SYSTEMATIC REVIEW OPEN

Check for updates

Natural language processing for mental health interventions: a systematic review and research framework

Matteo Malgaroli[1]✉, Thomas D. Hull[2], James M. Zech[2,3] and Tim Althoff[4]

Neuropsychiatric disorders pose a high societal cost, but their treatment is hindered by lack of objective outcomes and fidelity metrics. AI technologies and specifically Natural Language Processing (NLP) have emerged as tools to study mental health interventions (MHI) at the level of their constituent conversations. However, NLP's potential to address clinical and research challenges remains unclear. We therefore conducted a pre-registered systematic review of NLP-MHI studies using PRISMA guidelines (osf.io/s52jh) to evaluate their models, clinical applications, and to identify biases and gaps. Candidate studies (n = 19,756), including peer-reviewed AI conference manuscripts, were collected up to January 2023 through PubMed, PsycINFO, Scopus, Google Scholar, and ArXiv. A total of 102 articles were included to investigate their computational characteristics (NLP algorithms, audio features, machine learning pipelines, outcome metrics), clinical characteristics (clinical ground truths, study samples, clinical focus), and limitations. Results indicate a rapid growth of NLP MHI studies since 2019, characterized by increased sample sizes and use of large language models. Digital health platforms were the largest providers of MHI data. Ground truth for supervised learning models was based on clinician ratings ($n = 31$), patient self-report ($n = 29$) and annotations by raters ($n = 26$). Text-based features contributed more to model accuracy than audio markers. Patients' clinical presentation ($n = 34$), response to intervention ($n = 11$), intervention monitoring ($n = 20$), providers' characteristics ($n = 12$), relational dynamics ($n = 14$), and data preparation ($n = 4$) were commonly investigated clinical categories. Limitations of reviewed studies included lack of linguistic diversity, limited reproducibility, and population bias. A research framework is developed and validated (NLPxMHI) to assist computational and clinical researchers in addressing the remaining gaps in applying NLP to MHI, with the goal of improving clinical utility, data access, and fairness.

Translational Psychiatry (2023)13:309 ; https://doi.org/10.1038/s41398-023-02592-2

INTRODUCTION

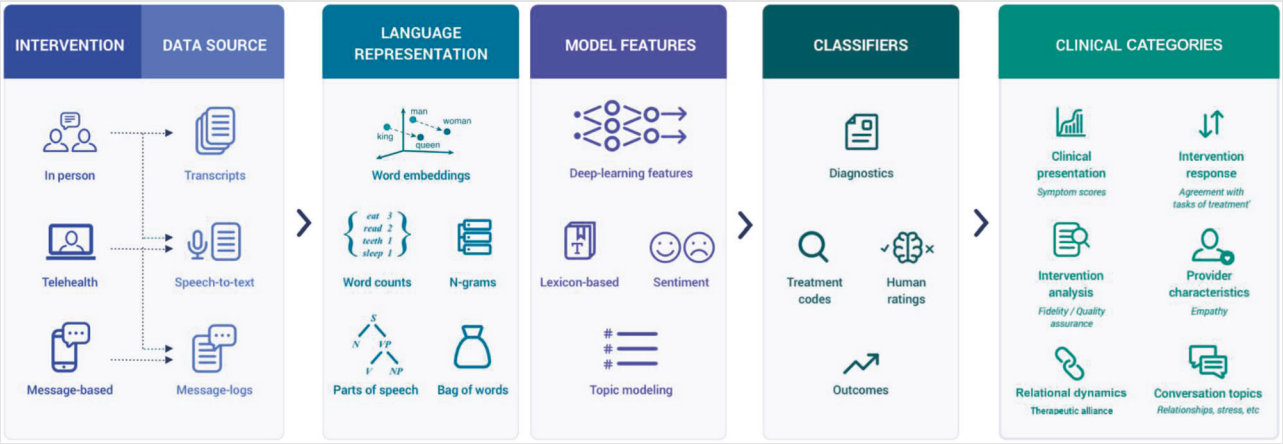
Neuropsychiatric disorders including depression and anxiety are the leading cause of disability in the world [1]. The sequelae to poor mental health burden healthcare systems [2], predominantly affect minorities and lower socioeconomic groups [3], and impose economic losses estimated to reach 6 trillion dollars a year by 2030 [4]. Mental Health Interventions (MHI) can be an effective solution for promoting wellbeing [5]. Numerous MHIs have been shown to be effective, including psychosocial, behavioral, pharmacological, and telemedicine [6–8]. Despite their strengths, MHIs suffer from systemic issues that limit their efficacy and ability to meet increasing demand [9, 10]. The first is the lack of objective and easily administered diagnostics, which burden an already scarce clinical workforce [11] with diagnostic methods that require extensive training. A second is variable treatment quality [12]. Widespread dissemination of MHIs has shown reduced effect sizes [13], not readily addressable through supervision and current quality assurance practices [14–16]. The third is too few clinicians [11], particularly in rural areas [17] and developing countries [18], due to many factors, including the high cost of training [19]. As a result, the quality of MHI remains low [14], highlighting opportunities to research, develop and deploy tools that facilitate diagnostic and treatment processes.

Recent innovations in the fields of Artificial Intelligence (AI) and machine learning [20] offer options for addressing MHI challenges. Technological and algorithmic solutions are being developed in many healthcare fields including radiology [21], oncology [22], ophthalmology [23], emergency medicine [24], and of particular interest here, mental health [25]. An especially relevant branch of AI is Natural Language Processing (NLP) [26], which enables the representation, analysis, and generation of large corpora of language data. NLP makes the quantitative study of unstructured free-text (e.g., conversation transcripts and medical records) possible by rendering words into numeric and graphical representations [27]. MHIs rely on linguistic exchanges and so are well suited for NLP analysis that can specify aspects of the interaction at utterance-level detail for extremely large numbers of individuals, a feat previously impossible [28]. Typically unexamined characteristics of providers and patients are also amenable to analysis with NLP [29] (Box 1). NLP for MHI began with prepackaged software tools [30], followed by more computationally intense deep neural networks [31], particularly large language

[1]Department of Psychiatry, New York University, Grossman School of Medicine, New York, NY 10016, USA. [2]Talkspace, New York, NY 10025, USA. [3]Department of Psychology, Florida State University, Tallahassee, FL 32306, USA. [4]Department of Computer Science, University of Washington, Seattle, WA 98195, USA.
✉email: matteo.malgaroli@nyulangone.org

Received: 7 March 2022 Revised: 31 August 2023 Accepted: 4 September 2023
Published online: 06 October 2023

Box 1. Overview and glossary of terms for Natural Language Processing (NLP)



Language Representation.

Word Embeddings: Words are mapped to a numeric vector space by an algorithm (e.g., word2vec) based on how they are used and their most frequent neighbor words in a large text dataset. Words have similar values when they co-occur in the same contexts, indicating a shared meaning.

Word counts (Unigrams): Single words analyzed based on their frequency.

N-grams: Language model consisting of sequences of *n*-number of words, to capture word context (e.g., the bigram "*not depressed*").

Part-of-Speech: Label words by their grammatical and syntactic functions.

Bag-of-Words/TF-IDF: Proportional frequency of words or *n*-grams to identify unique features of a text.

Model Features.

Deep Learning (DL) Features: DL algorithms are differentiated by number of layers, complexity, and model parameters. Language models are trained using large amounts of text data (e.g., all of Wikipedia), removing random words in a sentence, and learning to fill in the blank. This results in probabilistic models of language that can both interpret and produce text. Transformer architectures (e.g., BERT) also have attention mechanisms to help maintain context connections between distant words. Specific features for clinical tasks are generated by fine-tuning language models on domain-specific datasets.

Topic Modeling: Extracts and clusters common topics emerging in a text.

Lexicon Features: Matching text to a predefined word list made by human experts.

Sentiment: Matching text to emotions. Performed through dictionary methods, human raters, or pre-trained models.

Classifiers (model output)

Supervised models: Identify a category or outcome (e.g., diagnosis) after training on a dataset with examples. Human-labeled cases are known as the model's 'ground truth' and performance is measured against match with ground truth labels.

Unsupervised models: Derive features from a dataset based on the distribution.

models (i.e., attention-based architectures such as Transformers) [32], and other methods for identifying meaningful trends in large amounts of data. The diffusion of digital health platforms has made these types of data more readily available [33]. These data make it possible to study treatment fidelity [33], estimate patient outcomes [34], identify treatment components [35], evaluate therapeutic alliance [36], and gauge suicide risk [37] in a transformative way, sufficient to generate anticipation and apprehension regarding conversational agents [38]. Lastly, NLP has been applied to mental health-relevant contexts outside of MHI including social media [39] and electronic health records [40].

While these studies demonstrate NLP's research potential, questions remain about its impact on clinical practice. A significant limiting factor is the current separation between two communities of expertise: clinical science and computer science. Clinical researchers possess domain knowledge on MHI but have difficulty keeping up with the rapid advances in NLP. The clearest reflection of this separation is the continued reliance of clinical researchers on traditional expert-based dictionary methods [30] versus the ongoing state-of-the-art developments in large language models within computer science [32]. Accordingly, while prior reviews provided insights into the growing role of machine learning in mental health [25, 41], they did not include peer-reviewed manuscripts from AI conferences where many advances in NLP are reported. In addition, NLP pipelines were not deconstructed into algorithmic components, limiting the ability to identify distinctive model features. Meanwhile, computer scientists and computational linguists are driving developments in NLP that, while methodologically advanced, are typically limited in their applicability to clinical service provision.

We therefore conducted a systematic review of NLP studies for mental health interventions, examining their algorithmic and clinical characteristics to promote the intersection between computer and clinical science. Our aim was threefold: 1) classify NLP methods deployed to study MHI; 2) identify clinical domains and use them to aggregate NLP findings; 3) identify limitations of current NLP applications to recommend solutions. We examined each manuscript for clinical components (setting, aims, transcript source, clinical measures, ground truths and raters) and key features of the NLP pipeline (linguistic representations and features, classification models, validation methods, and software packages). Finally, we explored common areas, biases, and gaps in the current NLP applications for MHI, and proposed a research framework to address these limitations.

METHODS

Search protocol and eligibility

The systematic review followed the Preferred Reporting Items for Systematic Reviews and Meta-Analyses (PRISMA) guidelines. The review was pre-registered, its protocol published with the Open Science Framework (osf.io/s52jh). The review focused on NLP for *human-to-human* Mental Health Interventions (MHI), defined as psychosocial, behavioral, and pharmacological interventions aimed at improving and/or assessing mental health (e.g., psychotherapy, patient assessment, psychiatric treatment, crisis

counseling, etc.). We excluded studies focused solely on *human-computer* MHI (i.e., conversational agents, chatbots) given lingering questions related to their quality [38] and acceptability [42] relative to human providers. We also excluded social media and medical record studies as they do not directly focus on intervention data, despite offering important auxiliary avenues to study MHI. Studies were systematically searched, screened, and selected for inclusion through the Pubmed, PsycINFO, and Scopus databases. In addition, a search of peer-reviewed AI conferences (e.g., Association for Computational Linguistics, NeurIPS, Empirical Methods in NLP, etc.) was conducted through ArXiv and Google Scholar. The search was first performed on August 1, 2021, and then updated with a second search on January 8, 2023. Additional manuscripts were manually included during the review process based on reviewers' suggestions, if aligning with MHI broadly defined (e.g., clinical diagnostics) and meeting study eligibility. Search string queries are detailed in the supplementary materials.

Eligibility and selection of articles

To be included, an article must have met five criteria: (1) be an original empirical study; (2) written in English; (3) vetted through peer-review; (4) focused on MHI; and (5) analyzed text data that was gathered from MHI (e.g., transcripts, message logs). Several exclusion criteria were also defined: (a) study of human-computer interventions; (b) text-based data not derived from human-to-human interactions (i.e., medical records, clinician notes); (c) social media platform content (e.g., Reddit); (d) population other than adults (18+); (e) did not analyze data using NLP; or (f) was a book chapter, editorial article, or commentary. Candidate manuscripts were evaluated against the inclusion and exclusion criteria initially based on their abstract and then on the full-text independently by two authors (JMZ and MM), who also assessed study focus and extracted data from the full text. Disagreement on the inclusion of an article or its clinical categorization was discussed with all the authors following full-text review. When more than one publication by the same authors used the same study aim and dataset, only the study with the most technical information and advanced model was included, with others classified as a duplicate and removed. Reasons for exclusion were recorded.

Data extraction

Studies that met criteria were further assessed to extract clinical and computational characteristics.

Setting and data. The MHI used to generate the data for NLP analyses. Treatment modality, digital platforms, clinical dataset and text corpora were identified.

Study focus. Goal of the study, and whether the study primarily examined conversational data from patients, providers, or from their interaction. Moreover, we assessed which aspect of MHI was the primary focus of the NLP analysis.

Ground truth. How the concepts of interest were operationalized in each study (e.g., measuring depression as PHQ-9 scores). Information on raters/coders, agreement metrics, training and evaluation procedures were noted where present. Information on ground truth was identified from study manuscripts and first order data source citations.

Natural language processing components. We extracted the most important components of the NLP model, including acoustic features for models that analyzed audio data, along with the software and packages used to generate them.

Classification model and performance. Where multiple algorithms were used, we reported the best performing model and its metrics, and when human and algorithmic performance was compared.

Reproducibility. Information on whether findings were replicated using an external sample separated from the one used for algorithm training, interpretability (e.g., ablation experiments), as well as if a study shared its data or analytic code.

Limitations and biases. A formal assessment of the risk of bias was not feasible in the examined literature due to the heterogeneity of study type, clinical outcomes, and statistical learning objectives used. Emerging limitations of the reviewed articles were appraised based on extracted data. We assessed possible selection bias by examining available information on samples and language of text data. Detection bias was assessed through information on ground truth and inter-rater reliability, and availability of shared evaluation metrics. We also examined availability of open data, open code, and for classification algorithms use of external validation samples.

RESULTS

The initial literature screen delivered 19,756 candidate studies. After 4677 duplicate entries were removed, 15,078 abstracts were screened against inclusion criteria. Of these, 14,819 articles were excluded based on content, leaving 259 entries warranting full-text assessment. The screening process is reported in Fig. 1, with the final sample consisting of 102 studies (Table 1).

Study characteristics

Publication year. Results indicate a growth of NLP for MHI applications, with the first study appearing in 2010 and the majority being published between 2020–2022 (53.9%, $n = 55$). The median year of publication was 2020 (IQR = 2018–2021), a trend consistent with NLP advancements [32].

Setting and data. The majority of interventions consisted of synchronous therapy (53.9%, $n = 55$), with Motivational Interviewing as the most reported therapy modality ($n = 20$). These studies primarily involved face-to-face randomized controlled trials, traditional treatments, and collected therapy corpora (e.g., Alexander Street Corpus). Transcripts of clinical assessments, interviews, and structured tasks were another important source of textual data (20.6%, $n = 21$), elicited through the use of standardized prompts and questions. While most face-to-face studies used text data from manual transcripts, 18 studies used machine-transcription generated from audio sources [36, 43–59]. Online message-based interventions were the second largest setting (22.6%, $n = 23$), with text-data consisting of anonymized conversation logs between providers and patients. Sample sizes increased from less than 100 therapy transcripts [45, 60, 61] to over 100,000 [34, 62–64], with studies analyzing more than one million conversations [65, 66].

Ground truth. Clinicians provided ground truth ratings in the form of diagnoses, assessments, or suicide risk for 31 studies. Patients provided ground truth for 29 studies, through self-report measures of symptoms and functioning ($n = 22$), intervention feedback, and treatment alliance ratings. Students ($n = 9$), researchers ($n = 6$), crowd-workers ($n = 3$), and other raters ($n = 26$) provided treatment annotations and emotion/sentiment analysis. As the modal intervention, Motivational Interviewing Skills Codes (MISC) [67] annotations were the most prevalent source of provider/patient information. Thirty-two studies provided information on rater/coder agreement, with adequate inter-rater reliability across studies for frequent and aggregated codes. Only 20 studies provided information on the raters' training or selection, with Sharma et al., describing in detail an interactive training consisting of instructions, supervision, evaluation, and final selection [62]. Combined human and deep-learning-based approaches were also explored as an alternative to producing a large amount of treatment-related labels. In particular, Ewbank

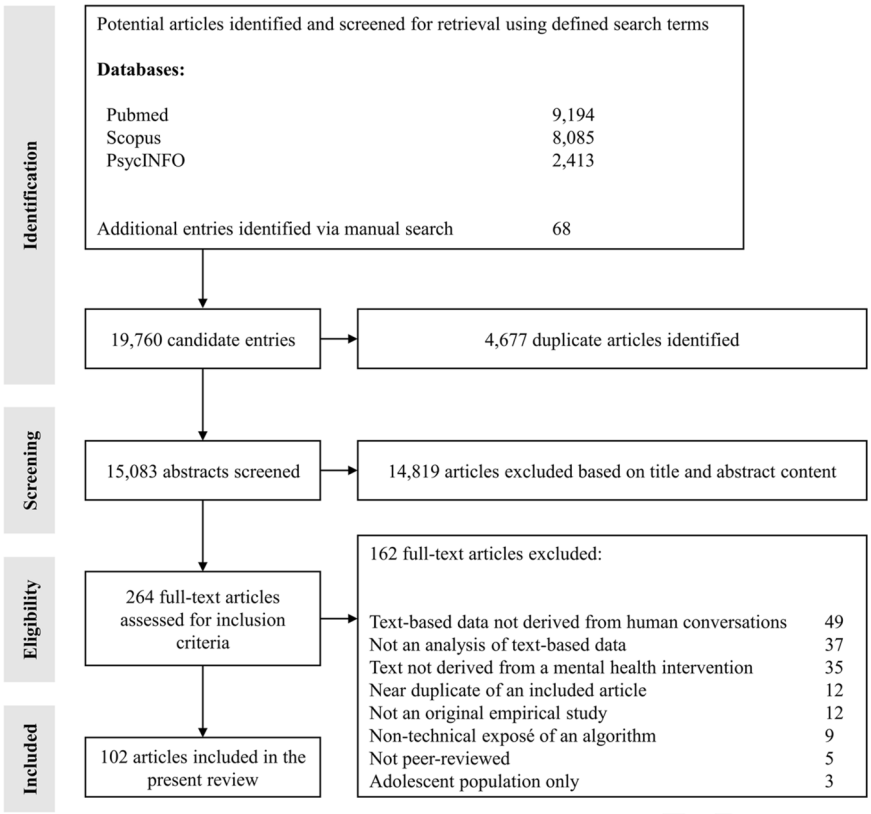


Fig. 1 PRISMA flow diagram.

and collegues [34, 35] used a hybrid approach to generate ground truth: human raters annotated a portion of sessions and an annotation model was based on their inputs to label a larger number of sessions.

Natural language processing and machine learning components

Multiple NLP approaches emerged, characterized by differences in how conversations were transformed into machine-readable inputs (linguistic representations) and analyzed (linguistic features). Linguistic features, acoustic features, raw language representations (e.g., tf-idf), and characteristics of interest were then used as inputs for algorithmic classification and prediction. Methods used mirrored the development of NLP tools through time (Fig. 2).

Language representation. The majority of studies ($n = 53$) tabulated the frequency of individual words through the use of lexicons or dictionaries. Forty-three studies (42.6%) used n-grams for language representation due to their simplicity and interpretability. Bag of Words and Term Frequency-Inverse Document Frequency (TF-IDF) were used by 30 studies to model word frequencies directly for classification purposes [37].

After raw word counts, Word Embeddings were the most commonly utilized language representation ($n = 49$, 48%), owing to its advantages for performing analytic operations. Lower-dimensional embeddings were primarily generated using word2-vec and GloVe algorithms. With recent advances in deep learning, more sophisticated Transformer architectures (e.g., RoBERTa) produced contextualized embeddings, where the representation of a word or token depends on its surrounding context.

