cpsy

an open access journal

Citation: Mukherjee, S. S., Yu, J., Won, Y., McClay, M. J., Wang, L., Rush, A. J., & Sarkar, J. (2020). Natural language processing-based quantification of the mental state of psychiatric patients. *Computational Psychiatry*, *4*, 76–106. https://doi.org/10.1162/cpsy_a_00030

DOI: https://doi.org/10.1162/cpsy_a_00030

Supporting Information: http://doi.org/10.1162/cpsy_a_00030

Received: 11 October 2019
Accepted: 6 November 2020

Competing Interests: The authors declare no conflict of interest.

Corresponding Author:
Sankha S. Mukherjee
sankha.mukherjee@gmail.com

The MIT Press

RESEARCH

Natural Language Processing-Based Quantification of the Mental State of Psychiatric Patients

Sankha S. Mukherjee[1], Jiawei Yu[1], Yida Won[1], Mary J. McClay[1], Lu Wang[1], A. John Rush[2,3,4], and Joydeep Sarkar[1]

[1]Holmusk Inc., Singapore

[2]National University of Singapore, Singapore

[3]Duke University School of Medicine, Durham, North Carolina, USA

[4]Texas Tech University Health Sciences Center, Permian Basin, Texas, USA

Keywords: natural language processing, psychiatry, mental health

ABSTRACT

Psychiatric practice routinely uses semistructured and/or unstructured free text to record the behavior and mental state of patients. Many of these data are unstructured, lack standardization, and are difficult to use for analysis. Thus, it is difficult to quantitatively analyze a patient's illness trajectory over time and his or her responsiveness to treatment, and it is also difficult to compare different patients quantitatively. In this article, experts in the field of psychiatry, along with machine learning models, have collaboratively transformed patient data available in status assessments generated by physicians into binary vector representations. Data from patients with mental health disorders collected within a real-world clinical setting from one of the largest behavioral electronic health record (EHR) systems in the United States have been used for generating these representations. The binary vector representation of these health records is shown to be useful in various clinical tasks, such as disease phenotyping, characterizing the suicidality of patients, and inferring diagnoses. To summarize, this approach can transform semistructured free-text summaries of patients' status assessments into a structured, quantifiable format, which enriches the data that reside within EHR systems. This allows for effective intra- and interpatient quantifications and comparisons, which are much needed in the field of mental health. With the aid of these binary representations, patients' mental states can be systematically tracked over time, as can their responses to medications at the individual and population levels.

INTRODUCTION

Mental disorders are among the most challenging illnesses to treat due to the paucity of biomarkers that identify and quantify the severity of disease, as is standard in other therapeutic areas. Diagnosis in mental health is based upon observations as specified by systems such as the *Diagnostic and Statistical Manual of Mental Disorders*, fifth edition (*DSM*-5; American Psychiatric Association, 2013) and the International Classification of Diseases (ICD; World Health Organization, 1993). Categorization within such systems predates modern neuroscience (Marshall, 2020), hence the contextualization of advances in neuroscience within such classification systems for disease and severity is rather difficult. Furthermore, unlike in other areas

of health care, external experiences and social constructs typically have a significant influence on the prognosis and efficacy of the treatment of mental health disorders.

There exists a significant gap between advances in neuroscience and their ultimate translation into treatment decisions. This has led to a call for a more rigorous, evidence-based system, called the Research Domain Criterion (RDoC; Cuthbert, 2014; Cuthbert & Insel, 2013; Cuthbert & Kozak, 2013), that will attempt to classify disorders based upon a combination of domains/constructs of behavior and mental capacity and units of analysis such as genes, molecules, cells, brain circuits, physiology, behavior, self-reports, and other paradigms. This is a radical step away from a purely behavior-based system to a more evidence-based system. Computational psychiatry has been proposed to be a bridge that can further accelerate the translation of neuroscientific research into clinical practice (Huys, Maia, & Frank, 2016). Greater adoption of RDoC and similar systems will enable large, multiscale models to be used in conjunction with traditional therapeutics to improve the treatment of mental health disorders.

The MindLinc Global Database (MGD) is a repository of longitudinal patient records with behavioral health disorders from more than 25 hospitals that use the MindLinc electronic health record (EHR; Beyer, Kuchibhatla, Gersing, & Krishnan, 2005). It is one of the largest longitudinal behavioral health databases in the world to capture clinical data in a real-world setting. MGD comprises records of more than 500,000 patents, with more than 14 million certified visits. As of early 2016, there were more than 42 million records of diagnostic data, 68 million records of prescription data, 22 million records of information pertaining to substance abuse, and 330 million records of mental status examinations.[1]

The Mental Status Examination (MSE; Koita, Riggio, & Jagoda, 2010; Martin, 1990) is an important clinical assessment in psychiatric practice. It is an essential tool that is used in part for mental health diagnostics. Clinicians assess a patient's mental state using various methods. Primarily, the patient's electronic medical record (EMR) history and the clinician's expert opinion on the patient's in-clinic behavior during the visit formulate the basis for the diagnosis, but it may be enhanced by conversations between the clinician and the patient and elements of the formal MSE results. For example, the MSE may assess a patient's memory, recall, and cognitive ability by asking the patient to serially count down from 100 in steps of 7 or the ability of a patient to formulate abstract thoughts by assessing the patient's ability to interpret proverbs. The result is a clinician note that is a mix of symptoms (what the patient complains about), signs (what the clinician observes about the patient), the patient's physical appearance, and responses to formal assessments of specific brain functions that are used in the clinic in a nonsystematic way but are borrowed from the full formal MSE. We hypothesize that clinical observations, reports by the patient, and the assessment of brain function, as achieved by using elements of the formal MSE, together may reflect the mental health and/or cognitive status of a patient. We have developed a way to systematically extract these data from each clinic visit, allowing the integration of information across visits and patients over time. For ease of discussion, we will call this system the status assessment (SA) for each clinical visit.

