended tools such as the Beck Depression Inventory-II (Beck et al., 1996) rely on numerical scales (Likert, 1932) to structure and standardize assessments and are widely used in clinical research and practice. While these methods have advanced replicability and reliability in psychological assessment, they can miss important individual differences by restricting responses to pre-defined categories, limiting the ability to capture the complexity of mental states (Kjell, Kjell, et al., 2024).

Recent advances in AI, particularly transformer-based LLMs (Vaswani et al., 2017), present promising solutions to these limitations (Kjell, Kjell, et al., 2024). LLMs excel in analyzing context-rich natural language with remarkable accuracy across diverse tasks (Devlin et al., 2019). Open-ended response formats, where patients describe their experiences in their own words, provide high-dimensional, context-rich information that remains underutilized in current assessment practices. Empirical studies highlight the potential of NLP-based analysis of open-ended responses, achieving moderate convergence with closed-ended rating scales using traditional NLP methods (Kjell et al., 2019) and nearing theoretical upper limits of accuracy with LLMs (Kjell et al., 2022). Preliminary research also highlights their potential for predicting clinically significant outcomes, including suicide risk (Matero et al., 2019; Mohammadi et al., 2019; Zirikly et al., 2019).

At the FBZ at Ruhr University Bochum, open-ended patient responses have been routinely collected from approximately 3,000 patients across pre-, post-, and follow-up therapy sessions. While previous studies have explored the use of NLP for mental health assessment, they have largely relied on social media language, limiting their clinical applicability. My project moves beyond the current state-of-the-art by applying LLMs to analyze this unique, large-scale, and longitudinal clinical dataset, assessing the relationship between patients' probed mental health responses and clinically relevant constructs such as diagnosis, symptom severity, and functional impairment. Additionally, it will predict key clinical outcomes such as treatment response and therapy goal attainment. Furthermore, the project seeks to generate clinically meaningful, data-driven insights that go beyond traditional diagnostic categories, offering a more nuanced and patient-centered understanding of mental health trajectories.

Building on the expertise of the HRG at Lund University, a pioneer in the development of probed language-based assessments, I will adapt and refine existing methods and previously validated NLP models at HRG to analyze these data, while also training state-of-the-art NLP models tailored to this unprecedented dataset. The Gateway Fellowship provides a unique opportunity to extend my PhD research on evidence-based diagnostic practices by integrating AI-driven approaches into clinical assessment. This project represents a critical step toward modernizing assessment methods by leveraging open-ended natural language formats. By addressing limitations in current practices and emphasizing a patient-centered approach, this research aims to contribute to improved outcomes for individuals receiving mental health care.

Methods

Preprocessing

To streamline data collection, an automated transcription pipeline was implemented: The handwritten text data is first recorded by trained employees of the FBZ adult outpatient clinic using a mobile audio recording device. Identifying features (e.g. names, dates of birth, location details) are replaced by placeholders during recording (anonymisation). The transcription is then carried out on local hardware using the open source tool Whisper Large v2 (https://github.com/openai/whisper), a state-of-the-art speech-to-text model (Radford et al., 2022). Each recording begins with a structured introduction, including a patient identification code, followed by responses to predefined questions. The transcription pipeline automatically processes all audio recordings, extracts the patient codes, and identifies responses to key questions. The data is then cleaned and structured into a tabular format for further analysis.

Measures

Sociodemographic and context measures

demographics include age, sex, marital status, relationship status, general education, vocational qualification, work ability

context factors include: Vorbehandlung, how therapy ended.

Responses from open-ended questions before therapy

At the start of therapy, patients complete two separate questionnaires designed to assess key aspects of their mental health concerns, functional impairments, and expectations for treatment. Questions 1–9 come from the first questionnaire (*Fragebogen zur Lebensgeschichte*), and questions 10–13 come from the second (*Eingangsfragebogen*). The questions include:

1. **Problem development:** 'Briefly describe how the problems for which you are seeking treatment have developed over time.' (geman original question: „Beschreiben Sie kurz, wie sich Ihre Probleme, wegen derer Sie eine Behandlung aufsuchen, im Laufe der Zeit entwickelt haben.")
2. **Extra stressors:** 'What causes you stress in addition to your everyday problems (e.g. finances, housing situation)?' (geman original question: „Was macht Ihnen zusätzlich zu Ihren Problemen im Alltag Stress (z. B. Finanzen, Wohnsituation)¿‘)
3. **Pre-onset changes:** 'Did something special change in your life before the onset of your symptoms? (e.g. death of an important person, divorce or separation, change in work situation or income, addition to the family)' (geman original question: „Hat sich vor dem Beginn Ihrer Beschwerden etwas Besonderes in Ihrem Leben verändert? (z. B. Tod einer wichtigen Bezugsperson, Scheidung oder Trennung, Veränderung der Arbeitssituation oder des Einkommens, Familienzuwachs)“)
4. **Event connection:** 'Do you see a connection between the event(s) and the development of your problems?' (geman original question: „Sehen Sie einen Zusammenhang zwischen dem Ereignis/den Ereignissen und der Entwicklung Ihrer Probleme¿‘)
5. **Physical symptoms:** 'Are there any physical side effects when your problems occur?' (geman original question: „Gibt es körperliche Begleiterscheinungen, wenn Ihre Probleme auftreten¿‘)
6. **Problem causes:** 'What do you think are the causes of your problems?' (geman original question: „Welche Ursachen sehen Sie für Ihre Probleme¿‘)
7. **Expected improvements:** 'What would improve in your life if you no longer had your problems?' (geman original question: „Was würde sich in Ihrem Leben verbessern,

wenn Sie ihre Probleme nicht mehr hätten¿')

8. **Environment response:** 'How does your environment (partner, family, friends, work colleagues) react to your problems?' (geman original question: „Wie reagiert Ihre Umwelt (Partner:in, Familie, Freund:innen, Arbeitskolleg:innen) auf die Probleme¿')
9. **No change required:** 'What should not change under any circumstances as a result of the therapy?' (geman original question: „Was sollte sich durch die Therapie auf keinen Fall verändern¿')
10. **Problem description:** 'Finally, please describe in your own words the problems for which you would like treatment.' (geman original question: „Beschreiben Sie zum Abschluss bitte noch einmal in eigenen Worten Ihre Probleme, deretwegen Sie eine Behandlung wünschen.")
11. **Impacted life areas:** 'In which areas of your life do these problems limit you (e.g. job, relationship)?' (geman original question: „In welchen Lebensbereichen schränken Sie diese Probleme ein (z. B. Beruf, Partnerschaft)¿')
12. **Therapy goals:** 'What would you like to achieve for yourself in therapy?' (geman original question: „Was möchten Sie in der Therapie für sich erreichen¿')

Psychometric measures

Clinical variables retrieved from the FBZ database include DSM-5 and ICD-10 diagnosis, symptom severity scores, therapist rated improvement scores, positive mental health measure, and therapeutic process variables.

Diagnosis. Diagnosis at the outpatient clinic is conducted using structured clinical interviews. These typically take place before therapy begins, usually at the fourth therapist–patient contact. The interview used is the Diagnostic Interview for Mental Disorders (Margraf et al., 2021), which covers the most frequent DSM-5 disorders encountered in outpatient therapy settings.

Beck-Depression-Inventory II..

Depression Anxiety Stress Scale 42.

Brief Symptom Inventory.

Symptom Checklist.

Childhood Trauma Questionnaire.

Clinical Global Impression.

Severity Scale. *Ziehen Sie Ihren gesamten Erfahrungsschatz an dieser Art von Kranken in Betracht, und geben Sie an, wie hoch Sie den jetzigen Grad der seelischen Erkrankung des Patienten einschätzen.*

Improvement Scale. Patient: *Beurteilen Sie dabei Ihre Zustandsänderung insgesamt, also nicht nur das Ergebnis der Behandlung. Bitte vergleichen Sie Ihren jetzigen Zustand mit dem zu Beginn der Behandlung, und geben Sie an, inwieweit sich Ihr Krankheitsbild geändert hat.*

Therapist: *Beurteilen Sie dabei die Zustandsänderung insgesamt, also nicht nur das Ergebnis der Behandlung. Bitte vergleichen Sie den jetzigen Zustand des Patienten mit dem zu Beginn der Behandlung, und geben Sie an, inwieweit sich das Krankheitsbild des Patienten geändert hat.*

Global Improvement

1. *Sind Ihre Erwartungen an diese Therapie (bislang) in Erfüllung gegangen?*
2. *Wie sehr hat Ihnen die Therapie insgesamt (bislang) genützt?*

Goal Attainment Scale

Bitte geben Sie an, in wieweit Sie das folgende Ziel bislang erreicht haben? [1-10]

Measurement time points

Analysis

The Sequential Evaluation with Model Pre-registration (Kjell, Ganesan, et al., 2024) framework will be implemented to ensure robust model development, mitigating overfitting and enabling unbiased performance evaluation. Additionally, evaluating models on prospective data will simulate real-world clinical deployment by assessing performance on new, unseen patient data.