Model features. The most common linguistic features were based on lexicons ($n = 43$) computing the frequency of words by their membership in categories designed by domain experts. This approach is exemplified by software such as LIWC [30], and owes its diffusion in clinical research to its ease of use and low technological requirements. Another prevalent NLP task was sentiment analysis ($n = 32$), which generated feature scores for emotions (e.g., joy, annoyance) that are derived from lexicon-based methods and pre-trained models (e.g., VADER). Topic modeling ($n = 16$) also emerged as a widely used approach to identify common themes across clinical transcripts.

Deep learning features. More recent technological developments saw the rise of features based on deep neural networks ($n = 40$). The adoption of large language models grew in parallel with increases in computational power, the development of dedicated code libraries (e.g., Pytorch and Tensorflow), and increased availability of large MHI corpora (Fig. 2). Transformer models were the most used language models given their ability to generate contextually-meaningful linguistic features from sequences of text through the use of attention mechanisms, and to study the flow of individual talk turns [68], as well as its effects on overall session estimates [48].

Models deployed include BERT and its derivatives (e.g., RoBERTa, DistillBERT), sequence-to-sequence models (e.g., BART), architectures for longer documents (e.g., Longformer), and generative models (e.g., GPT-2). Although requiring massive text corpora to initially train on masked language, language models build linguistic representations that can then be fine-tuned to downstream clinical tasks [69]. Applications examined include fine-tuning BERT for domain adaptation to mental health language (MentalBERT) [70], for sentiment analysis via transfer learning (e.g., using the GoEmotions corpus) [71], and detection of topics [72]. Generative language models were used for revising interventions [73], session summarizations [74], or data augmentation for model training [70].

Table 1. Summary of included studies.

Study Overview			Natural Language Processing Components			Machine Learning & Algorithmic Components			Results & Reproducibility	
Reference	**Study Aim**	**Data**	**Language Representation**	**Model Features**	**Algorithm / Software**	**Ground Truth**	**Prediction Model**	**Algorithm Performance Metric**	**Practical Results**	**Reproducibility**
Alonso-Sánchez et al., 2022a [87]	Monitor illness stage in schizophrenia	Transcripts of the Thematic Apperception Test from 46 individuals with first episode schizophrenia and 36 healthy controls at acute and after 6 months of treatment	Word counts; Word embeddings	Semantic similarity	: Covington Vector semantic tool (GloVe)	Provider: Diagnosis (Structured Clinical Interview for DSM-5) Patients: Symptoms (Positive and Negative Syndrome Scale-8)	Bayesian t-test	*Patients vs. controls*: BF_{10}=6.53/6.3 *PANSS-8 positive*: t=.39/-.1; BF_{10}=9.24/.44 *PANSS-8 negative*: t=.08/.59, BF_{10}=.18/1&7 (baseline/follow-up)	Semantic similarity from descriptive discourse could identify functional deficits in schizophrenia and follow the trajectory of negative symptoms	Code: Data: External Validation: Interpretability:
Alonso-Sánchez et al., 2022b [92]	Study semantic distance and neuroimaging correlates in first episode schizophrenia	Transcripts of Thematic Apperception Test from 30 untreated individuals with first episode schizophrenia and 30 healthy controls	Word counts; Word embeddings	Semantic similarity	: GloVe	Providers: Schizophrenia Diagnosis Researchers: functional imaging markers (connectivity, inferior frontal gyrus and ventral anterior temporal lobe)	Bayesian t-test; spectral dynamic causal modelling	*Patients vs. controls*: δ [95% CI]=[-1.71,-.59]; Bayes Factor= 1623 *Connectivity*: free energy= 80/0 (model with/out semantic similarity)	Cortical connectivity was better explained by differences in semantic similarity along with diagnosis, compared to diagnostic status alone.	Code: Data: External Validation: Interpretability:
Althoff et al., 2016 [33]	Analyze therapy discourse and predict outcome	15,555 messages from 408 crisis center counselors	Word counts; N-grams; TF-IDF	lexicon : sentiment ; topic modeling	: LIWC : VADER / : Hidden Markov Model	Patients: conversation helpfulness ("I feel better" / "I don't feel better")	Logistic Regression	Accuracy=.68; AUC=.72	More effective crisis counselors show greater adaptability and linguistic creativity, faster problem identification and solution generation, and higher rates of reflections and affirmations.	Code: - Data: - External Validation: - Interpretability: -
Arevian et al., 2019 [75]	Monitor serious mental illnesses longitudinally	1101 recordings of phone calls between case managers and 47 patients from a community-based clinic	Word counts; N-grams; part-of-speech	lexicon : sentiment ; latent semantic analysis; acoustic features	: LIWC : Praat	Providers: clinical state (Global assessment rating, 1-10) Patients: symptoms (BASIS-24, SF-12 mental health subscale)	Support Vector Machine	*Clinical state* (concurrent/forecast): rho=.78/.33 *BASIS-24 score*: rho=.25 *SF-12 score*: rho=.25	Analyzing patient speech samples may be a feasible adjunct to patient monitoring in community-based clinical settings.	Code: - Data: - External Validation: - Interpretability: -
Aswamenakul et al., 2018 [43]	Classify patient utterance-level inclinations (behavioral change)	187 audio recordings and 193 transcripts of motivational interviewing sessions	Word counts; Word embeddings	lexicon ; acoustic features	: GloVE : LIWC : COVAREP : Speechmatics	Students: Motivational Interviewing Skills Code (MISC 2.0) annotations	Logistic Regression	Accuracy=.63; F1=.57; Precision=.56; Recall=.57	Using both acoustic and linguistic features can increase the accuracy of speech classifier algorithms.	Code: - Data: - External Validation: - Interpretability: Model ablation
Atkins et al., 2012 [136]	Identify psychotherapy topics	2,251 transcripts from couples therapy sessions and between-session (118 couples)	Word counts; bag-of-words	topic modeling	: Latent Dirichlet Allocation	Couple Interaction Rating System (CIRS) annotations	Sparse Logistic Regression	Accuracy=.65-.70	Topic models are a promising method for classifying linguistic behaviors in psychotherapy.	Code: - Data: - External Validation: - Interpretability: -
Atkins et al., 2015 [114]	Evaluate treatment fidelity	148 transcripts of motivational interviewing sessions for substance abuse (29,590 talk turns)	Word counts; N-grams; bag-of-words	topic modeling	: Latent Dirichlet Allocation	Motivational Interviewing Skills Code (MISC 2.1) annotations	Labeled Topic Modeling	AUC=.72	The reliability of therapist acts classifier algorithms can vary widely by the specific linguistic behavior being classified (e.g., open-ended questions, complex reflections).	Code: - Data: - External Validation: - Interpretability: -
Atzil-Slonim et al., 2021 [127]	Predict ruptures and functioning using conversation topics	873 transcripts of psychodynamic psychotherapy transcripts from 58 clients treated by 52 therapists	Word counts; bag-of-words; part-of-speech	topic modeling	: YAP : MALLET	Patients: functioning (Outcome Rating Scale, ORS; score<24); Providers: ruptures (Post Session Questionnaire; score>1)	Sparse Multinomial Logistic Regression	*ORS*: Accuracy=.76 *Ruptures*: Accuracy=.65	Topic models can provide insights into the therapeutic process with relatively small datasets.	Code: - Data: - External Validation: - Interpretability: -
Baggott et al., 2015 [106]	Investigate the effects of MDMA on speech	35 transcripts of the Interpersonal Perception Task from participants who were administered MDMA or a placebo	Word counts; bag-of words; part-of-speech	lexicon	: NLTK, Gensim (Python 2.7) : LIWC (v1.11)	Providers: MDMA vs. placebo administration	Random Forest	Accuracy=.72; Sensitivity=.71; Specificity=.80	Participants who receive MDMA-assisted therapy speak much more about death but also used more prosocial words versus controls.	Code: - Data: - External Validation: - Interpretability: -
Bantilan et al., 2020 [37]	Detect and score suicidal ideation	85,216 message therapy transcripts from 1,864 Talkspace users	N-grams; TF-IDF	lexicon	: : NLTK (Python)	Providers: Suicide risk assessment (i.e., "no risk", "risk factors", "suicidal ideation", "method, and plan")	Logistic Regression	AUC=.83	In the context of a text-based psychotherapy platform, client suicide risk can reliably be estimated through a natural language processing algorithm.	Code: - Data: - External Validation: - Interpretability: -
Burkhardt et al., 2021 [109]	Identify linguistic indicators of behavioral activation	~2.5 million message therapy transcripts from Talkspace users	Word counts; Word embeddings	lexicon : sentiment	: Semantic Vectors (Java) : LIWC : Empath	Patients: Symptoms (PHQ-9; Behavioral Activation for Depression Scale)	Mixed-effects model	-	Both previously established linguistic markers of depression (i.e., LIWC) and novel linguistic indicators of behavioral activation were strongly associated with depression symptom scores.	Code: - Data: - External Validation: - Interpretability: -
Burkhardt et al., 2022 [71]	Predict depression and anxiety severity	6,500 messaging therapy transcripts from Talkspace users; GoEmotions corpus	N-grams; part-of-speech; Word embeddings	deep learning ; lexicon : sentiment	/ : BERT GoEmotions / : LIWC	Patients: Symptoms (PHQ-9; GAD-7; Diagnosis: score ≥10)	Transformer (BERT)	*Diagnosis*: AUC=.67/.66; F1=.52/.46; Precision=.62/.57; Recall=.45/.38 *Symptoms scores*: R2= .81/.78 (PHQ-9/GAD-7)	Deep learning features outperformed word counting methods for symptoms prediction. Depression severity was associated with differences in grief, pride, excitement, relief, and disgust.	Code: - Data: - External Validation: - Interpretability: SHAP values
Can et al., 2016 [115]	Identify utterance-level interventions (reflections)	57 transcripts of motivational interviewing sessions from 3 clinical trials	N-grams; part-of-speech	-	-	Students: Motivational Interviewing Skills Code (MISC 2.1) annotations (counselor: reflections)	Maximum Entropy Markov Model	F1=.81; Sensitivity=.93; Specificity=.90; Precision=.73	Therapy speech acts classifier algorithms can be more reliable given more contextual data.	Code: - Data: - External Validation: - Interpretability: -
Cao et al., 2019 [116]	classify and forecast utterance-level treatment fidelity	353 transcripts of motivational interviewing sessions	Word embeddings	deep learning	: GloVE : ELMo	Motivational Interviewing Skills Code (MISC 2.1) annotations	Transformer (ELMo)	*MISC categories*: F1=.65/.53 *MISC forecast*: F1=.31/.44, Recall=.77 (therapist/client codes)	Therapy speech acts classifier algorithms can be enhanced by using a modified versions of analog behavior coding systems.	Code: Available Data: - External Validation: - Interpretability: Model ablation
Carcone et al., 2019 [83]	Automate treatment segmentation	37 transcripts of motivational interviewing with patients in treatment for weight-loss	Word counts	deep learning ; lexicon	: LIWC : CNN	Minority Youth Sequential Code for Observing Process Exchanges (MY-SCOPE) treatment codes	Support Vector Machine	*Caregiver codes*: Accuracy=.66, F1=.64 *human inter-rater reliability*: K=.66 *human-algorithm agreement*: K=.61	A speech acts classifier algorithm was able to reach human-level inter-rater reliability.	Code: - Data: - External Validation: - Interpretability: -
Carrillo et al., 2016 [86]	Classify bipolar patients using speech	40 Transcripts of the Structured Clinical Interview for DSM-5 Disorders from 40 patients (20 diagnosed with Bipolar Disorder)	Word counts; part-of-speech	sentiment ; lexicon	/ : Dictionary of Affect in Language	Providers: Bipolar disorder diagnosis (Structured Clinical Interview for DSM-IV)	Logistic Regression	AUC=.70; F1=.62	Accurately classifying patients on emotional presentation may be less algorithmically challenging than classifying based on mental health diagnoses.	Code: Available Data: - External Validation: - Interpretability: -
Carrillo et al., 2018 [95]	Classify depressed patient and predict remission	35 transcripts from 17 psilocybin trial patients and 18 healthy controls of an autobiographical memory interview (pre-treatment)	Word counts	sentiment	: Emotional Analysis algorithm	Patients: Symptoms (Quick Inventory of Depressive Symptoms QIDS; treatment response: QIDS reduction ≥50%)	Gaussian Naive Bayes Classifier	*Patients vs control*: Accuracy=.83, Precision=.82, Recall=.82 *Treatment response*: Accuracy=.85, Precision=.75	A machine learning algorithm can reliably differentiate depressed patients from healthy controls and distinguish treatment responders from non-responders.	Code: Available Data: - External Validation: - Interpretability: -
Chaoua et al., 2018 [137]	Identify psychotherapy topics	1729 transcripts of psychotherapy sessions (340,455 talk turns)	Word counts; bag-of-words	lexicon : topic modeling	: NLTK (Python) : Topic Modeling Toolbox, Palmetto	-	Partially Labeled Latent Dirichlet Allocation	coherence ≥ 50%	Therapy patients are more likely to topic-switch away from parenting, friendship, and sexual dysfunction, while counselors are more likely to topic-switch away from medication and patient-counselor relations.	Code: Available Data: - External Validation: - Interpretability: -
Chen et al., 2019 [76]	Automate and predict utterance-level treatment fidelity	337 audio recordings and transcripts of motivational interviewing sessions for substance abuse	Word embeddings	deep learning ; acoustic features	: BiLSTM : Kaldi (C++)	Motivational Interviewing Skills Code (MISC 2.0) annotations	GraphSVM	Accuracy=.63; F1=.59	Graph-based machine learning models may outperform strict hierarchical classification approaches to transcript analysis.	Code: Available Data: - External Validation: - Interpretability: -
Chen et al., 2022a [48]	Evaluate CBT sessions quality	1,118 transcripts of CBT sessions from the Beck Community Initiative corpus; 4,268 transcripts sessions from a university counseling center	Word counts; TF-IDF; Word embeddings	deep learning	: doc2vec, glove : BERT, BigBird, Longformer LSTM : CORE-MI	Providers: Cognitive Therapy Rating Scale (CTRS) annotations. (High session quality: CTRS ≥ 40)	Transformer (BERT + LSTM)	F1=75.1; RMSE=8.09; MAE=6.27	A hierarchical deep learning framework examining treatment segments can improve session-level quality monitoring.	Code: - Data: - External Validation: - Interpretability: attention weights
Chen et al., 2022b [119]	Evaluate motivational interviewing treatment fidelity	345 transcripts of motivational interviewing sessions concerning alcohol and drug abuse problems; Switchboard-DAMSL (SwDA) corpus	Word embeddings	deep learning	: BERT, Reptile	Motivational Interviewing Skills Code (MISC) utterance-level codes	Transformer (Reptile meta-learning + BERT)	*Therapist codes*: UAR=.66 *Patient codes*: UAR=.56	Psychotherapy fidelity classifiers may perform better through meta-learning using public datasets from related domains	Code: Data: External Validation: - Interpretability:
Christian et al., 2021 [128]	Identify linguistic features of ruptures	54 transcripts of psychodynamic psychotherapy sessions from 27 client-therapist pairs	Word counts	lexicon : sentiment	: : DAAP	Ruptures (Segmented Working Alliance Inventory - Observer scale; score <4)	Correlation, T test	-	During periods of therapeutic rupture, patients and therapists show a decrease in emotional engagement.	Code: Available Data: - External Validation: - Interpretability: -
Corcoran et al., 2018 [89]	Detect psychosis risk	93 transcripts of open-ended interviews with adult and adolescent participants at risk for psychotic disorder from two sites	Word counts; part-of-speech	latent semantic analysis	: NLTK (Python)	Providers: Psychosis diagnosis (Structured Interview for Prodromal Syndromes; Scale of Prodromal Symptoms)	Logistic Regression	*Psychosis onset*: Accuracy=.79 *Patients vs control*: Accuracy=.72	Psychosis onset was associated with decreased semantic coherence, greater variance in coherence, and reduced usage of possessive pronouns.	Code: - Data: - External Validation: Cross-site validation Interpretability: -
Crangle et al., 2019 [82]	Label emotions in speech	Audio recordings and transcripts of 18 couples therapy sessions (3 couples total)	Word counts	acoustic features	: MATLAB	Emotion category (anger, sadness, joy, tension, neutral) and intensity (low, medium, high)	Random Forest	76-99% pairwise recognition rate	Couples therapy audio recordings can be a fruitful context for building emotion recognition algorithms.	Code: - Data: - External Validation: - Interpretability: -
Demiris et al., 2022 [49]	Predict anxiety and quality of life scores	Audio recordings and transcripts of 124 sessions with family caregivers of hospice patients	N-grams	acoustic features	/ : Automated Speech Recognition (fine-tuned DeepSpeech2)	Patients: Symptoms (GAD-7; score ≥10) and Caregiver Quality of Life Index-Revised (CQLI-R; score≥32)	Logistic Regression	*GAD-7*: Accuracy=.89; Precision= 92; Recall= .88 *CQLI-R*: Accuracy=.76; Precision=.73; Recall=.79	Quality of life in older adults can be classified through linguistic and acoustic features from automated speech recognition system.	Code: - Data: - External Validation: - Interpretability: -
Ding et al., 2022 [70]	Identify cognitive distortions in serious mental illness	Messaging therapy transcripts from 39 individuals with serious mental illness (7,354 messages); test data augmentation (Easy Data Augmentation, Back Translation, GPT-2)	Word embeddings	deep learning	: BERT, MentalBERT	Providers: Cognitive distortions (mental filter, jumping to conclusions, catastrophizing, should statements, overgeneralization)	Transformer (BERT, MentalBERT)	*Distortions*: macro-AUPRC=.38 *human inter-rater reliability*: K=.39-.51	Domain-specific language models may be ideally applied to classify frequently-occurring cases. Data augmentation improved detection of rare classes with base models but GPT-2-based augmentation harmed model performance	Code: - Data: - External Validation: - Interpretability: -
Dirkse et al., 2015 [94]	Identify linguistic features of patients' messages	767 messages from 59 iCBT patients being treated for GAD	Word counts	lexicon : sentiment	/ : LIWC	Patients: Symptoms (PHQ-9; GAD-7; Panic Disorder Symptoms Scale)	Repeated-measures Poisson regression	-	Increased patient usage of negative emotion words were significant predictors of symptom ratings and decreased over the course of therapy treatment.	Code: - Data: - External Validation: - Interpretability: -
Dore & Morris, 2018 [135]	Evaluate linguistic coordination	~1.16 million messages from 169,376 users of an online emotional support platform	Word counts; part-of-speech	lexicon : sentiment ; latent semantic analysis	: LIWC : NRC Emotion Lexicon	Patients: support effectiveness (bad, okay, good), emotional change (better, same, worse)	Regularized Regression	-	Emotional support may be more effective when there is greater linguistic synchrony between providers and help-seekers.	Code: - Data: - External Validation: - Interpretability: -
Ewbank et al., 2019 [34]	Evaluate treatment fidelity and predict outcome	89,394 messages from 14,899 patients receiving iCBT	Word embeddings	deep learning	: word2vec : BiLSTM	Providers: annotations (CBT interventions) Patients: Symptoms (PHQ-9; GAD-7)	Transformer(BiLSTM) Logistic Regression	*Annotations*: Precision=.52-1; Sensitivity=.15-1; Specificity=.79-1 *Human interrater agreement*: κ=.54	Higher rates of agenda setting, and therapist praise were positively associated with symptom improvement, while greater therapist empathy and risk checking was linked to worse outcomes.	Code: - Data: - External Validation: - Interpretability: -