Within the MGD, most physicians appear to summarize patients' clinical assessments as a single component, which includes the formal MSE, ad hoc observations, and symptoms,

[1] The anonymization of clinical records is performed on-site, following "Safe Harbor" specifications of the Health Insurance Portability and Accountability Act (1996), with the full consent of the individual hospitals, before it is extracted and put into a common database. These anonymized data are available to all hospitals that also contribute data to the MGD for academic research.

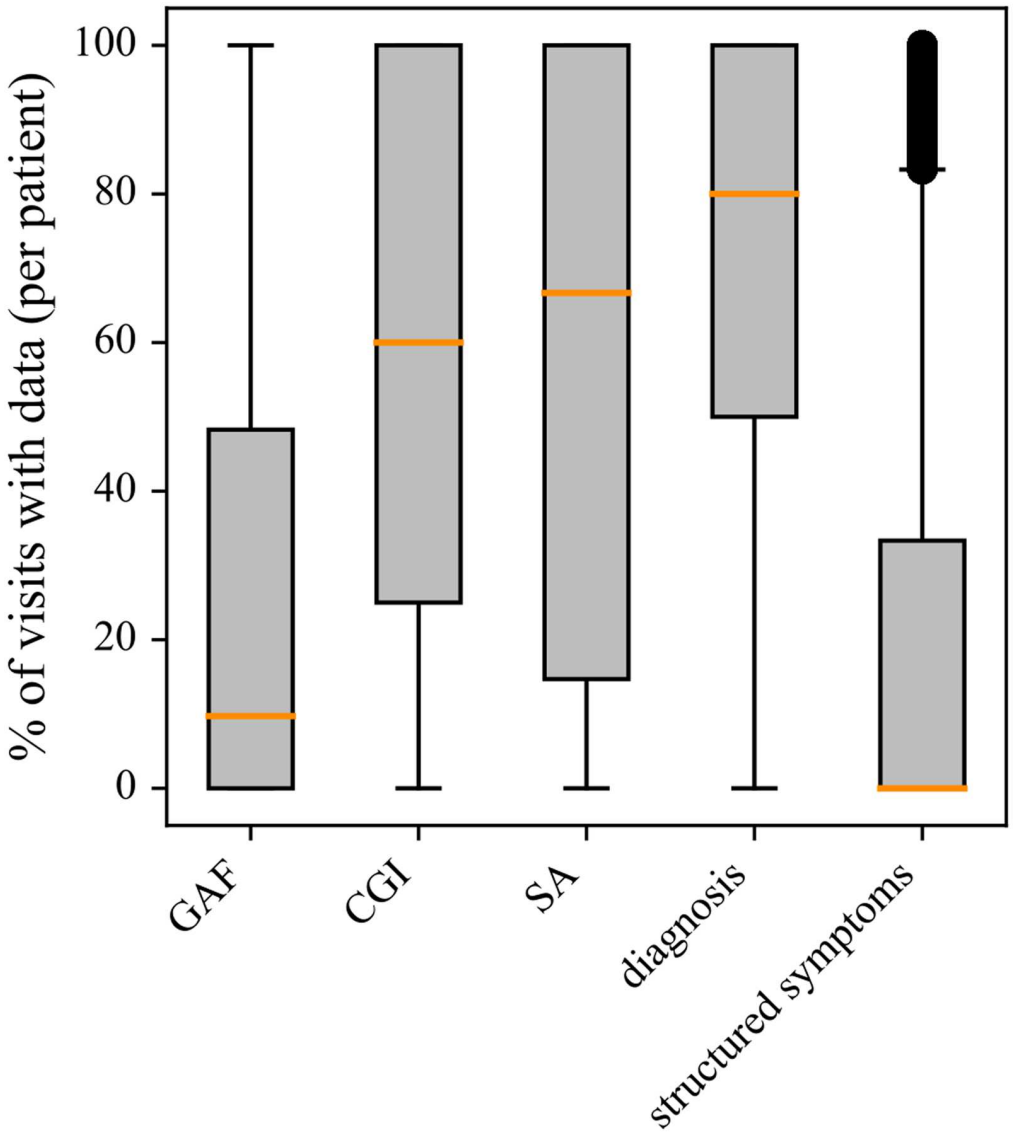

Figure 1. Boxplots showing the distribution of the percentage of visits that record different psychiatric assessments within the MindLinc Global Database (MGD) over the entire population of patients. It is worth noting that the Clinical Global Impression (CGI), status assessment (SA), and diagnosis are generally recorded for the majority of the visits, while the Global Assessment of Functioning (GAF) and symptom information are infrequently recorded within the MGD.

within the SA rather than to record symptom information within structured EHR fields. Therefore, structured symptom information is unavailable for most encounters (median 0%), whereas the SA is available for the majority of the visits (median 67%), along with the diagnosis (median 80%) and the Clinical Global Impression (CGI; Guy, 1976) score (median 60%), as shown in Figure 1.