During the model development phase, preprocessing pipelines will be finalized, and exploratory models developed using advanced cross-validation techniques. Contextual embeddings derived from pretrained LLMs will be linked to clinical outcomes using state-of-the-art prediction models, including ridge regression (Hoerl & Kennard, 1970), lasso regression (Tibshirani, 1996), and random forests (Ho, 1995). The final pipelines will be pre-registered for evaluation (e.g., https://aspredicted.org/).

In the evaluation phase, pre-registered models will be tested on held-out datasets, enabling unbiased performance assessments. This phase will include detailed error analysis to identify potential biases and limitations, as well as comparative analyses to benchmark model performance against the HRG's models.

Furthermore, topic modeling techniques (e.g., Latent Dirichlet Allocation; (Blei et al., 2003), or BERTopic; (Grootendorst, 2022)) will be employed to explore themes in patient-generated text. These analyses will provide valuable clinical insights into patients' subjective experiences, including their perceived problems, impairments, and goals. By analyzing these insights, I aim to highlight commonalities and differences in patient narratives across diverse populations or conditions. This process may also reveal nuanced linguistic cues that correlate with clinical outcomes, offering a richer understanding of patient perspectives and informing personalized care strategies.

Results

References

Beck, A. T., Steer, R. A., & Brown, G. K. (1996). *BDI-II: Beck Depression Inventory manual* (2nd ed.). Psychological Corporation.

Blei, D. M., Ng, A. Y., & Jordan, M. I. (2003). Latent dirichlet allocation. *Journal of Machine Learning Research*, *3*(null), 993–1022.

Bonnin, G., Kröber, S., & Brachel, R. von. (2024). Die Klassifikation psychischer Störungen mit diagnostischen Interviews. In T. Teismann, P. Thoma, S. Taubner, A. Wannemüller, & K. von Sydow (Eds.), *Klinische Psychologie und Psychotherapie: Ein verfahrensübergreifendes Lehr- und Lernbuch*. Hogrefe.

Chorpita, B. F., Daleiden, E. L., Ebesutani, C., Young, J., Becker, K. D., Nakamura, B. J., Phillips, L., Ward, A., Lynch, R., Trent, L., et al. (2011). Evidence-based treatments for children and adolescents: An updated review of indicators of efficacy and effectiveness. *Clinical Psychology: Science and Practice*, *18*(2), 154–172.

Devlin, J., Chang, M.-W., Lee, K., & Toutanova, K. (2019). BERT: Pre-training of deep bidirectional transformers for language understanding. In J. Burstein, C. Doran, & T. Solorio (Eds.), *Proceedings of the 2019 conference of the north American chapter of the association for computational linguistics: Human language technologies, volume 1 (long and short papers)* (pp. 4171–4186). Association for Computational Linguistics. https://doi.org/10.18653/v1/N19-1423

Grootendorst, M. (2022). *BERTopic: Neural topic modeling with a class-based TF-IDF procedure*. https://arxiv.org/abs/2203.05794

Ho, T. K. (1995). Random decision forests. *Proceedings of 3rd International Conference on Document Analysis and Recognition*, *1*, 278–282 vol.1. https://doi.org/10.1109/ICDAR.1995.598994

Hoerl, A. E., & Kennard, R. W. (1970). Ridge Regression: Biased Estimation for Nonorthogonal Problems. *Technometrics*, *12*(1), 55–67. https://doi.org/10.1080/00401706.1970.10488634

Jensen-Doss, A., & Weisz, J. R. (2008). Diagnostic agreement predicts treatment process and outcomes in youth mental health clinics. *Journal of Consulting and Clinical Psychology*, *76*(5), 711–722. https://doi.org/10.1037/0022-006X.76.5.711

Kjell, O. N. E., Ganesan, A. V., Boyd, R., Oltmanns, J. R., Rivero, A., Feltman, S., Carr, M. A., Luft, B. J., Kotov, R., & Schwartz, H. A. (2024). *Demonstrating high validity of a new AI-language assessment of PTSD: A sequential evaluation with model pre-registration*. PsyArXiv. https://doi.org/10.31234/osf.io/xw24e

Kjell, O. N. E., Kjell, K., Garcia, D., & Sikström, S. (2019). Semantic measures: Using natural language processing to measure, differentiate, and describe psychological constructs. *Psychological Methods*, *24*(1), 92–115. https://doi.org/10.1037/met0000191