Table 1. continued

Ewbank et al., 2020 [35]	Evaluate patient treatment response	Messages from 25,366 patients treated with iCBT	Word embeddings	deep learning	word2vec BiLSTM	Researchers: utterance-level patient change-talk Patients: Symptoms (PHQ-9; GAD-7); engagement (2+ sessions)	Transformer(BiLSTM) Logistic Regression	*Change-talk*: F1=.22–.94; Precision=.17–.97; Recall=.33–.91 *Human interrater agreement*: κ=.4–.6 *Human-algorithm agreement*: κ=.4–.5	More patient change talk was associated with greater treatment engagement, while neutral and counter-change talk was associated with lower engagement.	Code: - Data: - External Validation: - Interpretability: -
Flemotomos et al., 2021a [53]	Evaluate cognitive behavioral therapy (CBT) treatment fidelity	1,118 recordings of CBT sessions from the Beck Community Initiative; 4,268 transcripts of sessions from a university counseling center	Word embeddings; TF-IDF	deep learning	BERT CORE-MI	Providers: Cognitive Therapy Rating Scale (CTRS) codes	Transformer (BERT+ gated recurrent units)	CTRS session score ≥40: F1=.73	A combination of different linguistic representations and therapy metadata may be best suited for classifying therapeutic speech acts	Code: - Data: - External Validation: - Interpretability: -
Flemotomos et al., 2021b [52]	Evaluate utterance-level interventions	5,097 recordings of university center counseling sessions and transcripts from Alexander Street Press	Word embeddings	deep learning; acoustic features	Bidirectional LSTM ; CORE-MI	Motivational Interviewing Skills Code (MISC 2.5) annotations	Transformer (Bidirectional LSTM with attention)	MISC utterance-level: F1=.52	Automatic classification of psychotherapy session content can be performed using audio data.	Code: - Data: - External Validation: - Interpretability: -
Gaut et al., 2017 [138]	Identify topics and emotions in psychotherapy	1,181 transcripts of psychotherapy from Alexander Street Press	Word counts; N-grams; bag-of-words	topic modeling	Latent Dirichlet Allocation	Alexander Street Press: session content codes Coders: talk turn content (anger, anxiety, depression, low self-esteem, suicidality)	Labeled Topic Modeling	*Session codes*: AUC=.79 *Model talk-turns codes*: AUC=.67–.75, Precision=.25–.44 *Human coding reliability*: AUC=.81–.87, Precision=.45–.68	Therapy speech acts classifier can be improved by using topic modeling but is not yet as reliable as humans for talk-turn annotations	Code: - Data: - External Validation: - Interpretability: -
Gibson et al., 2016 [124]	Predict therapists' empathy	337 transcripts of motivational interviewing sessions for substance abuse	Word embeddings	deep learning	word2vec (Python) LSTM	Empathy (Motivational Interviewing Treatment Integrity manual; high: score>4); Motivational Interviewing Skills Code (MISC) annotations	LSTM	*Empathy*: UAR=.79 *MISC*: F1=.26/.64; Precision=.35/.68; Recall=.23/.62 (MISC28/MISC8)	A simplified motivational interviewing coding scheme outperformed the original Motivational Interviewing Skills Code (MISC) in a speech acts classification algorithm.	Code: - Data: - External Validation: - Interpretability: -
Glauser et al., 2019 [103]	Classify history of suicidality	122 transcripts of the Mini International Neuropsychiatric Interview administered to study participants	Word counts; N-grams	lexicon	LIWC	Providers: Suicidality and anxiety/depression/bipolar diagnosis (Mini International Neuropsychiatric Interview)	Support Vector Machine	AUC=.78	Natural language processing can reliably identify current or lifetime history of suicidality and depression in people with epilepsy.	Code: - Data: - External Validation: - Interpretability: -
Goldberg et al., 2020a [126]	Evaluate therapist facilitative interpersonal skills	165 transcripts of participants engaging in the Facilitative Interpersonal Skills task	Word counts; TF-IDF	-	-	Students: Facilitative Interpersonal Skills Task ratings (FIS)	Elastic net	*FIS*: rho=.27–.53; R2=.13–.24 *Human interrater agreement*: ICC=.87–.93 *Human-algorithm agreement*: ICC=.31–.60	Automated assessment of constructs like provider interpersonal skills may be best applied in more standardized contexts (e.g., a behavioral task) versus naturalistic psychotherapy.	Code: - Data: - External Validation: - Interpretability: -
Goldberg et al., 2020b [36]	Estimate therapeutic alliance	1,235 audio recordings from psychotherapy sessions of 386 clients seen by 4 therapists	N-grams; TF-IDF; Word embeddings	-	Sent2vec (Python) ; Kaldi (C++)	Patients: Therapeutic alliance (Working Alliance Inventory-Short Form)	Ridge Regression	Spearman's P=.15	An algorithm could predict client-rated therapeutic alliance scores from transcript data.	Code: - Data: - External Validation: - Interpretability: -
He et al., 2015 [97]	Identify linguistic markers of PTSD	300 transcripts of study participants' self-narrative remote interviews	N-grams, bag-of-words	-	: NLTK (Python)	Providers: PTSD diagnosis (Structured Clinical Interview for DSM-IV)	Product Score Model	AUC=.94; Accuracy=.82; Recall=.85; Specificity=.81	Textual assessment on self-narratives is a promising tool for early stage diagnosis of PTSD.	Code: - Data: - External Validation: - Interpretability: -
Hoogendoorn et al., 2017 [141]	Predict treatment outcome from patient-therapist email exchanges	Emails from 69 German speaking patients with Social Anxiety Disorder	Word counts; bag-of-words, part-of-speech	sentiment; topic modeling	: NLTK (Python) : pattern.de (Python) : Latent Dirichlet allocation	Patients: Symptoms (Social Phobia Scale; reliable change index yes/no)	Logistic Regression	AUC=.83; F1=.82; Precision=.90; Recall=.75	Six weeks of usage and content of patient emails predicted reliable improvement in an internet-based treatment	Code: - Data: - External Validation: - Interpretability: -
Howes et al., 2014 [96]	Predict depression and anxiety severity	882 treatment patient-therapist dialogues in text-based iCBT treatment	Word counts; N-grams; bag-of-words	lexicon; sentiment; topic modeling	; LIWC, Weka : MALLET	Patients: Symptoms (PHQ-9)	Support Vector Machine	*Baseline symptoms*: F1=.74 *Non-improving*: F1=.71	Standard topic, sentiment and emotion modelling can be usefully applied to online text therapy dialogue.	Code: - Data: - External Validation: - Interpretability: -
Hudon et al., 2022 [121]	Classify themes in virtual reality therapy	162 transcripts of Avatar Therapy (AT) sessions of 18 patients with treatment-resistant schizophrenia	TF-IDF	-	: Scikit-learn (Python)	Researchers: annotations of 28 AT therapeutic themes.	Support Vector Machine	*patient/therapist themes*: F1=.15–.79 / 1-1 *human inter-rater agreement*: Scott's Pi=.59 *human-algorithm agreement*: Scott's Pi=.65	Automated annotation reached an agreement in the same range as human agreement and may be useful to study digital interventions for serious mental illness	Code: - Data: - External Validation: - Interpretability: -
Hull et al., 2021 [64]	Analyze the impact of COVID-19 on anxiety and depression symptoms	219,156 message therapy transcripts from 169,889 Talkspace users	N-grams, bag-of-words	topic modeling	; SpaCy, Textacy (Python) : Latent Dirichlet allocation	Patients: Symptoms (PHQ-9, GAD-7)	Labeled Topic Modeling	-	Therapy treatment seekers after March, 2020 presented with more severe intake anxiety levels than before the COVID-19 outbreak.	Code: - Data: - External Validation: - Interpretability: -
Imel et al., 2015 [140]	Classify therapy by theoretical modality and identify topics	1,398 transcripts of psychotherapy sessions from Alexander Street Press; 148 transcripts from motivational interviewing sessions from 5 separate trials	Word counts; bag-of-words, part-of-speech	topic modeling	: Latent Dirichlet allocation	Providers: Treatment approach (Psychodynamic / CBT / Humanistic / Other / Drug therapy)	Random Forest	Accuracy=.87	Topic models can be used to discriminate between different psychotherapeutic approaches.	Code: - Data: - External Validation: - Interpretability: -
Just et al., 2020 [84]	Study speech abnormalities in non-affective psychosis	60 Transcripts of the Narrative Emotion Task from 40 individuals with schizophrenia or schizoaffective disorder and 20 healthy controls	Word counts; Word embeddings	sentiment; lexicon; referential abnormalities; neologisms	; Deepset, GloVE (Python) ; LIWC	Providers: Formal thought disorder (Scale for the Assessment of Positive/Negative Symptoms)	Multinomial Logistic Regression	*Patient vs control*: Accuracy=.70/.50 (psychosis with/out thought disorder)	Algorithmically derived measures of linguistic coherence failed to predict non-affective psychosis.	Code: - Data: - External Validation: - Interpretability: -
Kshirsagar et al., 2017 [106]	Detect crisis situations	106,000 text entries from users of an online emotional support platform	Word embeddings	deep learning	; GloVE, Spacy (Python) ; RNN + attention	Crowdworkers: state of crisis yes/no (e.g., suicide)	Transformer (RNN + Attention)	F1=.80	Attention-based neural networks may be favored over logistic regression in the context of rapidly identifying crisis language in real-time.	Code: - Data: - External Validation: - Interpretability: LIME
Lee et al., 2019 [113]	Automate treatment segmentation	4,000 transcripts of psychotherapy sessions from Alexander Street Press	Word counts; N-grams; Word embeddings	deep learning; lexicon; sentiment; dialogue features (speaker change, turn index, utterance position)	; CNN ; DAAP, LIWC ; SentiWordNet	Students: Treatment fidelity (clinical codes)	Support Vector Machine	Accuracy=.75; F1=.70; Precision=.71; Recall=.70	Support vector machines may be better suited to therapy act classification tasks than convolutional neural networks.	Code: - Data: - External Validation: - Interpretability: model ablation
Liu et al., 2021 [122]	Classify conversational stages and behaviors in crisis texts	10,809,178 message transcripts of 271,445 conversations from Shout mental health crisis text line (169 labeled for classification)	Word embeddings	deep learning	; Longformer	Providers: message-level annotations of conversational stage and behavioral codes Users: autism diagnosis (self-declared)	Transformer (Longformer)	*conversational stages*: accuracy=.89–.90 *conversational behaviors*: accuracy=.077; 1-hamming=.5; LRAP=.95 *self-identified diagnosis*: accuracy=.95	NLP can predict conversation stages and behaviors at a message-level, which could potentially be used as markers of conversation quality in text crisis conversations	Code: - Data: - External Validation: - Interpretability: LIME
Malins et al., 2022 [110]	Classify talk-turn level patient activation	128 transcripts from CBT sessions delivered remotely to 53 patients with severe health anxiety	Word counts; N-grams	lexicon; sentiment	; NLTK, TextBlob (Python)	Researchers: Consultation Interactions Coding Scheme annotations	bagged trees	Accuracy=.81; precision=.87; recall=.75; F1=.87	NLP can be used to discriminate between high and low patient activation from turns of speech. Including key stakeholders in model development can enhance NLP predictive accuracy and clinical utility	Code: - Data: - External Validation: - Interpretability: -
Mao et al., 2022 [77]	Predict depression diagnosis and severity ratings	189 Audio recordings and transcripts of clinical interviews from the Distress Analysis Interview Corpus dataset	Word embeddings	acoustic features; deep learning	; GloVe ; COVAREP ; CNN, LSTM, BiLSTM	Patients: Symptoms (PHQ-8*)	Transformer (Bidirectional LSTM with attention)	Accuracy= 96; F1=.96; sensitivity=.98; specificity=1	A multimodality approach combining audio and text features outperformed single modality models for automated depression detection	Code: - Data: Publicly available External Validation: - Interpretability: -
Martinez et al., 2019 [46]	Estimate working alliance	802 transcripts of university counseling sessions (31 university counselors and 204 clients)	N-grams, bag-of-words	personae topic modeling	: CoreNLP (Python) ; Kaldi (C++)	Patients: Therapeutic alliance (Working Alliance Inventory-Short Form)	Linear Mixed Effect Models	MSE=.69	Working alliance strength can be predicted by an algorithm trained to label therapists and clients based on in-session personas.	Code: - Data: - External Validation: - Interpretability: -
Mehta et al., 2022 [123]	Classify therapist interventions by theoretical orientation	243 transcripts from psychotherapy sessions collected from a university counseling center	Word embeddings	deep learning	; RoBERTa	Students: Multitheoretical List of Therapeutic Interventions (MULTI-30) talk-turn annotations	Transformer (RoBERTa)	*MULTI-30 codes*: Accuracy=.79; $F1_{micro}$=.50; $F1_{weighted}$=.79 *Human inter-rater reliability*: K=.37–.63	Large language models may be used to adequately classify interventions in psychotherapy interventions but gaps in performance remain	Code: - Data: - External Validation: - Interpretability: Qualitative analysis of model predictions
Mieskes et al., 2019 [44]	Identify therapist utterances	35 transcripts of psychotherapy for German patients with schizophrenia	Word counts; part-of-speech	lexicon; sentiment; disfluences	: DKPRo Core ; Weka ; ELAN	Providers: Treatment qualities (e.g., positive focus)	Random Forest	F1=.77; Recall=.78; Precision=.77	Machine learning can be used to distinguish between therapist and client utterances.	Code: Available Data: - External Validation: - Interpretability: -
Min et al., 2021 [54]	Compare manual and automated transcriptions for classifying therapist utterances	213 recordings and transcripts from counseling sessions	Word counts; Word embeddings	Lexicon; sentiment; acoustic features; deep learning	; LIWC ; Jewer ; Google Speech-to-Text	Motivational Interviewing Treatment Integrity (MITI) codes	Transformer (BERT)	*manual transcripts*: Accuracy=.69; F1=.01–.88 *speech-to-text*: Accuracy=.56; F1=.00–.78 *speech-to-text+context*: Accuracy=.58; F1=.28	Speech-to-text can result in noisy transcriptions with high word error rate, but this issue can be alleviated by providing the classifier with local context (i.e., previous and following utterances).	Code: Data: External Validation: - Interpretability:
Miner et al., 2020 [47]	Test accuracy of speech recognition	100 audio recordings of therapy sessions (100 patients and 78 therapists)	Word embeddings	depression-specific lexicon; semantic distance	word2vec ; Google Cloud Speech-to-Text	Manually-generated transcripts	Two-tailed Welch's t-test and Mann–Whitney U-test	*Overall transcription*: word error rate=25% *Depression keywords transcription*: recall=.80; precision=.83 *Harm-related sentences*: word error rate= 34%	Automatic speech recognition may support understanding of language patterns but may not be ready for individual-level safety surveillance.	Code: Available Data: - External Validation: - Interpretability: -
Miner et al., 2022 [50]	Evaluate therapist language as it relates to patient symptoms	98 transcripts from psychotherapy sessions collected from 24 college counseling sites	Word counts; N-grams	Lexicon; sentiment	; EmoLex ; LIWC ; Google Cloud Speech-to-Text	Patients: Symptoms (PHQ-9)	Logistic regression, PCMCI	PHQ-9≥10: accuracy=.72	Therapists' language responds to patients' speech and diagnosis; therapist and client linguistic patterns display temporally complex interactions	Code: Publicly available Data: - External Validation: - Interpretability:
Mota et al., 2022 [90]	Examine the interaction between linguistic connectedness and emotion with negative symptoms	57 transcripts of responses to emotion-eliciting pictures from 24 treatment-seekers for psychotic symptoms and 33 healthy controls	Word counts; N-grams; part-of-speech	Sentiment; graph features (word node connectedness)	; LIWC ; speechgraph	Patients: Symptoms (PANSS)	Partial Spearman correlation	Rho=.50, p=.02	A speech elicitation protocol (based on positive affective pictures) captured linguistic connectivity features directly linked with negative symptoms.	Code: Data: External Validation: Interpretability:
Nasir et al., 2019 [134]	Measure therapist-patient linguistic coordination	145 transcripts of motivational interviewing sessions; 574 transcripts of couples therapy sessions	Bag-of-words; Word embeddings	word mover distance, linguistic distance	word2vec (Python)	Therapist Empathy (Motivational Interviewing Treatment Integrity code); couple positive/negative affect (Social Support Interaction Rating System codes)	Spearman correlation	*Empathy*: rho= -.26 *Positive/Negative affect*: rho= - .31/.34	Normalized Conversational Linguistic Distance may be a useful measure of interpersonal behavior in the context of psychotherapy.	Code: - Data: - External Validation: Two independent samples Interpretability: -
Nitti et al., 2010 [61]	Identify therapy conversational patterns	43 transcripts of psychodynamic psychotherapy sessions between a single therapist and patient	Word counts; part-of-speech; bag-of-words	lexicon; discourse flow; topic modeling	/; T-Lab 5.3	Dialogue stages (Two-stage semiotic model; deconstructive/constructive)	Perceptron Neural Network	-	Two distinct stages emerged over course of the single patient's progress in psychotherapy.	Code: - Data: - External Validation: - Interpretability: -
Nook et al., 2022 [112]	Assess the impact of linguistic distancing on therapy outcomes	6,229 Messaging therapy transcripts from Talkspace users	N-grams	Lexicon	; LIWC	Patients: Symptoms (PHQ-9; GAD-7)	Mixed-effects modeling	*Between-person*: R2=.04 *Within-person*: R2=.01	Patients' linguistic distance increased over the course of therapy and was related to symptom reductions. However, no consistent evidence emerged that linguistic distance mediated clinical outcomes	Code: Publicly available Data: Partially available External Validation: - Interpretability: -