Although the SA allows clinicians to enter information about the mental health of the patient in a very flexible manner, it also promotes the accumulation of unstructured data in place of structured fields. In MGD, for example, the information for SA is present as a (category, sign) tuple. Among the 25 hospitals that contribute data to the MGD, both the categories and

the signs lack standardization. SA data are categorized into 69 separate categories, many of which represent the same information, and should be regrouped into a smaller set of standard categories. The sign is usually represented as a free-text description. SA is one of the richest categories of data available from real-world psychiatric practice, but it is practically impossible to use for analytical purposes because of its unstructured nature. From a psychiatric perspective, the SA provides a granular, multidimensional representation of the mental state of patients at the time of consultation. Hence, significant improvements in our understanding of different mental state disorders can be achieved if the information contained in this SA is appropriately harnessed.

In the last decade, great progress has been made in the field of natural language processing (NLP). This is especially true with the advent of deep learning–based NLP. Currently, some of the best algorithms for part-of-speech tagging (Huang, Xu, & Yu, 2015), parsing (D. Chen & Manning, 2014; Dyer, Ballesteros, Ling, Matthews, & Smith, 2015; Zhou et al., 2017; Zhu, Zhang, Chen, Zhang, & Zhu, 2013), named entity recognition (Chiu & Nichols, 2016; Passos, Kumar, & McCallum, 2014), sentiment classification (Mikolov, Sutskever, Chen, Corrado, & Dean, 2013), machine translation (Koehn, Och, & Marcu, 2003), and contextual embeddings (Q. Chen, Zhu, Ling, Wei, & Jiang, 2016; Liu, Shen, Duh, & Gao, 2017) are done with deep learning–based NLP. Most medical databases, including those within hospitals, clinics, and pharma companies, are rife with textual data, including structured and semistructured fields, clinician notes, reports, and so on. More recently, language models, such as Deep Bidirectional Transformers for Language Understanding (BERT; Alsentzer et al., 2019; Devlin, Chang, Lee, & Toutanova, 2018), Generative Pretraining Transformer (GPT; Radford et al., 2019), and XLNet, have been proposed that are able to disambiguate the context of the meanings of the same words used in different contexts. Context-specific models based on BERT have been generated for the medical domain using articles from Wikipedia, PubMed, and PMC, in the form of BioBERT (Lee et al., 2019). It has been shown that clinical notes have language and nomenclature that are different from the text present in the journal articles of PubMed and PMC. Another pretrained BERT-based model, called ClinicalBERT (Alsentzer et al., 2019), has been trained from clinical notes available in the MIMIC (Johnson et al., 2016) database. ClinicalBERT has been shown to outperform BERT and BioBERT in NLP tasks catering to clinical notes.

Deep learning–based NLP has been used for mining text data and obtaining meaningful information from clinical notes, as well as from social media, and falls under the purview of information extraction from free text (Jing, 2012). Deep learning–based NLP has been used in many diverse applications, such as recognizing adverse drug interactions from social media (Wunnava, Qin, Kakar, Rundensteiner, & Kong, 2018), predicting health care trajectories from medical records (Pham, Tran, Phung, & Venkatesh, 2017), predicting early psychiatric readmission from discharge summaries (Rumshisky et al., 2016), extracting symptoms of severe mental illness from clinical texts, text-based phenotyping, subdomain classification (Jackson et al., 2017), drug–drug interactions (Xu, Shi, Zhao, & Zheng, 2018), and hospital mortality prediction (Wray et al., 2018). In this article, a combination of NLP and the expertise of a subject matter expert (SME) has been used for creating a system that will allow for the conversion of the SA data into a standardized binary vector. This standardization will not only allow different patients from disparate hospitals to be compared with one another but also allow these vectors to be used in analytical and machine learning (ML) algorithms to better predict clinical outcomes.

In the section "Method," the SA data are described in detail. Furthermore, the reclassification of categories and free text within the SA is also described. Since the data are not

error-free, methodologies used to preprocess the free text for correcting the most prominent errors have also been described. Apart from above, the NLP neural network architecture is explained in detail in the section "Method." The section "NLP Classification Model" lays out the detailed training process and the testing of the long short-term memory (LSTM)-based NLP model, and it compares the performance against other traditional ML models. In the section "SA Vector Use Cases," a few use cases have been described that utilize the SA vector output of the NLP model. These use cases vary in complexity, from the development of simple suicidality scales to clinical phenotyping. These use cases show the value of a quantitative set of multidimensional patient symptoms in a physiologically relevant quantitative framework.

METHOD

SA Overview

Inside the MindLinc EHR, the SA is the most abundant clinical assessment performed by the clinicians and is one of the largest sources of data in MGD. A screenshot of the EHR front end resulting in the generation of the SA data is shown in Figure 2.