Kjell, O. N. E., Kjell, K., & Schwartz, H. A. (2024). Beyond rating scales: With targeted evaluation, large language models are poised for psychological assessment. *Psychiatry Research*, *333*, 115667. https://doi.org/10.1016/j.psychres.2023.115667

Kjell, O. N. E., Sikström, S., Kjell, K., & Schwartz, H. A. (2022). Natural language analyzed with AI-based transformers predict traditional subjective well-being measures approaching the theoretical upper limits in accuracy. *Scientific Reports*, *12*(1), 3918. https://doi.org/10.1038/s41598-022-07520-w

Likert, R. (1932). A technique for the measurement of attitudes. *Arch. Psychol.*, *140*(55).

Lutz, W., Schwartz, B., & Delgadillo, J. (2022). Measurement-Based and Data-Informed Psychological Therapy. *Annual Review of Clinical Psychology*, *18*(1), 71–98. https://doi.org/10.1146/annurev-clinpsy-071720-014821

Margraf, J., Cwik, J. C., Brachel, R. von, Suppiger, A., & Schneider, S. (2021). *DIPS open access 1.2: Diagnostisches interview bei psychischen störungen*. https://doi.org/10.46586/rub.172.149

Matero, M., Idnani, A., Son, Y., Giorgi, S., Vu, H., Zamani, M., Limbachiya, P., Guntuku, S. C., & Schwartz, H. A. (2019). Suicide risk assessment with multi-level dual-context language and BERT. In K. Niederhoffer, K. Hollingshead, P. Resnik, R. Resnik, & K. Loveys (Eds.), *Proceedings of the sixth workshop on computational linguistics and clinical psychology* (pp. 39–44). Association for Computational Linguistics. https://doi.org/10.18653/v1/W19-3005

Mohammadi, E., Amini, H., & Kosseim, L. (2019). CLaC at CLPsych 2019: Fusion of neural features and predicted class probabilities for suicide risk assessment based on online posts. In K. Niederhoffer, K. Hollingshead, P. Resnik, R. Resnik, & K. Loveys (Eds.), *Proceedings of the sixth workshop on computational linguistics and clinical psychology* (pp. 34–38). Association for Computational Linguistics. https://doi.org/10.18653/v1/W19-3004

Radford, A., Kim, J. W., Xu, T., Brockman, G., McLeavey, C., & Sutskever, I. (2022). Robust Speech Recognition via Large-Scale Weak Supervision. *arXiv Preprint arXiv: 2212.04356*. https://arxiv.org/abs/2212.04356

Tibshirani, R. (1996). Regression shrinkage and selection via the lasso. *Journal of the Royal Statistical Society: Series B (Methodological)*, *58*(1), 267–288. https://doi.org/10.1111/j.2517-6161.1996.tb02080.x

Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., Kaiser, L., & Polosukhin, I. (2017). Attention is All you Need. *Advances in Neural Information*

Processing Systems, *30*.

Wampold, B. E., & Imel, Z. E. (2015). *The great psychotherapy debate: The evidence for what makes psychotherapy work*. Routledge.

World Health Organization. (2017). *Depression and other common mental disorders: Global health estimates*.

Zirikly, A., Resnik, P., Uzuner, Ö., & Hollingshead, K. (2019). CLPsych 2019 shared task: Predicting the degree of suicide risk in Reddit posts. In K. Niederhoffer, K. Hollingshead, P. Resnik, R. Resnik, & K. Loveys (Eds.), *Proceedings of the sixth workshop on computational linguistics and clinical psychology* (pp. 24–33). Association for Computational Linguistics. https://doi.org/10.18653/v1/W19-3003

Time-point	DU-DI	DU-Prä	KZT1-DU4	KZT1-DUPost	KZT2-DUPost	LZT1-DUPost	LZT2-DUPost	Kat6
Explanation	Pre-therapy, 4th contact	Pre-therapy, 6th contact	4th therapy session	12th therapy session	24th therapy session	45th therapy session	60th therapy session	6 month after therapy
Diagnosis	X							
Demographics		X						
BSI	X		X	X	X	X	X	X
BDI-II		X	X	X	X	X	X	X
DASS-42		X	X	X	X	X	X	X
SCL-9K								
CTQ		X						
CGI-S		X						
CGI-I								
Glob-Pt								
GAS								

Figure 1

The step-by-step process of patient narrative analysis, from preprocessing to data evaluation.