Table 1. continued

Norman et al., 2020 [107]	Identify personality and emotional changes during PTSD treatment	Asynchronous digital treatment transcripts from 23 combat veterans and military sexual assault survivors	Word counts	sentiment ☺; lexicon 🅻	🅻: LIWC, IBM Watson Personality Insights ☺: IBM Watson Tone Analyzer	Patients: Symptoms and functioning (PHQ-9, GAD-7, PCL-5, PSOMS, PTGI, COPE, SWEMWS)	rMANOVA	-	An algorithm can detect shifts in personality traits, personal values and needs, and emotional expressiveness throughout mental health treatment.	Code: - Data: - External Validation: Two independent samples Interpretability: -
Pulaniyappan et al., 2019 [91]	Analyze speech connectedness and neuroimaging correlates in severe mental illness	Transcripts of Thematic Apperception Test responses from 34 patients with schizophrenia and 22 with patients bipolar disorder	Word counts; part-of-speech 🧠	Graph features (word node connectedness)	🧠: speechgraph	Researchers: functional imaging markers (gyrification index, variance of the degree centrality of the core hub)	Pearson correlation	R=.49, p<.001	Graph theory can help establish the relationship of linguistic connectedness with behavioral, functional, and neuroimaging correlates.	Code: Data: External Validation: Interpretability:
Park et al., 2019 [111]	Classify patient utterances	1,488 transcripts from counseling conversations on Trost - a Korean online counseling platform	Word embedding 🅸	deep learning 𝒜ₙ	🅸: seq2seq 𝒜ₙ: ConvMFIT network	Providers: intervention response (factual, anecdotal, problems, change, process)	Conversation Model Fine-Tuning Network	F1=.64; Precision=.72; Recall=.60	Therapy speech acts classifier algorithms can be built based off of partially pre-trained conversational models.	Code: - Data: - External Validation: - Interpretability: -
J. Park et al., 2021 [131]	Classify emotional valence	353 transcripts from primary care visits between 350 patients and 84 physicians	Word embeddings	deep learning 𝒜ₙ	𝒜ₙ: Hierarchical RNN, AllenNLP	Students: emotional valence of utterances (-3 to +3)	Hierarchical Recurrent Neural Network (RNN)	One vs Rest correlation =.60/.60 (humans/RNN)	A neural network can reach human-level performance at predicting emotional valence.	Code: - Data: - External Validation: - Interpretability: -
Perez-Rosas et al., 2017a [117]	Classify and predict therapist techniques	227 transcripts of motivational interviewing	Word counts; N-grams; part-of-speech 🧠	lexicon 🅻; linguistic style matching	🧠: Stanford Parser 🅻: LIWC	Providers: Motivational Interviewing Treatment Integrity Manual (MITI 4.0) annotations	Support Vector Machine	*Reflections (simple/complex)*: F1=.61/.63 *Human inter-rater reliability*: ICC=.89/.97	Richer linguistic features can improve classification of counselor utterances.	Code: - Data: - External Validation: - Interpretability: Model ablation
Perez-Rosas et al., 2017b [78]	Identify and predict counselor session-level empathy	276 transcripts and audio recordings from motivational interviewing counseling sessions	N-grams, bag-of-words	lexicon 🅻; linguistic style matching & coordination; topic modeling 💬; acoustic features 🔊	💬: Meaning Extraction Helper 🅻 : LIWC 🔊: OpenEar toolkit (EmoLarge), Voicebox Speech Processing	Empathy (Motivational Interviewing Treatment Integrity Manual 4.1; high: score>3)	Random Forest	*Empathy (high/low)*: Accuracy=.80; F1=.87/.71 *Human inter-rater reliability*: ICC=.60	More empathic counselors speak considerably less than non-empathic counterparts and exhibit higher linguistic style coordination and vocal pitch correlation with clients.	Code: - Data: - External Validation: - Interpretability: -
Pérez-Rosas et al., 2018 [55]	Identify markers of session quality	151 recordings from web-sourced motivational interviewing sessions	Word counts; N-grams	lexicon 🅻; sentiment ☺; conversation features (turns, turn word ratios)	🅻/☺: LIWC, CoreNLP (Python) 🔏: Google Speech-to-Text	Students: counseling quality (Motivational Interviewing Treatment Integrity annotations)	Support Vector Machine	*Counseling quality*: Accuracy=.87, F1=.87 *Human inter-rater reliability*: ICC=.94-.96	Standard linguistic features (e.g., N-grams) can be equally predictive of counseling conversation quality as behavioral coding systems (e.g., MITI).	Code: - Data: Available External Validation: - Interpretability:
Pérez-Rosas et al., 2019 [56]	Identify linguistic features of high quality counselors	259 recordings from web-sourced motivational interviewing sessions	Word counts; N-grams	sentiment ☺; lexicon 🅻; conversation features (turn word ratios, questions, reflections)	🅻/☺: LIWC, Opinion Finder, Wordnet Affect, CoreNLP (Python) 🔏: Google Speech-to-Text	Students: counseling quality (Motivational Interviewing Treatment Integrity annotations)	Support Vector Machine	Accuracy=.88, F1=.86/.90 (Low/High quality) *Human inter-rater reliability*: ICC=.94-.96	High quality motivational interviewing is characterized by more balanced word exchange, higher use of reflective listening, better language mirroring, and more focused conversations on behavior change.	Code: - Data: - External Validation: - Interpretability: Model ablation
Provoost et al., 2019 [101]	Evaluate a sentiment analysis algorithm	Messages from 493 participants enrolled in a text-based iCBT trial for depression	Word embeddings 🅸	sentiment ☺	☺: Sentimentics	Students: sentiment (-1 to 1); emotion (annoyance, acceptance, pensiveness, optimism, serenity)	Intra-class correlations	*Human-algorithm agreement*: ICC=.55, K=.58 *Human inter-rater reliability*: ICC=.58, Krippendorff's α=.51	Sentiment analysis may be a promising tool for psychotherapy classification, but humans' low baseline interrater reliability can pose a challenge to evaluating classifier algorithms.	Code: - Data: - External Validation: - Interpretability: -
Ramakrishna et al., 2018 [45]	Identify instances of humor	96 transcripts of motivational interviewing sessions from 6 clinical trials (26,428 total utterances and 2,251 instances of humor)	Word counts (rhymes, alliterations); Word embeddings 🅸	deep learning 𝒜ₙ; lexicon🅻	🅸: GloVE 𝒜ₙ: LSTM 🅻: Wordnet 🔏: CMU Pronunciation Dictionary	Humorous utterance (provider-patient shared "laughter" transcript tags)	LSTM Neural Network	F1=.70	A recursive neural network may outperform support vector machines baseline by having higher recall for speech classification tasks.	Code: - Data: - External Validation: - Interpretability: -
Salmi et al., 2022 [72]	Identify conversation topics of help seekers	8,589 chat conversations from a suicide prevention helpline in the Netherlands	TF-IDF 🄵; Word embeddings 🅸	deep learning 𝒜ₙ; topic modeling 💬	💬/𝒜ₙ: BERTopic 🅸/𝒜ₙ: Sentence-BERT	Relative change in topics pre/post COVID-19 lockdown	Transformer (BERTopic)	-	BERTopic captured changes in topics and was the most suitable method for analyzing chat messages compared to other topic modeling methods	Code: - Data: - External Validation: - Interpretability: -
Shapira et al., 2020a [108]	Study associations between patient linguistic behaviors and treatment outcome	873 transcripts of psychodynamic psychotherapy (58 clients treated by 52 therapists)	Word counts; part-of-speech 🧠	sentiment ☺	🧠: YAP (Go) ☺: ad hoc lexicon	Patients: Distress (Outcome Rating Scale, Outcome Questionnaire-45)	Mixed effects Model	*Distress*: ηp2=.08/.02/.05 (pre/during/pre-to-post treatment) *Human inter-rater reliability*: Fleiss' K=.95	Therapy sessions with lower frequency of first-person words and more positive emotional words are associated with lower next session patient distress.	Code: - Data: - External Validation: - Interpretability: -
Sharma et al., 2020b [62]	Classify empathic responses (utterance-level)	Message transcripts of 235,000 supportive conversations from a peer-to-peer support platform (TalkLife)	Word embeddings 🅸	deep learning 𝒜ₙ	𝒜ₙ : bi-encoder RoBERTa (domain-specific pretraining)	Crowdworkers: Empathy (EPITOME: emotional reaction, interpretation, exploration; scores 0 to 2); rationale (tokens leading to empathy annotation)	Transformer (RoBERTa)	*Empathy*: F1=.74/.67/.73 *Rationale*: token F1=.68/.68/.65 (reactions/interpretations/explorations)	On average, peer supporters do not become more empathic over time. Women are more empathic with other women than men are with other men.	Code: Available Data: Available External Validation: - Interpretability: model ablation
Sharma et al., 2021 [73]	Suggest edits to peer supporters to provide more empathic responses	Message transcripts of 3.33 million mental health-related interactions from a peer-to-peer support platform (TalkLife)	Word embeddings 🅸	deep learning 𝒜ₙ	🅸/𝒜ₙ: BERT 𝒜ₙ: PROVIDER (bi-encoder RoBERTa), DialoGPT, MIME, BART, Deep latent sequence model	Providers: 180 empathic rewritings Algorithm: changes in empathy (EPITOME rating scale)	Transformer + reinforcement learning (PROVIDER)	*Human vs algorithm rewriting*: BLEU=.14 *Rewriting*: empathy=+1.6; perplexity=-7.3; specificity=-.90; distinct-1/-.08, distinct-2/-.42; Levenshtein distance =.97	AI-assisted text rewriting of human-generated responses may be an effective approach to balancing the benefits and risks of using artificial intelligence in mental health settings.	Code: Data: External Validation: - Interpretability: Ablation, human evaluation
Schultebraucks et al., 2022 [79]	Predict PTSD and depression 1 month after emergency room admission	81 recordings from open-ended interviews with patients admitted to an emergency department following a life-threatening traumatic event	N-grams	Lexicon 🅻; sentiment ☺; deep learning 𝒜ₙ; facial coding (OpenFace); acoustic features 🔊	🅻/☺: Receptiviti, LIWC 🔏/𝒜ₙ: DeepSpeech 🔊: PRAAT	Patients: Symptoms (PTSD: PCL-5; diagnosis: score≥33. Depression: CESD; diagnosis: score ≥23)	Deep neural network	*Diagnosis*: AUC=.90/.86; F1=.83/.82; Precision=.83/.83; Recall=.84/.82; *Symptom Scores*: R2=.60/.62 (PTSD/Depression)	Combined audio, facial, and textual features analysis can accurately predict future provisional diagnostic status in minimally structured clinical contexts.	Code: - Data: - External Validation: - Interpretability: SHAP
Shidara et al., 2022 [139]	Classify clients' mental health issues	974 text message conversations held over LINE, an online counseling platform	N-grams; TF-IDF, Word embeddings 🅸	deep learning 𝒜ₙ Lexicon 🅻; topic modeling 💬	𝒜ₙ: BERT 💬: Latent Dirichlet Allocation 🅻: LIWC	Conversation topics (mental health issues vs. other issues)	Logistic regression	AUC=.70; F1=.65; precision=.61; recall=.68	TF-IDF methods can be used as a parsimonious and interpretable means of classifying conversational data.	Code: Data: External Validation: - Interpretability:
Si et al., 2019 [88]	Classify prior episode of psychosis	25 transcripts of semi-structured interview dialog with 25 participants with one prior episode of psychosis; 67,093 text messages from healthy controls	Word embeddings 🅸	deep learning 𝒜ₙ	🅸: word2vec 𝒜ₙ: CNN	Providers: Schizophrenia diagnosis	Convolutional neural network	F1=.99	A neural network can be used to distinguish between semi-structured interview transcripts and text-messages from healthy controls.	Code: Available Data: Available External Validation: - Interpretability: -
Singla et al., 2018 [80]	Evaluate utterance-level interventions	148 audio recordings and transcripts of motivational interviewing sessions	Word embeddings 🅸	deep learning 𝒜ₙ; acoustic features 🔊	𝒜ₙ: biLSTM 🔊: Praat	Motivational Interviewing Skills Code (MISC 2.1) annotations	Transformer (biLSTM with attention)	F1=.60	Prosodic features may aid in the classification of speech acts above lexical feature analysis in therapy.	Code: - Data: - External Validation: - Interpretability: Model ablation
Son et al., 2021 [98]	Predict PTSD symptoms trajectories among 9/11 responders	124 transcripts of oral history interviews from 9/11 first responders	Word counts, N-grams, bag-of-words	Lexicon 🅻; sentiment ☺; topic modeling 💬	🅻: LIWC 💬/☺: Differential Language Analysis Toolkit	Patient: Symptoms (PCL-IV)	Linear regression	β=.31/-.37/-.36 (anxious language / first-person plural / word length)	Greater use of first-person plural pronouns and longer words were linked to decreased future PTSD symptoms	Code: Data: External Validation: - Interpretability: -
Spruit et al., 2022 [85]	Provide clinical diagnoses	108 transcripts of interviews from the Dutch Verhalenbank ("Storybank") dataset	Word counts; N-grams; part-of-speech 🧠; Word embeddings 🅸	deep learning 𝒜ₙ Lexicon 🅻	🅻: LIWC 🧠: SpaCy 𝒜ₙ: RoBERTa/RoBERT, fastText	Providers: Mental health diagnoses	Random-Forest	*Diagnosis*: accuracy=.95; K=.89 *Diagnosis (multiclass)*: accuracy=.43; K=.30	In a comparison of different methods to generate linguistic representations, features from LICW and Spacy were the most accurate in classifying mental health diagnoses.	Code: Publicly available Data: Publicly available External Validation: - Interpretability: LIME
Srivastava et al., 2022 [74]	Evaluate a Counseling Summarization algorithm	212 transcripts from publicly available counseling conversations	N-grams, Word embeddings 🅸	deep learning 𝒜ₙ	🅸/𝒜ₙ: DistilBERT 𝒜ₙ: BERT, DialoGPT +BART, BiLSTM, T5	-	Transformer (ConSum)	Rouge-1=.45; Rouge-2=.16; Rouge-L=.25; BLEURT=.34; QUESTEVAL=.25	Domain-enriched transformer architectures can generate, acceptable counseling conversations summaries.	Code: Data: External Validation: - Interpretability: Ablation, human evaluation
Syzdek, 2020 [134]	Assess patient emotional sentiment	114 transcripts of psychotherapy sessions from the American Psychological Association's Therapy in Action series	Word counts	sentiment ☺	☺: syuzhet (R)	Providers: Orientation (behavioral, cognitive, humanistic, integrative, psychodynamic)	Hierarchical linear model analysis	-	Client sentiment tended to decrease over the course of the therapy session but increase between sessions.	Code: - Data: - External Validation: - Interpretability: -
Tanana et al., 2016 [118]	Evaluate human-algorithm agreement in coding interventions	341 transcripts of motivational interviewing sessions from 6 clinical trials	N-grams; part-of-speech 🧠; Word embeddings 🅸	deep learning 𝒜ₙ; discrete sentence features	🅸: GloVE 🧠: Stanford Parser 3.5.2 𝒜ₙ: Recursive Neural Network	Motivational Interviewing Skills Code (MISC 2.1) annotations	Discrete Sentence Feature Model	*Human-algorithm agreement*: Cohen's K=0- 9 / ICC=0-1 *Human inter-rater reliability*: Cohen's K=.1- 9 / ICC=.3-1 (utterance-/session-level)	Classifier algorithms had higher agreement with human coders for session-level codes, but had lower agreement at the utterance level.	Code: - Data: - External Validation: - Interpretability: -
Tanana et al., 2021 [102]	Compare sentiment classifier models	2,354 transcripts of psychotherapy sessions from Alexander Street Press corpus	Word counts; N-grams; Word embeddings 🅸	deep learning 𝒜ₙ; lexicon🅻; sentiment ☺	🅻/☺: LIWC 𝒜ₙ: BERT	Crowdworkers: Emotional valence (negative/neutral/positive)	Transformer (BERT)	*Valence*: F1=.66 *Human-algorithm agreement*: K=.48 *Human inter-rater reliability*: K=.42	A BERT-based model can exceed human performance on a sentiment rating task.	Code: Partially available Data: - External Validation: - Interpretability: -
Tasca et al., 2022 [133]	Detect talk-turn level defensive mechanisms	192 transcripts of Adult Attachment Interviews with 158 female participants (92 with binge-eating disorder)	Word embeddings 🅸	deep learning 𝒜ₙ	𝒜ₙ: RoB-RT	Raters: Defense Mechanism Rating Scale (DMRS) codes	Transformer (RoBERTa)	*All defenses*: Accuracy=.74, AUC=.82; F1=.61; PR-AUC=.60; Precision=.51, Recall=.77 *Human inter-rater reliability*: ICC=.76-.83	Transformer models could reduce coders' workload by highlighting relevant talk-turns to manually code specific defense mechanisms.	Code: Data: External Validation: - Interpretability:
Tavabi et al., 2020 [57]	Classify patient utterance-level inclinations (behavioral change)	219 audio recordings and transcripts of motivational interviewing sessions (219 clients and 12 therapists)	Word counts; Word embeddings 🅸	deep learning 𝒜ₙ; lexicon 🅻; acoustic features 🔊	𝒜ₙ: BERT 🅻: LIWC 🔊: eGeMAPS, VGGish 🔏: Speechmatics	Motivational Interviewing Skills Code (MISC 2.0) annotations	Transformer (BERT + VGGish bimodal fusion network)	F1=.72/.53/.71 (Text/Speech/Multimodal data)	Side-by-side comparison of speech-based and text-based therapy acts classification models showed text-based to be more robust.	Code: - Data: - External Validation: - Interpretability: -
Tsakalidis et al., 2021 [129]	Detect therapist-patient ruptures	873 transcripts psychodynamic psychotherapy session (Hebrew) from 68 clients	Word counts; N-grams, part-of-speech 🧠	-	🧠: YAP (Go)	Providers + Patients: Alliance ruptures (Post-Session Questionnaire)	Logistic regression	*Rupture*: F1=.62/.47 (client/therapist)	A rupture identification algorithm performed significantly better in cases where both the client and therapist agreed there had been a working alliance rupture.	Code: - Data: - External Validation: - Interpretability: -
Tseng et al., 2017 [100]	Estimate session-level negativity	588 transcripts of couples therapy sessions (134 couples); OpenSubtitles dataset	Word embeddings 🅸	deep learning 𝒜ₙ	𝒜ₙ: LSTM-RNN 🅸: word2vec (Python)	Raters: Session-level Negativity (from Couples Interaction and Social Support Interaction Rating Systems)	Transformer (LSTM-RNN + attention)	*Human-algorithm agreement*: Krippendorff's alpha=.84; MSE: 1.37 *Human inter-rater reliability*: Krippendorff's alpha=.82	Using embedding sequences from conversation models as input features achieves high inter-rater agreement with human annotators.	Code: - Data: - External Validation: Out-of-domain validation Interpretability: -
Wadden et al., 2021 [63]	Compare conversational features with and without moderation	200,000 messages from 7,000 mental health conversations conducted over two online peer-support platforms	Word counts; N-grams; Word embeddings 🅸	deep learning 𝒜ₙ; lexicon 🅻 ; sentiment ☺	🅻/☺: LIWC, Convokit 𝒜ₙ/☺: Tensorflow toxicity identifier	Moderation status	-	-	Moderation tends to enhance users' linguistic coordination, decrease profanity, increase disclosure of negative emotions, and facilitate more perspective change versus an unmoderated environment.	Code: - Data: - External Validation: - Interpretability: -
Wawer et al., 2021 [51]	Detect autism and schizophrenia diagnoses and compare model performance with psychiatrists	168 transcripts of semi-structured assessments in Polish individuals (47 with schizophrenia and 37 with autism diagnoses, 84 controls)	N-grams; bag-of-words; Word embeddings 🅸	Lexicon 🅻; sentiment ☺	☺: SenticNet; AffectNet 🅻: Linguistic Category Model 🅸: Universal sentence Encoder	Providers: Diagnosis (Schizophrenia and Autism) Raters (four psychiatrists): diagnoses based on transcript data	Neural network	*Autism*: model accuracy = .63; human accuracy=.55-.62 *Schizophrenia*: model accuracy = .81; human accuracy=.69-.76	The best diagnostic classifiers outperformed psychiatrists in diagnosing individuals based solely on transcripts.	Code: Data: External Validation: - Interpretability: -