As shown, the SA within the EHR is categorized into several functional categories, such as appearance, attitude, cognition, intelligence, orientation, and perceptual, which may be amended by the physician adding more categories, as necessary. Each category has one or more items (called "signs") that describe the patient within that particular category. These signs are typically free-text entries entered by the physician, as shown at the bottom of Figure 2. This section has been zoomed-in within the picture so that it is easier to see the text box and the suggestion "Pick a category then type a sign." A full list of categories and the number of unique

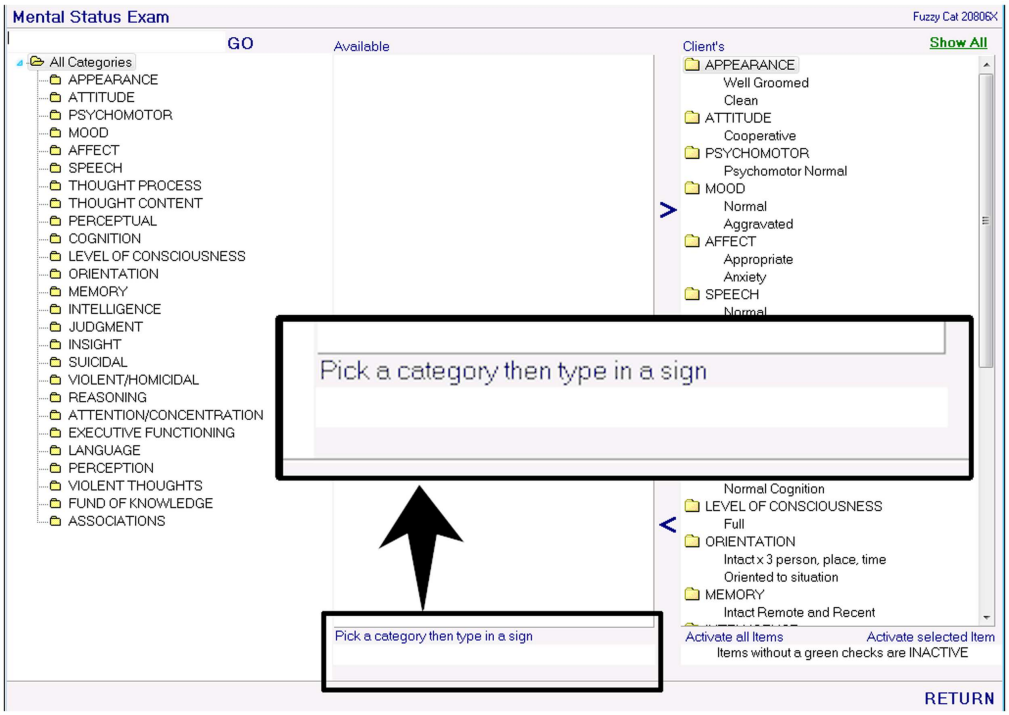

Figure 2. A screenshot of the electronic health record (EHR) software used for inputting status assessment (SA) data into the database.

Table 1. The different reclassified categories, along with the number of unique signs associated with each

		Total counts		Percentage validated		
ID	**Reclassified categories**	**Unique signs**	**Total signs**	**Unique**	**Database**	**Groups**
1	Abnormal or psychotic thoughts	60,115	53,715,764	0.83	1.08	1
2	Affect	29,559	22,417,497	4.27	97.78	1
3	Appearance	5,743	21,673,586	97.79	98.68	1
4	Association	158	538,919	100	100	1
5	Attention/concentration	2,704	7,551,799	23.56	99.85	1
6	Attitude	17,451	19,525,590	3.4	99	1
7	Cognition	7,639	10,789,531	0.26	0.67	3
8	Executive functioning	1,525	7,002,438	12.33	99.89	1
9	Fund of knowledge	162	710,150	83.95	99.99	2
10	Gait and station	440	659,938	9.32	24.68	3
11	Homicidal	3	9,509	100	100	3
12	Impulse control	19	531,981	100	100	3
13	Insight	3,733	15,389,888	2.76	3.79	2
14	Intelligence	1,809	12,563,500	2.6	6.09	3
15	Judgment	4,225	15,899,780	1.04	3.67	3
16	Language	946	6,086,767	10.47	99.82	2
17	Level of consciousness	1,199	9,909,668	8.26	99.89	2
18	Memory	6,302	12,884,403	0.4	4.24	3
19	Mood	40,377	26,680,794	0.26	94.36	2
20	Orientation	6,698	14,479,475	0.52	9.4	3
21	Psychomotor	21,183	13,837,858	0.47	94.71	2
22	Reasoning	1,534	6,494,647	6.45	99.54	2
23	Sensorium	301	220,005	32.89	65.9	2
24	Sleep	24	80,929	100	100	3
25	Speech	20,374	16,214,748	0.68	95.41	2
26	Suicidal	19,273	15,001,434	0.51	28.93	2
27	Violent thoughts	13,841	21,200,549	0.72	32.95	2

signs associated with each category are available in Table 1. These signs represent free-text descriptions of the patient associated with that particular category.

Examples of signs associated with the category "affect" are "able to smile and laugh appropriately," "abruptly tearful at times," and "actually appears euthymic but displays some anxiety." As is evident from these examples, this category represents the psychiatrists' impressions of the patient at a particular time.

On the other hand, another category, "orientation," represents the results of specific cognitive tests performed by physicians on the patient with the help of questionnaires to check their orientation to time, location, and so on. Examples of such questions are as follows:

- What is your full name?
- Where are we at (floor, building, city, county, and state)?

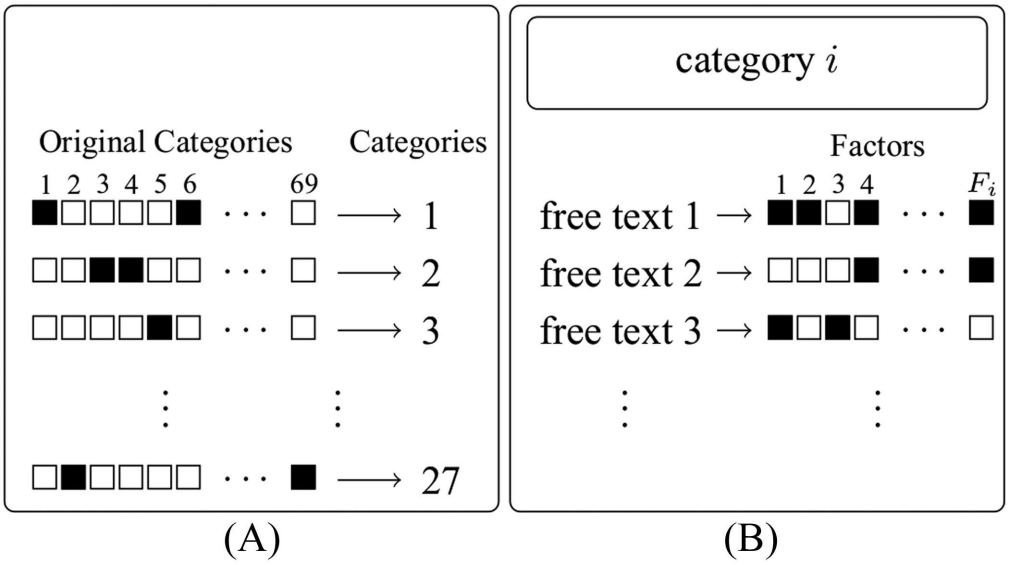

Figure 3. An overview of the process that is used for converting the semistructured status assessment (SA) into the reclassified SA vectors. A) The subject matter expert (SME) first reclassifies the original 69 categories into 27 categories. B) After the reclassification of the categories, the SME converts the free text (called signs) in each of the 27 categories into subcategories called factors.

- What is the full date today (date, month, year, day of the week, and season of the year)?
- How would you describe the situations we are in?

The patient's orientation and assessments resulting from the answers to these questions are entered into a database in the form of free text. Examples of entries for the orientation category are "intact oriented for all the four questions" and "intact oriented $\times 4$."

As mentioned before, the combination of inputs from a SME and a set of ML models, working in tandem with the SME, has been used for this vectorization process. The first part of this process involves human annotation, and this process is schematically depicted in Figure 3. The second part is a combination of a human classification process and machine learning, which progressively improves the ML algorithms over time. The overview of this is shown in Figure 4. These two steps are described in detail in the next two subsections.

Recategorization of SA

The original SA comprises a total of 69 categories. Since the data within the EMR are collated from several different hospitals, categories defined in one hospital are not identical to those defined in another. This is because the MindLinc database allows clinicians to define new categories as they deem necessary. Thus, in many cases, what should be represented by a single category is represented by a number of categories with similar names, each originating from a different hospital.

Since the SA comprises a (category, sign) tuple, the categories were standardized first. A SME first reclassified the original 69 categories into 27 by merging similar categories into a single category, as shown in Figure 3A. The details of the original 69 categories, and the new reclassified categories into which they have been merged, are provided in Appendix A of the Supporting Information. This is a many-to-one mapping and is performed once. The new