Table 1. continued

Wei et al., 2021 [66]	Analyze patient linguistic complexity	2.6 million message therapy transcripts from 7,170 Talkspace users	Word counts; N-grams; part-of-speech	lexicon 🕮 (Dale-Chall readability score, Coleman-Liau index, concreteness, Flesch-Kincaid grade level, lexical diversity)	-	Patients: Symptoms (GAD-7)	Linear mixed modeling	-	Clients who reported more anxiety showed reduced lexical diversity, but therapists used more complex language when communicating with anxious clients.	Code: - Data: - External Validation: - Interpretability: -
Weintraub et al., 2021 [99]	Classify emotional expressions	123 transcripts of semi-structured interviews of parents of youths who had active mood symptoms and a family history of bipolar disorder	Word counts; N-grams	lexicon 🕮; sentiment ☺	🕮/☺: LIWC	Researcher: expressed emotion (Five-minute Speech Sample coding system; high/low)	Support Vector Machine	Accuracy=.75.; Sensitivity=.69; Specificity=.81	Negative emotion words were associated with higher levels of expressed emotion.	Code: - Data: - External Validation: - Interpretability: -
Wiegersma et al., 2020 [60]	Identify emotionally salient therapy hotspots	44 transcripts of brief eclectic psychotherapy sessions for PTSD	Word counts; N-grams; part-of-speech 🏷; TF-IDF	lexicon 🕮; sentiment ☺; acoustic features 🔊	🏷: NLTK 🕮/☺: LIWC 🔊: Audacity, Praat, WebMaus	Hotspot during exposure therapy (Hotspots Manual)	Support Vector Machine	Accuracy=.56; F1=.52; Precision=.54; Recall=.56	Text and speech features could distinguish between emotional hotspots and non-hotspots in a training dataset but not an external dataset.	Code: - Data: - External Validation: - Interpretability: -
Wu et al., 2021 [68]	Identify high and low counselor empathy with a model trained from general conversations	21 transcripts from motivational interviewing sessions; emotional conversations from the Persona-based Empathetic Conversation (25,000) and Empathetic Dialogues datasets (23,100)	Word embeddings 🔢	deep learning ⛬	⛬: BERT; BART	Raters: utterance-level counselor empathy (high vs. low)	Transformer (BERT; BART)	Matthews correlation coefficient=.15	Models trained to identify empathic dialogs in non-counseling domains cannot be reliably used to identify empathy in the context of motivational interviewing.	Code: Data: Publicly available External Validation: Out-of-domain validation Interpretability:
Wu et al., 2022 [122]	Forecast therapist verbal acts in motivational interviewing	133 transcripts of motivational interviewing conversations from the ANNO-MI dataset	Word embeddings 🔢	deep learning ⛬	⛬: RoBERTa	Providers: dialogue action (reflection, question, input, other)	Transformer (RoBERTa)	F1=.40	Publicly available, professionally-transcribed counseling datasets can facilitate NLP research in psychotherapeutic contexts; therapist action forecasting is highly sensitive to conversation semantics.	Code: Data: Publicly available External Validation: - Interpretability:
Xezonaki et al., 2020 [93]	Detect markers of depression and suicide risk	1,262 transcripts from therapy sessions from Alexander Street Press; 189 transcripts of semi-structured clinical interviews	Word counts; N-grams; Word embeddings 🔢	deep learning ⛬; sentiment ☺	⛬: Hierarchical RNN ☺: LIWC, AFINN, Bing Liu, MPQA, Semeval15 NRC Emotion Lexicons	Alexander Street Press: Depression diagnosis Patients: Symptoms (PHQ-8)	Transformer (Hierarchical RNN + Attention)	F1=.72/.69 (training/development set)	Individuals diagnosed with depression use more affective language than their non-depressed counterparts.	Code: - Data: - External Validation: - Interpretability: -
Xiao et al., 2015 [59]	Evaluate therapist empathy	153 audio recordings of motivational interviewing sessions; transcripts of 1,200 psychotherapy sessions	N-grams	acoustic features 🔊	🔊/🖋: Kaldi (C++)	Session-level Empathy (Motivational Interviewing Treatment Integrity Manual 3.0; high: score>4.5)	Support Vector Machine	*Empathy*: Accuracy=.82; F1=.86; Precision=.81 Recall=.92 *Human inter-rater reliability*: ICC=.60, K=.74	Empathy prediction was more reliable when using transcript data versus audio recordings alone.	Code: - Data: - External Validation: - Interpretability: -
Xiao et al., 2016 [58]	Compare automated and manual evaluation of therapist empathy	353 audio recordings of motivational interviewing sessions; 1,200 transcripts of psychotherapy sessions	Word counts; N-grams	lexicon 🕮; acoustic features 🔊	🕮: Switchboard, WSJ 🔊/🖋: Kaldi (C++)	Session-level Empathy (Motivational Interviewing Treatment Integrity Manual 3.0; high: score>4.5)	Support Vector Machine	*Empathy*: Accuracy=.81 *Human-algorithm agreement*: r =.64; MSE=1.58 *Human inter-rater reliability*: ICC=.67, agreement ratio =90.7%	Automatic session level empathy ratings were not able to reach the reliability and accuracy of human raters.	Code: - Data: - External Validation: - Interpretability: -
S. Xu et al., 2018 [81]	Detect schizophrenia symptoms	Recordings of clinical semi-structured interviews administered to 50 patients with schizophrenia and 25 patients healthy controls	N-grams, bag-of-words 🄱; Word embeddings 🔢	Lexicon 🕮, acoustic features 🔊	🕮: LIWC , Diction 🔢: Doc2Vec 🄱: PV-DBOW 🔊: Kaldi, ASpIRE	Providers: schizophrenia negative symptoms using the Negative Symptom Assessment (NSA-16)	Ensemble (SVM, GradientBoost, AdaBoost, logistic regression, random forest)	*Patients vs controls*: Accuracy=.78; AUC=.81 *NSA-16*: accuracy=.66/.68/.78; AUC=.70/.74/.82 (items 2/6/15)	Schizophrenia patients show significant language differences captured by lexical features compared to healthy controls.	Code: Data: External Validation: - Interpretability:
Z. Xu et al., 2021 [105]	Detect and score suicide risk	5682 Cantonese conversations between help-seekers and counselors	Word embeddings 🔢	deep learning ⛬; knowledge graph features	🔢: word2vec (Python) ⛬: BiLSTM + knowledge graph + conversation encoder	Providers: Suicide risk ("crisis" / "non-crisis")	Bidirectional LSTM + knowledge graph	Precision=.65/.98; Recall=.87/.94	Semantic relations between words can be used to enhance the effectiveness of suicide risk detection algorithms.	Code: - Data: - External Validation: - Interpretability: -
Y. Xu et al., 2021 [130]	Identify premature departure in online counseling	575 online counseling sessions on the Cantonese *Open Up* platform	N-grams, part-of-speech 🏷	-	🏷: jieba	Raters: premature session departure	Logic-based pattern matching algorithm	*Departure*: F1= 0.92 *Human inter-rater reliability*: Krippendorff's α=.96	Logic-based pattern matching techniques may be used in identifying premature conversation terminations in the context of online counseling.	Code: Data: External Validation: Interpretability:
Zhang et al., 2019 [65]	Study change in therapist language with experience	1,055,924 crisis center text-based conversations from 3,475 counselors	N-grams; TF-IDF 🄱	linguistic diversity	🄱: ConvoKit (Python)	Providers: Experience (# sessions: new=20; 100=tenured=120)	Logistic Regression	Accuracy=.86	Crisis center counselors demonstrate more linguistic diversity across interactions as they gain experience, but also develop structured greeting and sign-off styles.	Code: Available Data: Available by application External Validation: - Interpretability: -
Zhang et al., 2020 [125]	Automate classification of counseling strategies	1.5 million crisis center text counseling conversations (~25 messages per conversation)	N-grams; TF-IDF 🄱; Word embeddings 🔢	-	🔢🄱: ConvoKit (Python)	Patients: conversation length & helpfulness ("I feel better" / "I don't feel better")	Unsupervised learning	-	Crisis conversations that focus on exploring help-seekers' current emotional challenges are rated as more helpful than those which focus on future strategies.	Code: Available Data: Available by application External Validation: - Interpretability: -

BASIS-24 Behavior and Symptom Identification Scale, *BERT* Bidirectional Encoder Representations from Transformers, *BiLSTM* Bidirectional Long-Short Term Memory, *CBT* Cognitive Behavioral Therapy, *CESD* Center for Epidemiological Studies—Depression scale, *CNN* Convolutional Neural Network, *COPE* Coping Orientation to Problems Experienced, *COVAREP* Collaborative Voice Analysis Repository, *DAAP* Discourse Attributes Analysis Program, *DSM-5* Diagnostic & Statistical Manual of Mental Disorders, Fifth Edition, ELMo Embeddings from Language Model, *GAD-7* General Anxiety Disorder-7, *GloVE* Global Vectors for Word Representation, *GPT-2* Generative Pre-trained Transformer 2, *LIME* Local interpretable model-agnostic explanations, *LIWC* Linguistic inquiry word count, *LSTM* Long-Short Term Memory, *MAE* Mean average error, *MALLET* Machine Learning for LanguagE Toolkit, *MentalBERT* a BERT pre-trained on mental health conversations, *MSE* Mean Squared Error, *NLTK* Natural Language Tool Kit, *PANSS-8* Positive and Negative Syndrome Scale, *PCL-5* PTSD Checklist for DSM-5, *PHQ-9* Patient Health Questionnaire-9, *PSOMS* Positive States of Mind Scale, *PTGI* Posttraumatic Growth Inventory, *PTSD* Post-traumatic stress disorder, *RMSE* Root mean squared error, *RNN* Recurrent Neural Network, *RoBERTa* Robustly Optimized BERT Pre-training Approach, *ROC AUC* Receiver operating characteristic area under the curve, *SF-12* Short Form Survey, *SVM* Support Vector Machine, *SWEMWS* Short Warwick-Edinburgh Mental Well-Being Scale, *TF-IDF* Term frequency—inverse document frequency, *VADER* Valence Aware Dictionary and sEntiment Reasoner, *YAP* Yet Another (natural language) Parser. NLP language representations and model features are associated with their respective software/algorithm by the following symbols: acoustic features, bag of words and TF-IDF, deep learning, lexicon, part-of-speech, sentiment analysis, speech-to-text, topic modeling, word embeddings.

Acoustic features. Beyond the use of speech-to-text transcripts, 16 studies examined acoustic characteristics emerging from the speech of patients and providers [43, 49, 52, 54, 57–60, 75–82]. The extraction of acoustic features from recordings was done primarily using Praat and Kaldi. Engineered features of interest included voice pitch, frequency, loudness, formants quality, and speech turn statistics. Three studies merged linguistic and acoustic representations into deep multimodal architectures [57, 77, 80]. The addition of acoustic features to the analysis of linguistic features increased model accuracy, with the exception of one study where acoustics worsened model performance compared to linguistic features only [57]. Model ablation studies indicated that, when examined separately, text-based linguistic features contributed more to model accuracy than speech-based acoustics features [57, 77, 78, 80].

Clinical research categories

Three primary sources of data emerged from the examined studies: conversational data from patients ($n = 45$), another set from providers ($n = 32$), and a third set from patient-provider interactions ($n = 21$). In addition, four studies focused on improving NLP data pipelines [47, 74, 44, 83] (Fig. 3). Each of the three data sources were further divided into two subgroups according to study aims. The resulting six clinical categories are discussed further below and composed the central concepts of the integrative framework presented in the discussion.

Patient analysis ($n = 45$)

Clinical presentation ($n = 34$): These studies assessed clinical characteristics evident in transcripts grounded in diagnostic ratings obtained by providers and self-reported symptoms from patients. The premise for these applications is the effect of neuropsychiatric disorders on speech (e.g., neologisms in schizophrenia), sentiment, and content (e.g., worry in anxiety) [29] that act as language-based markers of psychopathology. *Serious Mental Illness (SMI)*. Eleven SMI applications used NLP markers to identify psychosis [51, 81, 84, 85] and bipolar [86] diagnoses (Accuracy 0.70 and 0.85), monitor symptoms [75, 87], detect psychotic episodes ($F_1 = 0.99$) [88] and psychosis onset (AUC = 0.72) [89]. Negative symptoms [87, 90] and cognitive distortions [70] for SMI were detected using linguistic features, including connectedness emerging from graph analytics [90, 91]. Associations between linguistic features and neuroimaging were also examined [91, 92]. *Depression and Anxiety*. Examination of linguistic features showed

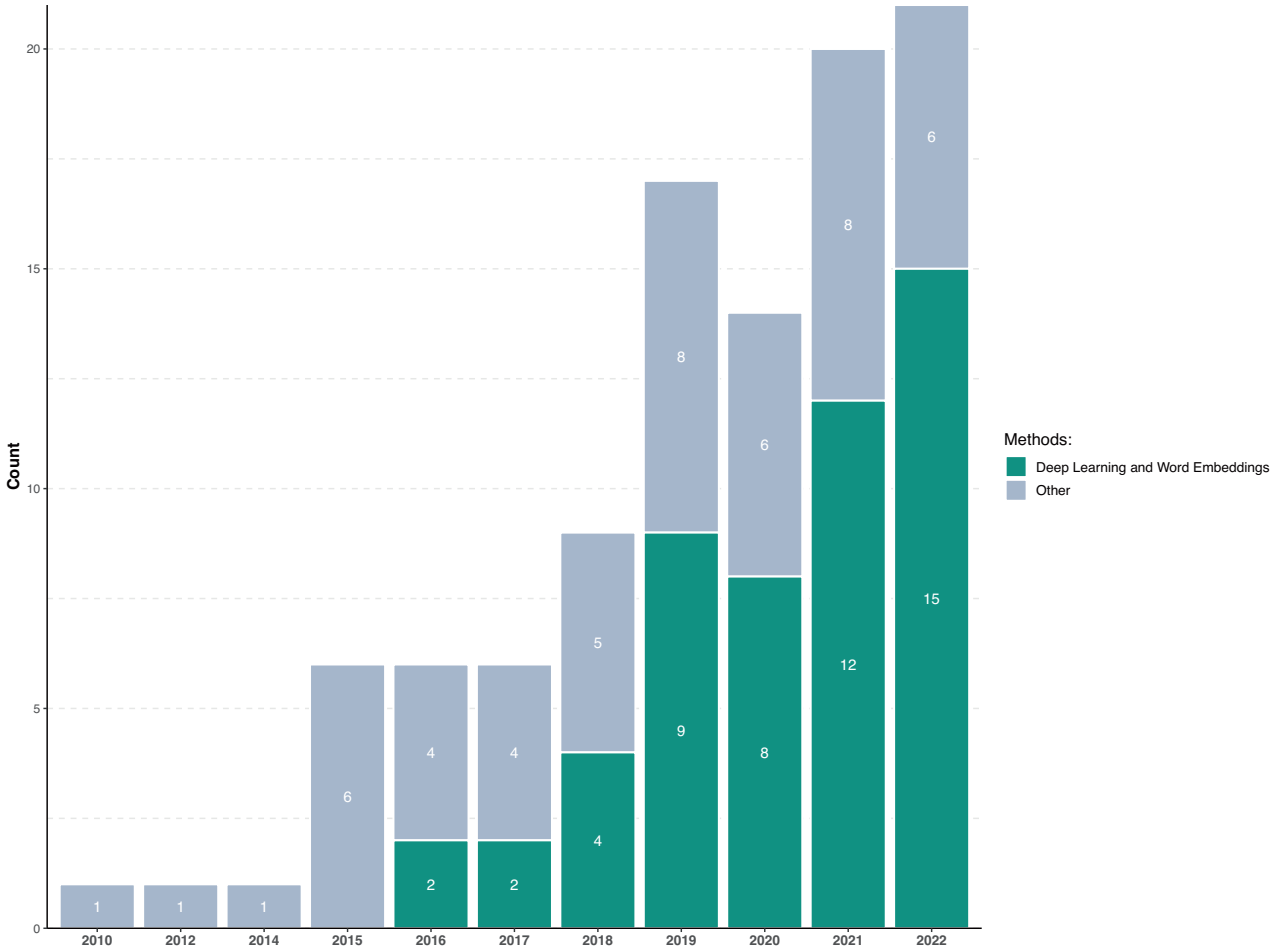


Fig. 2 Number of Articles Published per Year.

that lexical diversity [66], the use of more affective language [93] and negative emotions [93, 94] are markers of depression and anxiety [49], and can be used to predict outcomes (QIDS-16 scores, Accuracy = 0.85) [95]. Sentence embeddings [77], n-grams and topics [96] were also used to assess depression and anxiety severity. In addition, linguistic features were able to detect symptoms beyond those typically captured by diagnostic screenings [64]. *Post-Traumatic Stress Disorder* (PTSD). Three studies focused on analyzing open-ended trauma narrative to accurately identify PTSD diagnosis [97] and symptom trajectories [98]. Of note, linguistic features from narratives collected one month after life-threatening traumatic events were shown to be predictive of future PTSD (AUC = 0.90) [79]. *Affect Analysis.* Six manuscripts focused on the automatic examination of affect, a component of clinical mental status evaluations. These studies examined emotions at the session- and utterance-level [82], emotional involvement (e.g., warmth) [99], and negativity [100], and emotional distress including exposure therapy hotspots [60]. Sentiment analysis performed similarly to human raters (Cohen's $K = 0.58$) [101]. Across studies, the latest Transformer-based models were shown to capture emotional valiance [102] and associations with symptom ratings more accurately than other language features [71]. *Suicide Risk.* Another area of clinical interest was suicidality assessment ($n = 4$). While one study focused on lifetime history of suicidality [103], the majority used NLP to assess intentions of suicide or self-harm endorsed during interventions [37, 104, 105], one with sufficient accuracy to be deployed in a clinical setting (AUC = 0.83) [37].

Intervention response ($n = 11$): Eleven studies examined linguistic markers of patient response related to treatment administration [106], outcome [107, 108], patient activation [109, 110], and between-session fluctuation of symptoms [108]. One study identified linguistic markers of behavioral activation in the treatment of depression of 10,000 patients (PHQ-9 scores; R2 = 75.5%) [109]. Three studies captured within-session responses to MHI by examining patients' responses to provider interventions at *utterance-level* interactions [43, 57, 111]. Of note, Nook et al. showed that clustering a sample of 6,229 patients based on linguistic distance captured differences both in symptoms severity and treatment outcomes [112].

Provider analysis ($n = 32$)

Intervention monitoring ($n = 20$): Most provider analyses focused on monitoring treatment fidelity. These studies segmented interventions into utterance-level elements based on treatment protocols. The majority of treatment fidelity studies examined adherence to Motivational Interviewing (MI) in clinical trial and outpatient settings [52, 54–56, 76, 113–120], with Flemotomos et al. also implementing automated MI fidelity evaluation in practice [52]. Taking advantage of the generative properties of Transformer models, Cao et al. [116]. designed a system that identified MI interventions (MISC codes; $F_1 = 0.65$) and then forecasted the most likely upcoming intervention (based on the session's history), with the goal of guiding providers. Other treatment fidelity studies examined the fidelity of Cognitive Behavioral Therapy (CBT) [34, 35, 53, 48] and digital

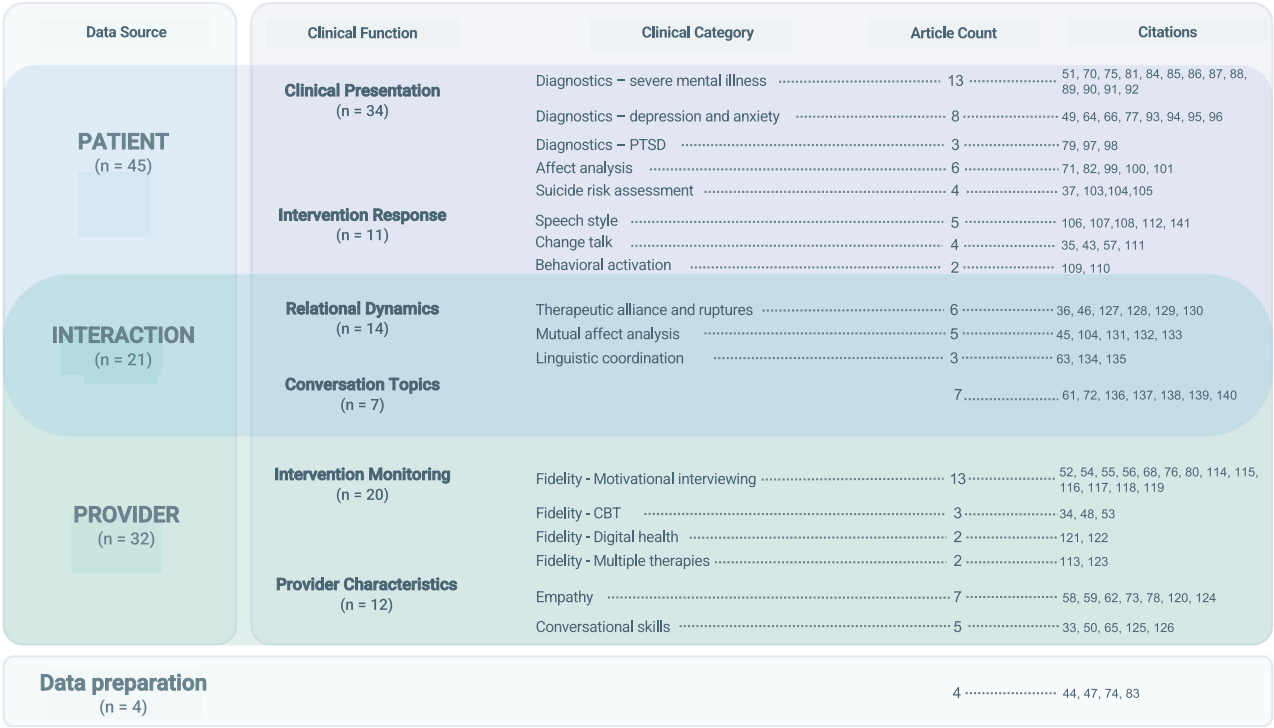


Fig. 3 Clinical research categories of reviewed manuscripts.

health [121, 122] interventions, with two examining dialogue acts distinguishing different psychotherapy approaches [113, 123]. Treatment fidelity studies primarily relied on human annotators to produce session-level behavioral codes to then train treatment segmentation models. These codes describe the structure of a session compared to the treatment's typical structure, which does not directly provide evidence for the effectiveness of specific interventions. A demonstration of the potential of directly examining treatment transcripts was shown in a study by Perez-Rosas et al. [55], where combined n-grams, lexicon, and linguistic features were as predictive of patient-rated quality as the use of behavioral codes ($F_1 = 0.87$ with/out human annotations). Language models can also be used to generate treatment fidelity labels: Ewbank and colleagues [34] automatically segmented CBT sessions into intervention components (e.g., Socratic questioning), with varying degrees of accuracy ($F_1 = 0.22$-0.94). They then showed how algorithmically identified CBT factors differentially increased the likelihood of engagement and symptom improvement (GAD-7 & PHQ-9 scores) for 17,572 patients.