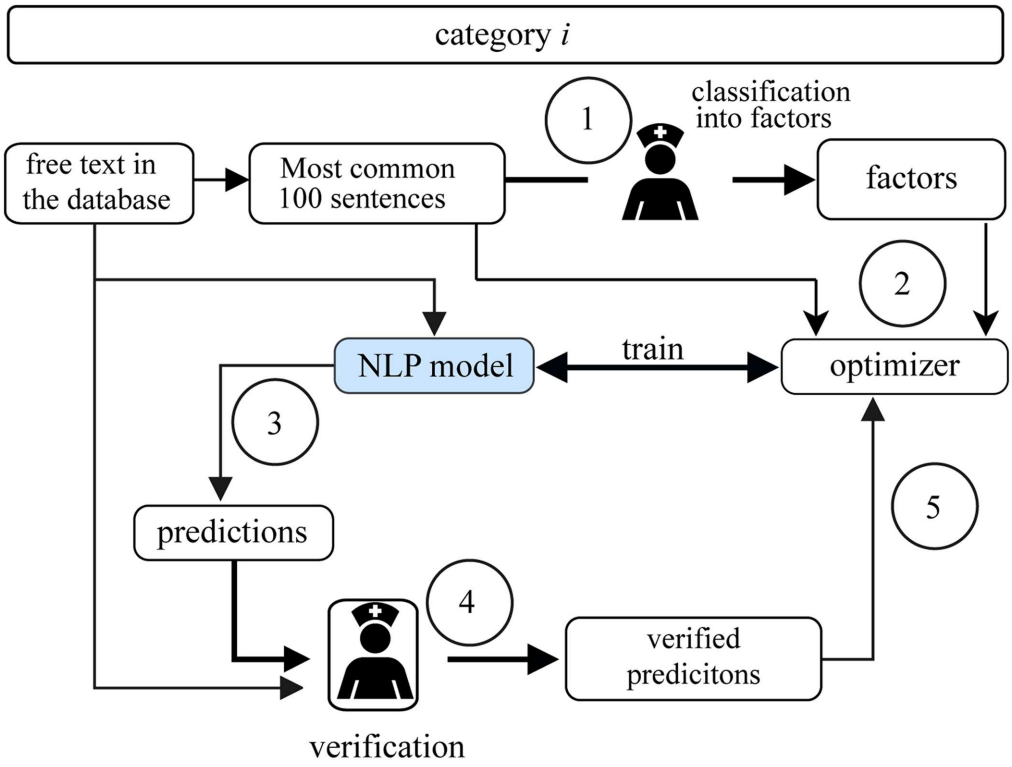

Figure 4. Details of the process of conversion of status assessment (SA) signs into reclassified SA factors for each category. 1) The manual classification of the most common 100 sentences is performed first. 2) Then, an optimizer is used for training the parameters of a natural language processing (NLP) model. 3) This optimized NLP model is then used for generating a set of new predictions, 4) which is verified by the subject matter expert (SME). 5) This new information is used for reoptimizing the NLP model. The three Steps 3, 4, and 5 define an iterative process, which, over a number of iterations, provides a reasonable set of optimized models.

reclassified categories created in this manner are subsequently known as "category," unless otherwise specified.

SME Categorization

For every category, the SME generates a set of subcategories after reviewing the set of available signs within that category, as shown in Figure 3B. These subcategories represent the major classifications that summarize the major aspects of information that clinicians typically want to capture within that category. For example, the major subcategories into which the category "language" is categorized are "intact," "neutral/unable to categorize," "repetition intact," "issues with repetition," "object naming intact," "issues with object naming," "impaired," "nonverbal/mute," "minimally verbal," and "issues related to DD." With modern NLP technologies, it is possible to convert free text describing, for example, the linguistic abilities of a patient into one of the aforementioned subcategories. A combination of these subcategories would be able to describe the state of a person within that particular category. A subcategory within a category is called a "factor" of a particular category. A list of factors associated with each category is tabulated in Appendix B of the Supporting Information.

After classifying the original categories into the reclassified categories, the signs in each category are subsequently categorized into factors, as shown in Figure 4. As mentioned before,

a sign represents free text that has been entered by the physician describing the patient. An example of such a sign within the category "abnormal or psychotic thoughts" is "he says he sees a chicken and fish eating each other in my office." This is classified by the SME as the factor "experiencing hallucinations (visual)," in the segment shown in Figure 4, Step 1. The factor_id represents the index of a factor within a category. Notice that the factor_id of this particular factor is 5, as represented in Row 6 of Appendix B of the Supporting Information. Hence, this particular string is represented by the vector [0, 0, 0, 0, 0, 1, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0] such that the value of the sixth index is set to 1 and the values of all the other digits are set to 0. It is important to remember that the patient might be simultaneously experiencing auditory hallucinations as well. A patient experiencing both auditory and visual hallucinations will be represented by the vector [0, 0, 0, 0, 1, 1, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0] (referred to as the categorical vector hereafter), wherein the fifth and the sixth digits of the vector are set to 1 and the rest are set to 0, in accordance with factors defined in Appendix B of the Supporting Information.

As one might well imagine, the amount of information that is available within the different categories is different. It is interesting to note that many clinicians use the same phrase for describing a particular mental health condition. For example, the phrase "not present" occurs more than 11 million times while describing patients' "violent thoughts." Similarly, "average" occurs more than 9 million times while describing "intelligence," and "without hallucinations" occurs more than 8 million times while describing the "abnormal or psychotic thoughts" of a patient. It goes to reason that classifying the most common signs would allow significant portions of the SA database to be mapped into meaningful categories. This provides a ground truth against which different NLP models may be trained. Once the SME has categorized a few of the most popular signs, a NLP algorithm is optimized to learn these classifications, as shown in Figure 4, Step 2. This allows the NLP algorithm to predict the next few most populous signs, as shown in Figure 4, Step 3. These predictions are then validated by the SME to generate a new set of data representing ground truth, as shown in Figure 4, Step 4, which can then be used to reoptimize the NLP algorithm, as shown in Figure 4, Step 5. Once more signs have been validated by the SME, the models are retrained to incorporate these new data.

Table 1 shows the total number of signs available within different categories, along with the total number of unique signs available in each category. The percentage of unique signs that have been validated, and the resultant percentage of the total number of signs in the database that have been validated as a result, for each category, are tabulated in the same table. As can be seen, 15 out of the 27 categories have over 95% of the signs validated.

The final SA vector (SAV) is generated by concatenating the entire set of 27 categorical vectors into a single vector that is 241 bits long. The bit positions for each of the categories and factors in the reclassified SAs are tabulated in Appendix B of the Supporting Information. The factor in the reclassified SA vector is one that is assigned by the SME if available, else it is one that is predicted by the NLP algorithm.

Preprocessing

Preprocessing comprises of all the steps that ultimately lead to the conversion of sentences into a list of word vectors (Bengio, Ducharme, Vincent, & Janvin, 2003). Before cleaning the signs, all signs are grouped into unique signs independent of the category to which they belong so that each unique sentence is cleaned only once. This results in 267,337 unique signs. Note that the same sign may appear in multiple categories. For instance, "mildly impaired" appears in nine different categories but is preprocessed only once for expediency.