Provider characteristics ($n = 12$): Empathy. Seven studies focused on the assessment of empathy, given its role in establishing treatment alliance [16]. Early models assessed session-level empathy by examining behavioral codes [58, 59, 78, 124]. Contextual language models and larger datasets examined utterance-level empathy [68], also in specific expressive forms (i.e., reactions, interpretations, and explorations) with sufficient accuracy (EPITOME codes; $F_1 = 0.74$) [62]. Similarly, Zhang and Danescu-Niculescu-Mizil [125] designed a model using 1.5 million crisis center conversations to identify whether providers responded to patients' empathetically versus advancing toward concrete resolutions. In addition to measuring empathy, Transformer-based architectures have emerged as a tool for augmenting providers' empathy, with one study using generative language models to suggest more empathic rewriting of text-based interventions [73]. *Conversational skills.* Five manuscripts examined the linguistic ability of providers [33, 50, 65, 125, 126]. One study [33] generated a model from 80,885 counseling interventions to extract therapist conversational factors, and showed how differences in content and timing of these factors predicted outcome (patient-reported helpfulness; AUC = 0.72). Importantly, conversational markers not only captured between-provider differences, but also found within-provider differences related to patients' diagnoses [50] and as they gained clinical experience over time [65].

Patient-provider interaction analysis ($n = 21$)

Relational dynamics ($n = 14$): *Therapeutic alliance and ruptures.* The study of patient-provider interactions primarily focused on analyzing therapeutic alliance, given its association with treatment outcomes [16]. Six studies sought to determine ratings of alliance strength [36, 46] or moments of rupture [127–130]. In one application, the NLP model detected patient ruptures unidentified by providers [129] and associated ruptures with decreases in emotional engagement between providers and patients [128]. *Mutual affect analysis.* Five interaction studies examined provider-therapist emotional convergence [131, 132] during the intervention, including defense mechanisms [133] and humor [45]. Tanana and colleagues examined a large corpus of therapy transcripts [102], and showed that an attention-based architecture captured therapists' and patients' valence interactions and their context more accurately ($K = 0.48$) than previous lexicon methods ($K = 0.25$ and $K = 0.31$) and human raters ($K = 0.42$). *Linguistic coordination.* Three studies focused on semantic similarity and linguistic coordination in therapeutic dyads given its association with positive outcomes [63, 134]. Researchers examined the association of linguistic coordination with affective behaviors across different treatment settings [134], and the role of linguistic synchrony in more effective interventions [135].

Conversation topics ($n = 7$): Seven studies focused on identifying conversational themes emerging from treatment interactions [61, 136–138], including identifying functioning issues [139] and capturing conversational changes following the COVID-19 pandemic [72]. Imel et al. [140]. also showed how treatment topics accurately reflect differences in therapeutic approaches (Accuracy =0.87).

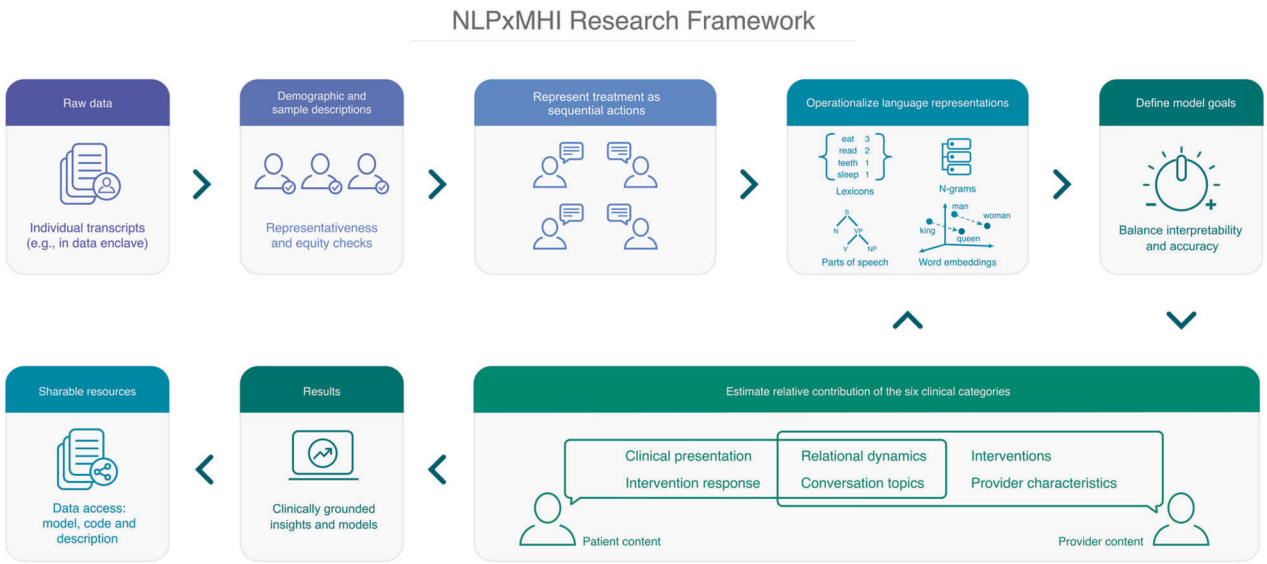


Fig. 4 NLPXMHI Framework workflow. A MHI transcripts corpus is reviewed for its representativeness. If deemed appropriate for the intended setting, the corpus is segmented into sequences, and the chosen operationalizations of language are determined based on interpretability and accuracy goals. Model features for the six distinct clinical categories are designed. If necessary, investigators may adjust their operationalizations, model goals and features. If no changes are needed, investigators report results for clinical outcomes of interest, and support results with sharable resources including code and data.

Limitations of reviewed studies

Bias towards English. Included studies overwhelmingly featured NLP analyses of English transcripts. English was the only source of conversational data for 87.3% of studies ($n = 89$). Of the remaining 13 manuscripts, three were Dutch [60, 85, 101], three were Hebrew [108, 127, 129], two were Cantonese [105, 130], two were German [44, 141], and Italian [61], Mandarin [88], and Polish [51] were each analyzed in a single study. This lack of linguistic diversity poses important questions on whether findings from the examination of English conversations can be generalized to other languages.

Limited reproducibility. While we reported algorithm performance where available, studies used different types of ground truth (e.g., psychiatrist-assessed [51] vs self-reported [122] autism) and reported different evaluation metrics (e.g., F-scores vs ROC AUC), which did not allow for meaningful direct comparisons across all studies. The examined studies were also limited in their availability of open data and open code: only a fraction of studies made their computational code ($n = 16$; 15.7%) or their data ($n = 8$; 7.8%) available. Although several studies provided graphical representations of their model architecture [52], information on algorithmic implementation, model hyper-parameters, and random seeds were typically left under-specified. Five deep learning studies mitigated this limitation by utilizing an interpretability algorithm to elucidate their models [71, 79, 85, 104, 122]. While data unavailability is understandable given concerns for patient privacy and remains a significant challenge to future work, the absence of detailed model information, shared evaluation metrics, and code is a critical obstacle to the replication and extension of findings to new clinical populations.

Population bias. A third limitation was the lack of sample diversity. Data for the studies were predominantly gathered from the US. Moreover, the majority of studies didn't offer information on patient characteristics, with only 40 studies (39.2%) reporting demographic information for their sample. In addition, while many studies examined the stability and accuracy of their findings through cross-validation and train/test split, only 4 used external validation samples [89, 107, 134] or an out-of-domain test [100]. In the absence of multiple and diverse training samples, it is not clear to what extent NLP models produced shortcut solutions based on unobserved factors from socioeconomic and cultural confounds in language [142].

DISCUSSION

In this systematic review we examined 102 applications of Natural Language Processing (NLP) for Mental Health Interventions (MHI) to evaluate their potential for informing research and practice on challenges experienced in the mental healthcare system. NLP methods are uniquely positioned to enhance language tasks with the potential to reduce provider burden, improve training and quality assurance, and more objectively operationalize MHI. To advance research in these areas, we highlight six clinical categories that emerged in the review. For the patient: (1) clinical presentation, including patient symptoms, suicide risk, and affect; and (2) intervention response, to monitor patient response during treatment. For the provider: (3) intervention monitoring, to evaluate the features of the administered treatment; and (4) provider characteristics, to study the person of the therapist and their conversational skills and traits. For patient-provider interactions: (5) relational dynamics, to evaluate alliance and relational coordination; and (6) conversational topics, to determine treatment content. In terms of language models, studies showed a shift from word count and frequency-based lexicon methods to context-sensitive deep neural networks. The growth of context-sensitive analyses appeared to follow increased prevalence of digital platforms and large corpora generated by telemedicine MHI. Acoustic features were another promising source of treatment data, although linguistic content was a richer source of information in the reviewed studies. Research in this area demonstrated progress in the areas of diagnostics, treatment specification, and the identification of contributors to outcome including the quality of the therapeutic relationship and markers of change for the patient. We propose integrating these disparate contributions into a single framework (NLPxMHI) to summarize promising avenues for increasing the utility of NLP for mental health service innovation.

NLPxMHI research framework

The goal of the NLPxMHI framework (Fig. 4) is to facilitate interdisciplinary collaboration between computational and clinical

researchers and practitioners in addressing opportunities offered by NLP. It also seeks to draw attention to a level of analysis that resides between micro-level computational research [44, 47, 74, 83, 143] and macro-level complex intervention research [144]. The first evolves too quickly to meaningfully review, and the latter pertains to concerns that extend beyond techniques of effective intervention, though both are critical to overall service provision and translational research. The process for developing and validating the NLPxMHI framework is detailed in the Supplementary Materials.

Demographic and sample descriptions for representativeness, fairness, and equity. Recent challenges in machine learning provide valuable insights into the collection and reporting of training data, highlighting the potential for harm if training sets are not well understood [145]. Since all machine learning tasks can fall prey to non-representative data [146], it is critical for NLPxMHI researchers to report demographic information for all individuals included in their models' training and evaluation phases. As noted in the Limitations of Reviewed Studies section, only 40 of the reviewed papers directly reported demographic information for the dataset used. The goal of reporting demographic information is to ensure that models are adequately powered to provide reliable estimates for all individuals represented in a population where the model is deployed [147]. While the US-based population bias for papers in the review may not be easily overcome through expansion of international population data, US domestic research can reduce hidden population bias by reporting language-relevant demographic data for the samples studied, as such data may signal to other researchers' findings influenced by dialect, geography, or a host of other factors. In addition to reporting demographic information, research designs may require over-sampling underrepresented groups until sufficient power is reached for reliable generalization to the broader population. Relatedly, and as noted in the Limitation of Reviewed Studies, English is vastly overrepresented in textual data. There does appear to be growth in non-English corpora internationally and we are hopeful that this trend will continue. Within the US, there is also some growth in services delivered to non-English speaking populations via digital platforms, which may present a domestic opportunity for addressing the English bias.

There are additional generalizability concerns for data originating from large service providers including mental health systems, training clinics, and digital health clinics. These data are likely to be increasingly important given their size and ecological validity, but challenges include overreliance on particular populations and service-specific procedures and policies. Research using these data should report the steps taken to verify that observational data from large databases exhibit trends similar to those previously reported for the same kind of data. This practice will help flag whether particular service processes have had a significant impact on results. In partnership with data providers, the source of anomalies can then be identified to either remediate the dataset or to report and address data weaknesses appropriately. Another challenge when working with data derived from service organizations is data missingness. While imputation is a common solution [148], it is critical to ensure that individuals with missing covariate data are similar to the cases used to impute their data. One suggested procedure is to calculate the standardized mean difference (SMD) between the groups with and without missing data [149]. For groups that are not well-balanced, differences should be reported in the methods to quantify selection effects, especially if cases are removed due to data missingness.

Represent treatment as sequential actions. We recommend representing treatment as sequential actions taken by providers and patients, instead of aggregating data into timeless corpora, to reduce unnecessary noise, enhancing the precision of effect estimates for intervention studies [52, 71, 109]. The reviewed studies highlight the potential benefits of embedding textual units into time-delimited sequences [52]. Longitudinal designs, while admittedly more complex, can reveal dynamics in intervention timing, change, and individual differences [150], that are otherwise lost. For example, the relationship between a specific intervention and outcome is intricate, as timing and context are important moderators of beneficial effects [113, 114]. There are no universal rules for determining how to sequence data, however the most promising avenues are: (1) turn taking; (2) the span between outcome measures; (3) sessions; or (4) clinically meaningful events arising from within, or imposed from outside the treatment.

Operationalize language representations and estimate contribution of the six clinical categories. The systematic review identified six clinical categories important to intervention research for which successful NLP applications have been developed [151–155]. While each individually reflects a significant proof-of-concept application relevant to MHI, all operate simultaneously as factors in any treatment outcome. Integrating these categories into a unified model allows investigators to estimate each category's independent contributions—a difficult task to accomplish in conventional MHI research [152]—increasing the richness of treatment recommendations. To successfully differentiate and recombine these clinical factors in an integrated model, however, each phenomenon within a clinical category must be operationalized at the level of utterances and separable from the rest. The reviewed studies have demonstrated that this level of definition is attainable for a wide range of clinical tasks [34, 50, 52, 54, 73]. Utterance-level operationalization exists for some therapy frameworks [153, 154], which can serve as exemplars to inform the specification process for other treatment approaches that have yet to tie aspects of speech to their proposed mechanisms of change and intervention. For example, it is not sufficient to hypothesize that cognitive distancing is an important factor of successful treatment. Researchers must also identify specific words in patient and provider speech that indicate the occurrence of cognitive distancing [112], and ideally just for cognitive distancing. This process is consonant with the essentials of construct and discriminant validity, with others potentially operative as well (e.g., predictive validity for markers of outcome, and convergent validity for related but complementary constructs). As research deepens in this area, we expect that there will be increasing opportunities for theory generation as certain speech elements, whether uncategorizable or derived through exploratory designs, remain outside of operationalized constructs of known theory.

Define model goals: interpretability and accuracy. Model interpretability is used to justify clinical decision-making and translate research findings into clinical policy [156]. However, there is a lack of consensus on the precise definition of interpretability and on the strategies to enhance it in the context of healthcare [157]. We suggest that enhancing interpretability through clinical review, model tuning, and generalizability is most likely to produce valid and trustworthy treatment decision rules [158] and to deliver on the personalization goals of precision medicine [159]. The reviewed studies show trade-offs between model performance and interpretability: lexicon and rule-based methods rely on predefined linguistic patterns, maximizing interpretability [33, 112], but they tend to be less accurate than deep learning models that account for more complex linguistic patterns and their context [71]. The interpretability of complex neural architectures, when deployed, should be improved at the *instance-level* to identify the words influencing model predictions. Methods include examining attention mechanisms, counterfactual explanations, and layer-wise relevance propagation. Surprisingly, only a handful of the reviewed studies implemented any of these techniques to enhance interpretability. Nevertheless, these

methods don't offer interpretation of the *overall* behavior of the model across all inputs. We expect that ongoing collaboration between clinical and computational domains will slowly fill in the gap between interpretability and accuracy through cyclical examination of model behavior and outputs. We also expect the current successes of large language models such as GPT-4 and LLaMa [160] to be further enhanced and made more clinically interpretable through training on data where relationships among the NLPxMHI categories and clinical outcomes is better understood. Meanwhile, the tradeoff between accuracy and interpretability should be determined based on research goals.

Results: clinically grounded insights and models. A sign of interpretability is the ability to take what was learned in a single study and investigate it in different contexts under different conditions. Single observational studies are insufficient on their own for generalizing findings [152, 161, 162]. Incorporating multiple research designs, such as naturalistic, experiments, and randomized trials to study a specific NLPxMHI finding [73, 163], is crucial to surface generalizable knowledge and establish its validity across multiple settings. A first step toward interpretability is to have models generate predictions from evidence-based and clinically grounded constructs. The reviewed studies showed sources of ground truth with heterogeneous levels of clinical interpretability (e.g., self-reported vs. clinician-based diagnosis) [51, 122], hindering comparative interpretation of their models. We recommend that models be trained using labels derived from standardized inter-rater reliability procedures from within the setting studied. Examples include structured diagnostic interviews, validated self-report measures, and existing treatment fidelity metrics such as MISC [67] codes. Predictions derived from such labels facilitate the interpretation of intermediary model representations and the comparison of model outputs with human understanding. Ad-hoc labels for a specific setting can be generated, as long as they are compared with existing validated clinical constructs. If complex treatment annotations are involved (e.g., empathy codes), we recommend providing training procedures and metrics evaluating the agreement between annotators (e.g., Cohen's kappa). The absence of both emerged as a trend from the reviewed studies, highlighting the importance of reporting standards for annotations. Labels can also be generated by other models [34] as part of a NLP pipeline, as long as the labeling model is trained on clinically grounded constructs and human-algorithm agreement is evaluated for all labels.

Another barrier to cross-study comparison that emerged from our review is the variation in classification and model metrics reported. Consistently reporting all evaluation metrics available can help address this barrier. Modern approaches to causal inference also highlight the importance of utilizing expert judgment to ensure models are not susceptible to collider bias, unmeasured variables, and other validity concerns [155, 164]. A comprehensive discussion of these issues exceeds the scope of this review, but constitutes an important part of research programs in NLPxMHI [165, 166].

Sharable resources: data access. The most reliable route to achieving statistical power and representativeness is more data, which is challenging in healthcare given regulations for data confidentiality and ethical considerations of patient privacy. Technical solutions to leverage low resource clinical datasets include augmentation [70], out-of-domain pre-training [68, 70], and meta-learning [119, 143]. However, findings from our review suggest that these methods do not necessarily improve performance in clinical domains [68, 70] and, thus, do not substitute the need for large corpora. As noted, data from large service providers are critical for continued NLP progress, but privacy concerns require additional oversight and planning. Only a fraction of providers have agreed to release their data to the public, even when transcripts are de-identified, because the potential for re-identification of text data is greater than for quantitative data. One exception is the Alexander Street Press corpus, which is a large MHI dataset available upon request and with the appropriate library permissions. Access to richer datasets from current service providers typically require data use agreements that stipulate the extent of data use for researchers, as well as an agreement between patients and service providers for the use of their data for research purposes. While these practices ensure patient privacy and make NLPxMHI research feasible, alternatives have been explored. One such alternative is a data enclave where researchers are securely provided access to data, rather than distributing data to researchers under a data use agreement [167]. This approach gives the data provider more control over data access and data transmission and has demonstrated some success [168].

Limitations

While this review highlights the potential of NLP for MHI and identifies promising avenues for future research, we note some limitations. Although study selection bias was limited by pre-registered review protocol and by inclusion of peer-reviewed conference papers, theoretical considerations suggest possible publication bias in the selection of the reported results toward positive findings (i.e., file-drawer effect). In particular, this might have affected the study of clinical outcomes based on classification without external validation. Moreover, included studies reported different types of model parameters and evaluation metrics even within the same category of interest. As a result, studies were not evaluated based on their quantitative performance. Future reviews and meta-analyses would be aided by more consistency in reporting model metrics. Lastly, we expect that important advancements will also come from areas outside of the mental health services domain, such as social media studies and electronic health records, which were not covered in this review. We focused on service provision research as an important area for mapping out advancements directly relevant to clinical care.

CONCLUSIONS

NLP methods hold promise for the study of mental health interventions and for addressing systemic challenges. Studies to date offer a large set of proof-of-concept applications, highlighting the importance of clinical scientists for operationalizing treatment, and of computer scientists for developing methods that can capture the sequential and context-dependent nature of interventions. The NLPxMHI framework seeks to integrate essential research design and clinical category considerations into work seeking to understand the characteristics of patients, providers, and their relationships. Large secure datasets, a common language, and fairness and equity checks will support collaboration between clinicians and computer scientists. Bridging these disciplines is critical for continued progress in the application of NLP to mental health interventions, to potentially revolutionize the way we assess and treat mental health conditions.