Tokenization is performed by splitting sentences on spaces. Stemming or lemmatization is not performed, as we wish to preserve as much of the meanings of the various words as possible, including their tenses.

After tokenization is performed, a vocabulary is generated, along with a list of words present in all the signs. Subsequently, Google's pretrained word2vec model is used for transforming the words in the vocabulary into word vectors. Not all words in the vocabulary have a corresponding word vector because of reasons such as misspellings, abbreviations, or the presence of numbers. Words that are not present in the pretrained model are dealt with in the following manner:

1. Common words, such as "a," "to," "and," and "of," are removed
2. Incorrectly spelled words have been corrected.
3. Contractions, abbreviations, and acronyms have been expanded.
4. Each digit in numbers has been replaced with the "#" character. This step is specific to the way in which Google's word2vec handles numbers. For example, the number 54 is replaced with the string "##."
5. Specific items have been substituted with descriptive words, typically associated with the "orientation" category. Examples of such substitutions are as follows:
   a. Anything representing a date (like 12/10/1996) has been replaced with the word "date."
   b. Expressions representing equations (such as 10 + 2 = 12) have been replaced with "maths."
   c. Hospital names and locations have been replaced with "place."
6. All punctuation has been removed.
7. Words with correct spelling but not vectorized by Google's word2vec model have been replaced by their synonyms.

Examples of signs and how they have been cleaned are shown in Table 2.

Table 2. A summary of the different processes used for preprocessing the free-textual signs

Sign	Clean sign	Tokens	Action
today there are no looseness of association or flight of ideas.	today there are no looseness association or flight ideas	{today, there, are, no, looseness, association, or, flight, ideas}	Removed "of"; removed "."
217 lbs on 10/27/11	### lbs on date	{###, lbs, on, date}	Replaced numbers with "#"; replaced "10/27/11" with "date"
history of a/v hallucinations reported and history of paranoid delusions reported	history audio visual hallucinations reported history paranoid delusions reported	{history, audio, visual, hallucinations, reported, history, paranoid, delusions, reported}	Replaced "a/v" with "audio visual"; removed "and" and "of"
slightly dishelved	slightly disheveled	{slightly, disheveled}	Replaced "dishelved" with "disheveled"
mmse 28/30	mini mental state examination ## over ##	{mini, mental, state, examination, ##, over, ##}	Replaced "mmse" with "mini mental state examination"; replaced fraction with"'## over ##"

NLP CLASSIFICATION MODEL

In this section, the classification model, along with its training procedure and associated results, is described. The results obtained from this model are also compared with models trained from traditional classification approaches, and it is shown that deep learning–based models for NLP are generally superior to classical models.

Architecture

Figure 5 illustrates the general training process of the NLP model for the *i*th category. As discussed earlier, free text for the *i*th category (s_i) is preprocessed, and each unique word in the sentence is mapped to a vector of length 300. All word vectors in a sentence are then concatenated into a two-dimensional matrix with a size of (n_w, 300), where n_w is the number of words in the sentence. Since sentences within a category have different lengths, the maximum number of words within all sentences in the category *i* is represented by *N*. Matrices for shorter sentences (those that have fewer than *N* words) are resized to (*N*, 300) by padding the original matrices with zeros.

As shown in Figure 6A, a basic NLP unit for category *i* takes in free text ($\mathbf{s}_i$) and classifies it into a binary output vector ($\mathbf{o}_i$). The detailed NLP unit architecture is presented in Figure 6B.

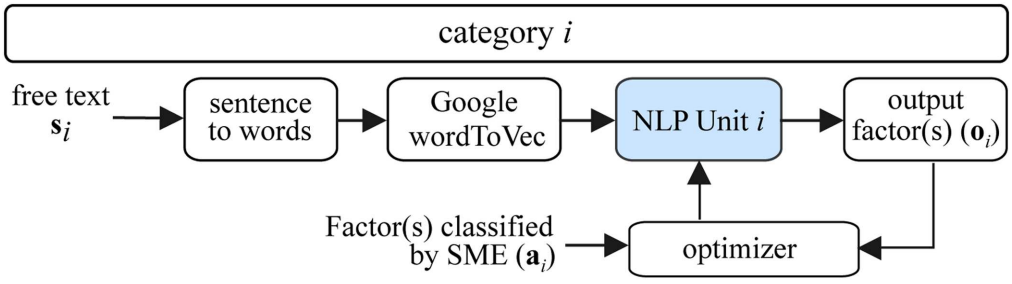

Figure 5. A general model training process for a free-text classification task.