DATA AVAILABILITY

Data for the current study were sourced from reviewed articles referenced in this manuscript. Literature search string queries are available in the supplementary materials.

REFERENCES

1. James SL, Abate D, Abate KH, Abay SM, Abbafati C, Abbasi N, et al. Global, regional, and national incidence, prevalence, and years lived with disability for 354 diseases and injuries for 195 countries and territories, 1990–2017: a systematic analysis for the Global Burden of Disease Study 2017. Lancet. 2018;392:1789–858.

2. Figueroa JF, Phelan J, Orav EJ, Patel V, Jha AK. Association of mental health disorders with health care spending in the medicare population. JAMA Netw Open. 2020;3:e201210.
3. Miranda J, McGuire TG, Williams DR, Wang P. Mental health in the context of health disparities. AJP. 2008;165:1102–8.
4. Health TLG. Mental health matters. Lancet Glob Health. 2020;8:e1352.
5. Association AP, others. American Psychiatric Association Practice Guidelines for the treatment of psychiatric disorders: compendium 2006. American Psychiatric Pub; 2006.
6. Cuijpers P, Driessen E, Hollon SD, van Oppen P, Barth J, Andersson G. The efficacy of non-directive supportive therapy for adult depression: a meta-analysis. Clin Psychol Rev. 2012;32:280–91.
7. Firth J, Torous J, Nicholas J, Carney R, Pratap A, Rosenbaum S, et al. The efficacy of smartphone-based mental health interventions for depressive symptoms: a meta-analysis of randomized controlled trials. World Psychiatry. 2017;16:287–98.
8. DeRubeis RJ, Siegle GJ, Hollon SD. Cognitive therapy versus medication for depression: treatment outcomes and neural mechanisms. Nat Rev Neurosci. 2008;9:788–96.
9. Cunningham PJ. Beyond parity: primary care physicians' perspectives on access to mental health care. Health Aff. 2009;28:w490–w501.
10. SAMHSA. Key substance use and mental health indicators in the United States: results from the 2019 National Survey on Drug Use and Health. 2021 https://digitalcommons.fiu.edu/srhreports/health/health/32.
11. Wang PS, Aguilar-Gaxiola S, Alonso J, Angermeyer MC, Borges G, Bromet EJ, et al. Use of mental health services for anxiety, mood, and substance disorders in 17 countries in the WHO world mental health surveys. Lancet. 2007;370:841–50.
12. Insel TR. Digital phenotyping: a global tool for psychiatry. World Psychiatry. 2018;17:276–7.
13. Johnsen TJ, Friborg O. The effects of cognitive behavioral therapy as an anti-depressive treatment is falling: a meta-analysis. Psychol. Bull. 2015;141:747.
14. Kilbourne AM, Beck K, Spaeth-Rublee B, Ramanuj P, O'Brien RW, Tomoyasu N, et al. Measuring and improving the quality of mental health care: a global perspective. World Psychiatry. 2018;17:30–8.
15. Tracey TJG, Wampold BE, Lichtenberg JW, Goodyear RK. Expertise in psychotherapy: an elusive goal? Am Psychol. 2014;69:218–29.
16. Wampold BE, Imel I Zac E. The great psychotherapy debate: Models, methods, and findings. 2nd ed. Routledge/Taylor & Francis Group; 2015.
17. Douthit N, Kiv S, Dwolatzky T, Biswas S. Exposing some important barriers to health care access in the rural USA. Public Health. 2015;129:611–20.
18. Saraceno B, van Ommeren M, Batniji R, Cohen A, Gureje O, Mahoney J, et al. Barriers to improvement of mental health services in low-income and middle-income countries. Lancet. 2007;370:1164–74.
19. Heisler EJ, Bagalman E. The Mental Health Workforce: A Primer. 2018. https://ecommons.cornell.edu/handle/1813/79417 (Accessed 7 Oct 2021).
20. LeCun Y, Bengio Y, Hinton G. Deep learning. Nature. 2015;521:436–44.
21. Hosny A, Parmar C, Quackenbush J, Schwartz LH, Aerts HJWL. Artificial intelligence in radiology. Nat Rev Cancer. 2018;18:500–10.
22. Schaffter T, Buist DSM, Lee CI, Nikulin Y, Ribli D, Guan Y, et al. Evaluation of combined artificial intelligence and radiologist assessment to interpret screening mammograms. JAMA Netw Open. 2020;3:e200265.
23. Hogarty DT, Mackey DA, Hewitt AW. Current state and future prospects of artificial intelligence in ophthalmology: a review. Clin Exp Ophthalmol. 2019;47:128–39.
24. Schultebraucks K, Shalev AY, Michopoulos V, Grudzen CR, Shin S-M, Stevens JS, et al. A validated predictive algorithm of post-traumatic stress course following emergency department admission after a traumatic stressor. Nat Med. 2020;26:1084–8.
25. Dwyer DB, Falkai P, Koutsouleris N. Machine learning approaches for clinical psychology and psychiatry. Annu Rev Clin Psychol. 2018;14:91–118.
26. Jurafsky D, Martin JH. Speech and Language Processing: An introduction to speech recognition, computational linguistics and natural language processing. 1st ed. Prentice-Hall; 2008.
27. Manning C, Schutze H. Foundations of statistical natural language processing. 1st ed. MIT Press; 1999.
28. Imel ZE, Caperton DD, Tanana M, Atkins DC. Technology-enhanced human interaction in psychotherapy. J Counseling Psychol. 2017;64:385.
29. Oyebode F Sims' symptoms in the mind: an introduction to descriptive psychopathology. Elsevier Health Sciences; 2008.
30. Tausczik YR, Pennebaker JW. The Psychological Meaning of Words: LIWC and Computerized Text Analysis Methods. J Lang Soc Psychol. 2010;29:24–54.
31. Cho K, van Merriënboer B, Bahdanau D, Bengio Y. On the properties of neural machine translation: Encoder–Decoder approaches. Proceedings of SSST-8, eighth workshop on syntax, semantics and structure in statistical translation. Doha, Qatar: Association for Computational Linguistics; 2014. p. 103–11.
32. Vaswani A, Shazeer N, Parmar N, Uszkoreit J, Jones L, Gomez AN, et al. Attention is All you Need. In: Guyon I, Von Luxburg U, Bengio S, Wallach H, Fergus R, Vishwanathan S, Garnett R (eds). Advances in Neural Information Processing Systems. Curran Associates, Inc. Vol. 30, 2017. https://proceedings.neurips.cc/paper_files/paper/2017/file/3f5ee243547dee91fbd053c1c4a845aa-Paper.pdf.
33. Althoff T, Clark K, Leskovec J. Large-scale analysis of counseling conversations: an application of natural language processing to mental health. Trans. Assoc. Comput. Linguistics. 2016;4:463–76.
34. Ewbank MP, Cummins R, Tablan V, Bateup S, Catarino A, Martin AJ, et al. Quantifying the association between psychotherapy content and clinical outcomes using deep learning. JAMA Psychiatry. 2020;77:35–43.
35. Ewbank MP, Cummins R, Tablan V, Catarino A, Buchholz S, Blackwell AD. Understanding the relationship between patient language and outcomes in internet-enabled cognitive behavioural therapy: a deep learning approach to automatic coding of session transcripts. Psychother Res. 2021;31:300–12.
36. Goldberg SB, Flemotomos N, Martinez VR, Tanana MJ, Kuo PB, Pace BT, et al. Machine learning and natural language processing in psychotherapy research: alliance as example use case. J Counseling Psychol. 2020;67:438–48.
37. Bantilan N, Malgaroli M, Ray B, Hull TD. Just in time crisis response: suicide alert system for telemedicine psychotherapy settings. Psychother Res. 2021;31:289–99.
38. Miner AS, Shah N, Bullock KD, Arnow BA, Bailenson J, Hancock J. Key Considerations for Incorporating Conversational AI in Psychotherapy. Front Psychiatry. 2019;10:746.
39. Chancellor S, De Choudhury M. Methods in predictive techniques for mental health status on social media: a critical review. npj Digit Med. 2020;3:1–11.
40. Vaci N, Liu Q, Kormilitzin A, Crescenzo FD, Kurtulmus A, Harvey J, et al. Natural language processing for structuring clinical text data on depression using UK-CRIS. Evid-Based Ment Health. 2020;23:21–26.
41. Aafjes-van Doorn K, Kamsteeg C, Bate J, Aafjes M. A scoping review of machine learning in psychotherapy research. Psychother Res. 2021;31:92–116.
42. Morris RR, Kouddous K, Kshirsagar R, Schueller SM. Towards an artificially empathic conversational agent for mental health applications: System design and user perceptions. J Med Internet Res. 2018;20:e10148.
43. Aswamenakul C, Liu L, Carey KB, Woolley J, Scherer S, Borsari B. Multimodal analysis of client behavioral change coding in motivational interviewing. In: Proc. 20th ACM international conference on multimodal interaction. ACM: Boulder CO: ACM; 2018. p. 356–60.
44. Mieskes M, Stiegelmayr A. Preparing data from psychotherapy for natural language processing. Proceedings of the eleventh international conference on language resources and evaluation (LREC 2018). Miyazaki, Japan: European Language Resources Association (ELRA); 2018. https://aclanthology.org/L18-1458.
45. Ramakrishna A, Greer T, Atkins D, Narayanan S. Computational modeling of conversational humor in psychotherapy. In: Interspeech 2018. ISCA; 2018. p. 2344–48.
46. Martinez VR, Flemotomos N, Ardulov V, Somandepalli K, Goldberg SB, Imel ZE, et al. Identifying Therapist and Client Personae for Therapeutic Alliance Estimation. In: *Interspeech 2019*. 2019, ISCA, pp 1901–5.
47. Miner AS, Haque A, Fries JA, Fleming SL, Wilfley DE, Terence Wilson G, et al. Assessing the accuracy of automatic speech recognition for psychotherapy. npj Digit Med. 2020;3:82.
48. Chen Z, Flemotomos N, Singla K, Creed TA, Atkins DC, Narayanan S An automated quality evaluation framework of psychotherapy conversations with local quality estimates. Computer Speech Lang. 2022;75:101380.
49. Demiris G, Oliver DP, Washington KT, Chadwick C, Voigt JD, Brotherton S, et al. Examining spoken words and acoustic features of therapy sessions to understand family caregivers' anxiety and quality of life. Int J Med Inform. 2022;160:104716.
50. Miner AS, Fleming SL, Haque A, Fries JA, Althoff T, Wilfley DE, et al. A computational approach to measure the linguistic characteristics of psychotherapy timing, responsiveness, and consistency. npj Ment Health Res. 2022;1:19.
51. Wawer A, Chojnicka I, Okruszek L, Sarzynska-Wawer J. Single and cross-disorder detection for autism and schizophrenia. Cogn Comput. 2022;14:461–73.
52. Flemotomos N, Martinez VR, Chen Z, Singla K, Ardulov V, Peri R, et al. Am I a good therapist? Automated evaluation of psychotherapy skills using speech and language technologies. CoRR, Abs. 2021;2102:10.3758.
53. Flemotomos N, Martinez VR, Chen Z, Creed TA, Atkins DC, Narayanan S. Automated quality assessment of cognitive behavioral therapy sessions through highly contextualized language representations. PLoS ONE. 2021;16:e0258639.
54. Min DJ, Pérez-Rosas V, Mihalcea R. Evaluating automatic speech recognition quality and its impact on counselor utterance coding. In: Proceedings of the seventh workshop on computational linguistics and clinical psychology: improving access. Association for Computational Linguistics; 2021. p. 159–68.
55. Pérez-Rosas V, Sun X, Li C, Wang Y, Resnicow K, Mihalcea R. Analyzing the quality of counseling conversations: the tell-tale signs of high-quality

counseling. In: Proceedings of the eleventh international conference on language resources and evaluation (LREC 2018). 2018. European Language Resources Association (ELRA): Miyazaki, Japan https://aclanthology.org/L18-1591 (Accessed 9 Mar2022).
56. Pérez-Rosas V, Wu X, Resnicow K, Mihalcea R. What makes a good counselor? learning to distinguish between high-quality and low-quality counseling conversations. Proceedings of the 57th annual meeting of the association for computational linguistics. Florence, Italy: Association for Computational Linguistics; 2019. p. 926–35.
57. Tavabi L, Stefanov K, Zhang L, Borsari B, Woolley JD, Scherer S, et al. Multimodal automatic coding of client behavior in motivational interviewing. In: Proceedings of the 2020 international conference on multimodal interaction. ACM: Virtual Event Netherlands; 2020. p. 406–13.
58. Xiao B, Huang C, Imel ZE, Atkins DC, Georgiou P, Narayanan SS. A technology prototype system for rating therapist empathy from audio recordings in addiction counseling. PeerJ Comput Sci. 2016;2:e59.
59. Xiao B, Imel ZE, Georgiou PG, Atkins DC, Narayanan SS. 'Rate My Therapist': automated detection of empathy in drug and alcohol counseling via speech and language processing. PLoS ONE. 2015;10:e0143055.
60. Wiegersma S, Nijdam MJ, van Hessen AJ, Truong KP, Veldkamp BP, Olff M. Recognizing hotspots in brief Eclectic psychotherapy for PTSD by text and audio mining. Eur J Psychotraumatol. 2020;11:1726672.
61. Nitti M, Ciavolino E, Salvatore S, Gennaro A. Analyzing psychotherapy process as intersubjective sensemaking: an approach based on discourse analysis and neural networks. Psychother Res. 2010;20:546–63.
62. Sharma A, Miner AS, Atkins DC, Althoff T. A computational approach to understanding empathy expressed in text-based mental health support. Association for Computational Linguistics; 2020. p. 5263–76.
63. Wadden D, August T, Li Q, Althoff T. The effect of moderation on online mental health conversations. Proc Int AAAI Conf Web Soc Media. 2021;15:751–63.
64. Hull TD, Levine J, Bantilan N, Desai AN, Majumder MS. Analyzing digital evidence from a telemental health platform to assess complex psychological responses to the COVID-19 pandemic: content analysis of text messages. JMIR Form Res. 2021;5:e26190.
65. Zhang J, Filbin R, Morrison C, Weiser J, Danescu-Niculescu-Mizil C. Finding your voice: the linguistic development of mental health counselors. Proc. 57th Annual Meeting of the Association for Computational Linguistics. Florence, Italy: Association for Computational Linguistics; 2019. p. 936–47.
66. Wei J, Finn K, Templeton E, Wheatley T, Vosoughi S. Linguistic complexity loss in text-based therapy. In: Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies. 2021. Association for Computational Linguistics; 2021. p. 4450–59.
67. Moyers T, Martin T, Catley D, Harris KJ, Ahluwalia JS. Assessing the integrity of motivational interviewing interventions: Reliability of the motivational interviewing skills code. Behav Cogn Psychother. 2003;31:177–84.
68. Wu Z, Helaoui R, Reforgiato Recupero D, Riboni D Towards Low-Resource Real-Time Assessment of Empathy in Counselling. In: *Proceedings of the Seventh Workshop on Computational Linguistics and Clinical Psychology: Improving Access*. 2021. Association for Computational Linguistics: Online, pp 204–16.
69. Alsentzer E, Murphy JR, Boag W, Weng W-H, Jin D, Naumann T, et al. Publicly Available Clinical. In Proceedings of the 2nd Clinical Natural Language Processing Workshop. Association for Computational Linguistics. 2019.
70. Ding X, Lybarger K, Tauscher J, Cohen T. Improving classification of infrequent cognitive distortions: domain-specific model vs. data augmentation. Proceedings of the 2022 conference of the North American chapter of the association for computational linguistics: human language technologies: Student Research Workshop. Seattle, Washington: Association for Computational Linguistics: Hybrid; 2022. p. 68–75.
71. Burkhardt H, Pullmann M, Hull T, Aren P, Cohen T. Comparing emotion feature extraction approaches for predicting depression and anxiety. Proceedings of the eighth workshop on computational linguistics and clinical psychology. Seattle, USA: Association for Computational Linguistics; 2022. p. 105–15.
72. Salmi S, Mérelle S, Gilissen R, van der Mei R, Bhulai S. Detecting changes in help seeker conversations on a suicide prevention helpline during the COVID– 19 pandemic: in-depth analysis using encoder representations from transformers. BMC Public Health. 2022;22:530.
73. Sharma A, Lin IW, Miner AS, Atkins DC, Althoff T. Towards facilitating empathic conversations in online mental health support: a reinforcement learning approach. In Proceedings of the Web Conference 2021, 2021. pp. 194–205.
74. Srivastava A, Suresh T, Lord SP, Akhtar MS, Chakraborty T. Counseling summarization using mental health knowledge guided utterance filtering. Proceedings of the 28th ACM SIGKDD conference on knowledge discovery and data mining. Washington, DC: ACM; 2022. p. 3920–30.
75. Arevian AC, Bone D, Malandrakis N, Martinez VR, Wells KB, Miklowitz DJ, et al. Clinical state tracking in serious mental illness through computational analysis of speech. PLoS ONE. 2020;15:e0225695.
76. Chen Z, Singla K, Gibson J, Can D, Imel ZE, Atkins DC, et al. Improving the prediction of therapist behaviors in addiction counseling by exploiting class confusions. ICASSP 2019—2019 IEEE international conference on acoustics, speech and signal processing (ICASSP). Brighton, United Kingdom: IEEE; 2019. p. 6605–9.
77. Mao K, Zhang W, Wang DB, Li A, Jiao R, Zhu Y, et al. Prediction of depression severity based on the prosodic and semantic features with bidirectional LSTM and Time Distributed CNN. IEEE Trans Affective Comput. 2022;1.
78. Pérez-Rosas V, Mihalcea R, Resnicow K, Singh S, An L. Understanding and predicting empathic behavior in counseling theraaspy. Proceedings of the 55th annual meeting of the association for computational linguistics (Volume 1: Long Papers). Vancouver, Canada: Association for Computational Linguistics; 2017. p. 1426–35.
79. Schultebraucks K, Yadav V, Shalev AY, Bonanno GA, Galatzer-Levy IR. Deep learning-based classification of posttraumatic stress disorder and depression following trauma utilizing visual and auditory markers of arousal and mood. Psychol Med. 2022;52:957–67.
80. Singla K, Chen Z, Flemotomos N, Gibson J, Can D, Atkins D, et al. Using prosodic and lexical information for learning utterance-level behaviors in psychotherapy. In: Interspeech 2018. ISCA; 2018. p. 3413-7.
81. Xu S, Yang Z, Chakraborty D, Tahir Y, Maszczyk T, Chua VYH, et al. Automatic verbal analysis of interviews with schizophrenic patients. 2018 IEEE 23rd international conference on digital signal processing (DSP). Shanghai, China: IEEE; 2018. p. 1–5.
82. Crangle CE, Wang R, Guimaraes MP, Nguyen MU, Nguyen DT, Suppes P. Machine learning for the recognition of emotion in the speech of couples in psychotherapy using the Stanford Suppes Brain Lab Psychotherapy Dataset. *CoRR* 2019. Preprint at http://arxiv.org/abs/1901.04110.
83. Carcone AI, Hasan M, Alexander GL, Dong M, Eggly S, Brogan Hartlieb K, et al. Developing machine learning models for behavioral coding. J Pediatr Psychol. 2019;44:289–99.
84. Just SA, Haegert E, Kořánová N, Bröcker A-L, Nenchev I, Funcke J, et al. Modeling Incoherent Discourse in Non-Affective Psychosis. Front Psychiatry. 2020;11:846.
85. Spruit M, Verkleij S, de Schepper K, Scheepers F. Exploring language markers of mental health in psychiatric stories. Appl Sci. 2022;12:2179.
86. Carrillo F, Mota N, Copelli M, Ribeiro S, Sigman M, Cecchi G, et al. Emotional Intensity analysis in Bipolar subjects. Preprint at http://arxiv.org/abs/1606.02231.
87. Alonso-Sánchez MF, Ford SD, MacKinley M, Silva A, Limongi R, Palaniyappan L. Progressive changes in descriptive discourse in First Episode Schizophrenia: a longitudinal computational semantics study. Schizophrenia. 2022;8:36.
88. Si D, Cheng SC, Xing R, Liu C, Wu HY. Scaling up prediction of psychosis by natural language processing. 2019 IEEE 31st International Conference on Tools with Artificial Intelligence (ICTAI), Portland, OR, USA, 2019. pp. 339–47, https://doi.org/10.1109/ICTAI.2019.00055.
89. Corcoran CM, Carrillo F, Fernández-Slezak D, Bedi G, Klim C, Javitt DC, et al. Prediction of psychosis across protocols and risk cohorts using automated language analysis. World Psychiatry. 2018;17:67–75.
90. Mota NB, Ribeiro M, Malcorra BLC, Atídio JP, Haguiara B, Gadelha A. Happy thoughts: What computational assessment of connectedness and emotional words can inform about early stages of psychosis. Schizophrenia Res. 2022;259:38–47.
91. Palaniyappan L, Mota NB, Oowise S, Balain V, Copelli M, Ribeiro S, et al. Speech structure links the neural and socio-behavioural correlates of psychotic disorders. Prog Neuro-Psychopharmacol Biol Psychiatry. 2019;88:112–20.
92. Alonso-Sánchez MF, Limongi R, Gati J, Palaniyappan L. Language network self-inhibition and semantic similarity in first-episode schizophrenia: A computational-linguistic and effective connectivity approach. Schizophrenia Res. 2022;259:97–103.
93. Xezonaki D, Paraskevopoulos G, Potamianos A, Narayanan S. Affective conditioning on hierarchical networks applied to depression detection from transcribed clinical interviews. Preprint at http://arxiv.org/abs/2006.08336.
94. Dirkse D, Hadjistavropoulos HD, Hesser H, Barak A. Linguistic analysis of communication in therapist-assisted internet-delivered cognitive behavior therapy for generalized anxiety disorder. Cogn Behav Ther. 2015;44:21–32.
95. Carrillo F, Sigman M, Fernández Slezak D, Ashton P, Fitzgerald L, Stroud J, et al. Natural speech algorithm applied to baseline interview data can predict which patients will respond to psilocybin for treatment-resistant depression. J Affect Disord. 2018;230:84–6.
96. Howes C, Purver M, McCabe R. Linguistic indicators of severity and progress in online text-based therapy for depression. Proceedings of the workshop on computational linguistics and clinical psychology: from linguistic signal to

clinical reality. Baltimore, Maryland, USA: Association for Computational Linguistics; 2014. p. 7–16.
97. He Q, Veldkamp BP, Glas CAW, de Vries T. Automated assessment of patients' self-narratives for posttraumatic stress disorder screening using natural language processing and text mining. Assessment. 2017;24:157–72.
98. Son Y, Clouston SAP, Kotov R, Eichstaedt JC, Bromet EJ, Luft BJ, et al. World Trade Center responders in their own words: predicting PTSD symptom trajectories with AI-based language analyses of interviews. Psychol Med. 2021;53:918–26.
99. Weintraub MJ, Posta F, Arevian AC, Miklowitz DJ. Using machine learning analyses of speech to classify levels of expressed emotion in parents of youth with mood disorders. J Psychiatr Res. 2021;136:39–46.
100. Tseng S-Y, Baucom B, Georgiou P. Approaching human performance in behavior estimation in couples therapy using deep sentence embeddings. In: Interspeech 2017. ISCA; 2017. p 3291–95.
101. Provoost S, Ruwaard J, van Breda W, Riper H, Bosse T. Validating automated sentiment analysis of online cognitive behavioral therapy patient texts: an exploratory study. Front Psychol. 2019;10:1065.
102. Tanana MJ, Soma CS, Kuo PB, Bertagnolli NM, Dembe A, Pace BT, et al. How do you feel? Using natural language processing to automatically rate emotion in psychotherapy. *Behav Res*. 2021. https://doi.org/10.3758/s13428-020-01531-z.
103. Glauser T, Santel D, DelBello M, Faist R, Toon T, Clark P, et al. Identifying epilepsy psychiatric comorbidities with machine learning. Acta Neurol Scand. 2020;141:388–96.
104. Kshirsagar R, Morris R, Bowman S. Detecting and explaining crisis. Proceedings of the fourth workshop on computational linguistics and clinical psychology—from linguistic signal to clinical reality. Vancouver, BC: Association for Computational Linguistics; 2017. p. 66–73.
105. Xu Z, Xu Y, Cheung F, Cheng M, Lung D, Law YW, et al. Detecting suicide risk using knowledge-aware natural language processing and counseling service data. Soc Sci Med. 2021;283:114176.
106. Baggott MJ, Kirkpatrick MG, Bedi G, de Wit H. Intimate insight: MDMA changes how people talk about significant others. J Psychopharmacol. 2015;29:669–77.
107. Norman KP, Govindjee A, Norman SR, Godoy M, Cerrone KL, Kieschnick DW, et al. Natural language processing tools for assessing progress and outcome of two veteran populations: cohort study from a novel online intervention for posttraumatic growth. JMIR Form Res. 2020;4:e17424.
108. Shapira N, Lazarus G, Goldberg Y, Gilboa-Schechtman E, Tuval-Mashiach R, Juravski D, et al. Using computerized text analysis to examine associations between linguistic features and clients' distress during psychotherapy. J Counseling Psychol. 2021;68:77–87.
109. Burkhardt HA, Alexopoulos GS, Pullmann MD, Hull TD, Areán PA, Cohen T. Behavioral activation and depression symptomatology: longitudinal assessment of linguistic indicators in text-based therapy sessions. J Med Internet Res. 2021;23:e28244.
110. Malins S, Figueredo G, Jilani T, Long Y, Andrews J, Rawsthorne M, et al. Developing an automated assessment of in-session patient activation for psychological therapy: codevelopment approach. JMIR Med Inf. 2022;10:e38168.
111. SPark S, Kim D, Oh A. Conversation model fine-tuning for classifying client utterances in counseling dialogues. 2019. IEEE 31st International Conference on Tools with Artificial Intelligence (ICTAI), IEEE. p. 339–47.
112. Nook EC, Hull TD, Nock MK, Somerville LH. Linguistic measures of psychological distance track symptom levels and treatment outcomes in a large set of psychotherapy transcripts. Proc Natl Acad Sci USA. 2022;119:e2114737119.
113. Lee F-T, Hull D, Levine J, Ray B, McKeown K. Identifying therapist conversational actions across diverse psychotherapeutic approaches. Proceedings of the sixth workshop on computational linguistics and clinical psychology. Minneapolis, Minnesota: Association for Computational Linguistics; 2019. p. 12–23.
114. Atkins DC, Steyvers M, Imel ZE, Smyth P. Scaling up the evaluation of psychotherapy: evaluating motivational interviewing fidelity via statistical text classification. Implement Sci. 2014;9:49.
115. Can D, Marín RA, Georgiou PG, Imel ZE, Atkins DC, Narayanan SS. "It sounds like…": A natural language processing approach to detecting counselor reflections in motivational interviewing. J Counseling Psychol. 2016;63:343–50.
116. Cao J, Tanana M, Imel ZE, Poitras E, Atkins DC, Srikumar V. Observing dialogue in therapy: categorizing and forecasting behavioral codes. Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics; 2019. p. 5599–611.
117. Pérez-Rosas V, Mihalcea R, Resnicow K, Singh S, Ann L, Goggin KJ, et al. Predicting counselor behaviors in motivational interviewing encounters. Proceedings of the 15th conference of the european chapter of the association for computational linguistics: volume 1, long papers. Valencia, Spain: Association for Computational Linguistics; 2017. p. 1128–37.
118. Tanana M, Hallgren KA, Imel ZE, Atkins DC, Srikumar V. A comparison of natural language processing methods for automated coding of motivational interviewing. J Subst Abus Treat. 2016;65:43–50.
119. Chen Z, Flemotomos N, Imel ZE, Atkins DC, Narayanan S. Leveraging open data and task augmentation to automated behavioral coding of psychotherapy conversations in low-resource scenarios. 2022. In Findings of the Association for Computational Linguistics: EMNLP 2022. 2022. p. 5787–95.
120. Wu Z, Helaoui R, Reforgiato Recupero D, Riboni D. Towards automated counselling decision-making: remarks on therapist action forecasting on the AnnoMI dataset. In: Interspeech 2022. ISCA; 2022. p. 1906–10.
121. Hudon A, Beaudoin M, Phraxayavong K, Dellazizzo L, Potvin S, Dumais A. Implementation of a machine learning algorithm for automated thematic annotations in avatar: A linear support vector classifier approach. Health Inform J. 2022;28:146045822211424.
122. Liu Z, Peach RL, Lawrance EL, Noble A, Ungless MA, Barahona M. Listening to mental health crisis needs at scale: using natural language processing to understand and evaluate a mental health crisis text messaging service. Front Digit Health. 2021;3:779091.
123. Mehta M, Caperton D, Axford K, Weitzman L, Atkins D, Srikumar V, et al. Psychotherapy is not one thing: simultaneous modeling of different therapeutic approaches. Proceedings of the eighth workshop on computational linguistics and clinical psychology. Seattle, USA: Association for Computational Linguistics; 2022. p. 47–58.
124. Gibson J, Can D, Xiao B, Imel ZE, Atkins DC, Georgiou P, et al. A deep learning approach to modeling empathy in addiction counseling; 2016. p. 1447-51.
125. Zhang J, Danescu-Niculescu-Mizil C. Balancing objectives in counseling conversations: advancing forwards or looking backwards. In: Proceedings of the 58th annual meeting of the association for computational linguistics. 2020. Association for Computational Linguistics 2020. p. 5276–89.
126. Goldberg SB, Tanana M, Imel ZE, Atkins DC, Hill CE, Anderson T. Can a computer detect interpersonal skills? Using machine learning to scale up the Facilitative Interpersonal Skills task. Psychother Res. 2021;31:281–8.
127. Atzil-Slonim D, Juravski D, Bar-Kalifa E, Gilboa-Schechtman E, Tuval-Mashiach R, Shapira N, et al. Using topic models to identify clients' functioning levels and alliance ruptures in psychotherapy. Psychotherapy. 2021. https://doi.org/10.1037/pst0000362.
128. Christian C, Barzilai E, Nyman J, Negri A. Assessing key linguistic dimensions of ruptures in the therapeutic alliance. J Psycholinguist Res. 2021;50:143–53.
129. Tsakalidis A, Atzil-Slonim D, Polakovski A, Shapira N, Tuval-Mashiach R, Liakata M. Automatic identification of ruptures in transcribed psychotherapy sessions. In: Proceedings of the seventh workshop on computational linguistics and clinical psychology: improving access. Association for Computational Linguistics; 2021. p. 122–8.
130. Xu Y, Chan CS, Tsang C, Cheung F, Chan E, Fung J, et al. Detecting premature departure in online text-based counseling using logic-based pattern matching. Internet Interventions. 2021;26:100486.
131. Park J, Jindal A, Kuo P, Tanana M, Lafata JE, Tai-Seale M, et al. Automated rating of patient and physician emotion in primary care visits. Patient Educ Couns. 2021;104:2098–2105.
132. Syzdek BM. Client and therapist psychotherapy sentiment interaction throughout therapy. Psychol Stud. 2020;65:520–30.
133. Tasca AN, Carlucci S, Wiley JC, Holden M, El-Roby A, Tasca GA. Detecting defense mechanisms from Adult Attachment Interview (AAI) transcripts using machine learning. Psychotherapy Res. 2022;33:757–67.
134. Nasir M, Chakravarthula SN, Baucom B, Atkins DC, Georgiou P, Narayanan S. Modeling interpersonal linguistic coordination in conversations using word mover's distance. In Interspeech. 2019, vol. 2019. pp. 1423–27.
135. Doré BP, Morris RR. Linguistic synchrony predicts the immediate and lasting impact of text-based emotional support. Psychol Sci. 2018;29:1716–23.
136. Atkins DC, Rubin TN, Steyvers M, Doeden MA, Baucom BR, Christensen A. Topic models: a novel method for modeling couple and family text data. J Fam Psychol. 2012;26:816.
137. Chaoua I, Recupero DR, Consoli S, Härmä A, Helaoui R. Detecting and tracking ongoing topics in psychotherapeutic conversations. In: AIH@IJCAI. 2018. p. 97–108.
138. Gaut G, Steyvers M, Imel ZE, Atkins DC, Smyth P. Content coding of psychotherapy transcripts using labeled topic models. IEEE J Biomed Health Informatics. 2017;21:476–87.
139. Shidara K, Tanaka H, Asada R, Higashiyama K, Adachi H, Kanayama D, et al. Linguistic features of clients and counselors for early detection of mental health issues in online text-based counseling. 2022 44th annual international conference of the IEEE engineering in medicine & biology society (EMBC). Glasgow, Scotland, United Kingdom: IEEE; 2022. p. 2668–71.
140. Imel ZE, Steyvers M, Atkins DC. Computational psychotherapy research: Scaling up the evaluation of patient–provider interactions. Psychotherapy. 2015;52:19–30.
141. Hoogendoorn M, Berger T, Schulz A, Stolz T, Szolovits P. Predicting social anxiety treatment outcome based on therapeutic email conversations. IEEE J Biomed Health Inform. 2017;21:1449–59.

142. Pace A, Luo R, Hirsh-Pasek K, Golinkoff RM. Identifying pathways between socioeconomic status and language development. Annu Rev Linguist. 2017;3:285–308.
143. Zhang XS, Tang F, Dodge HH, Zhou J, Wang F. Metapred: Meta-learning for clinical risk prediction with limited patient electronic health records. In Proceedings of the 25th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining (KDD '19). 2019. Association for Computing Machinery, New York, NY, USA, p. 2487–95. https://doi.org/10.1145/3292500.33307792019.
144. Skivington K, Matthews L, Simpson SA, Craig P, Baird J, Blazeby JM, et al. A new framework for developing and evaluating complex interventions: update of Medical Research Council guidance. BMJ. 2021;374.
145. O'Neil C Weapons of math destruction. Crown; 2016.
146. Gebru T, Morgenstern J, Vecchione B, Vaughan JW, Wallach H, Iii HD, et al. Datasheets for datasets. Commun ACM. 2021;64:86–92.
147. Hernandez-Boussard T, Bozkurt S, Ioannidis JP, Shah NH. MINIMAR (MINimum Information for Medical AI Reporting): developing reporting standards for artificial intelligence in health care. J Am Med Inform Assoc. 2020;27:2011–5.
148. Sterne JAC, White IR, Carlin JB, Spratt M, Royston P, Kenward MG. et al. Multiple imputation for missing data in epidemiological and clinical research: potential and pitfalls. BMJ. 2009;338:b2393.
149. Hicks JL, Althoff T, Sosic R, Kuhar P, Bostjancic B, King AC, et al. Best practices for analyzing large-scale health data from wearables and smartphone apps. npj Digit Med. 2019;2:45.
150. Mullet E, Chasseigne G. Assessing information integration processes: a comparison of findings obtained with between-subjects designs versus within-subjects designs. Qual Quant. 2018;52:1977–88.
151. Kazdin AE. Understanding how and why psychotherapy leads to change. Psychother Res. 2009;19:418–28.
152. Cuijpers P, Reijnders M, Huibers MJH. The role of common factors in psychotherapy outcomes. Annu Rev Clin Psychol. 2019;15:207–31.
153. Moyers TB, Miller WR, Hendrickson SML. How does motivational interviewing work? Therapist interpersonal skill predicts client involvement within motivational interviewing sessions. J Consulting Clin Psychol. 2005;73:590–8.
154. Pascual-Leone A. How clients "change emotion with emotion": A programme of research on emotional processing. Psychother Res. 2018;28:165–82.
155. Ohlsson H, Kendler KS. Applying causal inference methods in psychiatric epidemiology: a review. JAMA Psychiatry. 2020;77:637.
156. Tonekaboni S, Joshi S, McCradden MD, Goldenberg A. What clinicians want: contextualizing explainable machine learning for clinical end use. In: Machine learning for healthcare conference. PMLR; 2019. p. 359-80.
157. Widmer G, Kubat M. Learning in the presence of concept drift and hidden contexts. Mach Learn. 1996;23:69–101.
158. Lutz W, Stulz N, Martinovich Z, Leon S, Saunders SM. Methodological background of decision rules and feedback tools for outcomes management in psychotherapy. Psychother Res. 2009;19:502–10.
159. Delgadillo J, Lutz W. A development pathway towards precision mental health care. JAMA Psychiatry. 2020;77:889.
160. Touvron H, Lavril T, Izacard G, Martinet X, Lachaux M-A, Lacroix T, et al. Llama: open and efficient foundation language models. Preprint at https://arxiv.org/abs/2302.13971.
161. Cristea IA, Vecchi T, Cuijpers P. Top-down and bottom-up pathways to developing psychological interventions. JAMA Psychiatry. 2021;78:593–4.
162. Chorpita BF, Daleiden EL, Weisz JR. Identifying and selecting the common elements of evidence based interventions: a distillation and matching model. Ment Health Serv Res. 2005;7:5–20.
163. Sharma A, Lin IW, Miner AS, Atkins DC, Althoff T. Human–AI collaboration enables more empathic conversations in text-based peer-to-peer mental health support. Nat Mach Intell. 2023;5:46–57. https://doi.org/10.1038/s42256-022-00593-2.
164. Cunningham S. Causal Inference: The Mixtape. Yale University Press; 2021. https://mixtape.scunning.com/.
165. Weld G, West P, Glenski M, Arbour D, Rossi RA, Althoff T. Adjusting for confounders with text: challenges and an empirical evaluation framework for causal inference. In: Proceedings of the international AAAI conference on web and social media. 2022. p. 1109–20.
166. Zhang J, Mullainathan S, Danescu-Niculescu-Mizil C. Quantifying the causal effects of conversational tendencies. Proc ACM Hum-Computer Interact. 2020;4:1–24.
167. Lane J, Schur C. Balancing access to health data and privacy: a review of the issues and approaches for the future: balancing access to health data and privacy. Health Serv Res. 2010;45:1456–67.
168. MacAvaney S, Mittu A, Coppersmith G, Leintz J, Resnik P. Community-level research on suicidality prediction in a secure environment: overview of the CLPsych 2021 shared task. In: Proceedings of the seventh workshop on computational linguistics and clinical psychology: improving access. Association for Computational Linguistics; 2021. p. 70–80.

ACKNOWLEDGEMENTS

The authors thank Patricia Areán, Kyunghyun Cho, Trevor Cohen, Adam S. Miner, Eric C. Nook, and Naomi M. Simon for their contributions as expert panelists, guiding the development of the NLPxMHI framework with their incisive and constructive feedback. Their extensive combined expertise in clinical, NLP, and translational research helped refine many of the concepts presented in the NLPxMHI framework.

AUTHOR CONTRIBUTIONS

MM, TDH, JZ, and TA jointly contributed to Conceptualization, Writing—Original draft preparation, Reviewing and Editing. MM, TDH, and JZ contributed to Data curation, Analysis, and Visualization. All authors have approved the final manuscript.

FUNDING

MM's research was supported by the National Institutes of Health (NIH) and National Center for Advancing Translational Sciences (NCATS) and through grants # 1K23MH134068-01 and 2KL2TR001446-06A1, Talkspace, and by the American Foundation for Suicide Prevention through grant PRG-0-104-19. TDH's research was supported by National Institutes of Health Awards # R44MH124334 and R01MH125179-01. TA's research was supported by NIH grant R01MH125179, NSF grant IIS-1901386, NSF grant CNS-2025022, Bill & Melinda Gates Foundation (INV-004841), and the Office of Naval Research (#N00014-21-1-2154). The content is solely the responsibility of the authors and does not necessarily represent the official views of the NIH.

COMPETING INTERESTS

TDH is an employee and JZ is a contractor of the platform that provided data for 6 out of 102 studies examined in this systematic review. Talkspace had no role in the analysis, interpretation of the data, or decision to submit the manuscript for publication.

ADDITIONAL INFORMATION

Supplementary information The online version contains supplementary material available at https://doi.org/10.1038/s41398-023-02592-2.

Correspondence and requests for materials should be addressed to Matteo Malgaroli.