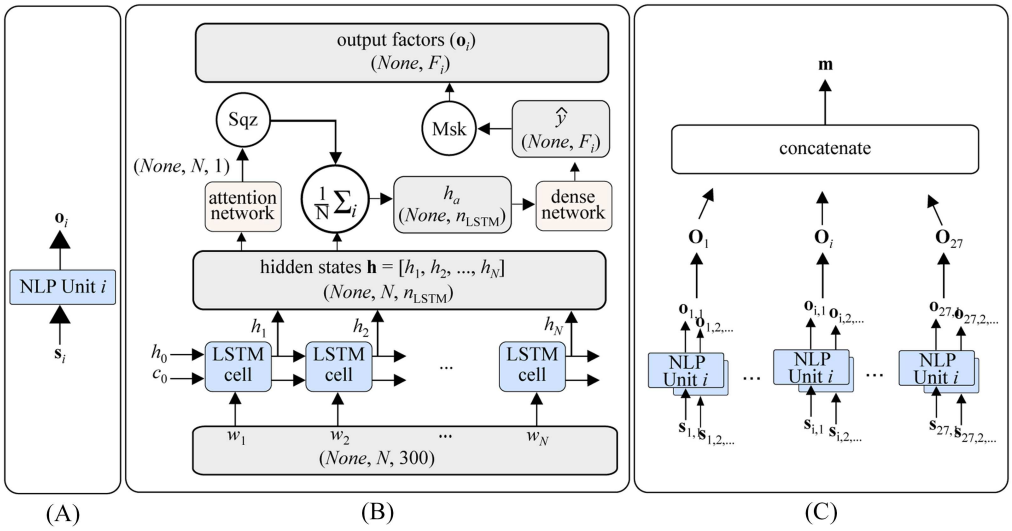

Figure 6. Natural language processing (NLP) unit. A) Basic NLP unit *i* for category *i* with input free text (s_i) and output factor (o_i). B) The detailed architecture of NLP unit *i*. Numbers in parentheses indicate the tensor shapes. C) Factors predicted from 27 NLP units concatenated as an integrated status assessment (SA) result (**m**).

A sequence of words (w_1, w_2, …, w_N) from a sign together with initial hidden (h_0) and memory cell states (c_0) is fed into a recurrent neural network (RNN; Lipton, Berkowitz, & Elkan, 2015) comprising a single LSTM cell. The NLP algorithm has been implemented using custom code written in TensorFlow Version 1.11 (Martín Abadi et al., 2016). A dynamic LSTM cell (Martín Abadi et al., 2016) has been used for constructing the model. Note that in TensorFlow, the first tensor dimension does not need to be declared during the definition of the model and will be dynamically adjusted during runtime, based on the batch size of the input to the model. This is generally provided by setting the first element of the shape of the tensor to *None* within the model definition. Initial states h_0 and c_0 are set as zero vectors. The input size is (*None*, *N*, 300), as shown in the parentheses, where (*N*, 300) is the size of a single sign matrix and *None* stands for the mini-batch size, which is automatically handled at runtime.

Series of hidden states (h_1, h_2, …, h_N) and memory cells (c_1, c_2, …, c_N) are generated as outputs from the LSTM cells. Note that the dynamic LSTM layers in TensorFlow map zero-padded inputs to zero vectors for the hidden states, and memory cells, when the length of the inputs (n_w) specifying the number of valid words present in each sequence is provided to this dynamic LSTM cell as an input. *N* hidden states from a batch of signs form a matrix with size (*None*, *N*, n_{LSTM}).

In this study, an attention layer is applied to combine information across all hidden states by calculating the weighted average of them (h_a). The weights for individual hidden states, which can be regarded as the amount of "attention" paid to every hidden state, are generated using a dense model, hereinafter referred to as the attention network, with all *N* hidden states as the inputs.

The attention network consists of two or three fully connected layers (the number of hidden layers in the attention network is a hyperparameter as discussed in the section "Model Training"), with tanh activation in all layers except the output layer, where either a sigmoid or a softmax activation is employed to confine the attention between 0 and 1. The choice of the activation function (sigmoid vs. softmax) is also a hyperparameter, and the results are presented in the section "Evaluation Metrics." The outputs of the attention layer, or the weights with size (*None*, *N*, 1), are then squeezed into size (*None*, *N*) before the elementwise multiplication with the hidden states to generate the hidden states with attention having size (*None*, *N*, n_{LSTM}). Finally, the hidden state with attention (h_a) is calculated by taking the mean of the *N* hidden states, yielding a vector h_a of size (*None*, n_{LSTM}).

To classify h_a into one of the F_i factors, it is passed through a dense network with two or three hidden layers with tanh activation, followed by a final dense layer with sigmoid activation. The output of the dense network ($\hat{y}$) with a size of (*None*, F_i) is passed through a mask layer to generate the final output (o_i). In the mask layer, elements greater than 0.5 in $\hat{y}$ are set to be 1, and the rest of the elements are set to 0.

The loss function for the optimizer to train on is the mean squared error between the predicted output (o_i) and the label (a_i), as shown in Equation 1, where Θ represents the model parameters, *m* is the number of samples, and F_i is the number of classes for category *i*. $y_k^{(j)}$ is the label for sample *j* and class *k*, while $\hat{y}_k^{(j)}$ is the output of the dense network. The loss of a cross-validation (CV) set is monitored during the training, and early stopping is applied to interrupt the training process if CV loss does not drop over 10 epochs:

$$J(\hat{y}|\Theta) = \frac{1}{m \times F_i} \Sigma_{j=1}^{m} \Sigma_{k=1}^{F_i} \left(y_k^{(j)} - \hat{y}_k^{(j)}\right)^2. \quad (1)$$

To avoid overfitting and improve the robustness of the model, Gaussian noise is added to the weights in the dense network (except for the output layer) during the training. Twenty-seven unique NLP units are trained independently and predict factors (o_i) for each category. The resultant predicted factors are concatenated to form an integrated SA vector (Figure 6C) with a length of 241, corresponding to the 241 factors (enumerated in Appendix B of the Supporting Information). Note that each category may have multiple signs and corresponding predictions of factors ($o_{i,1}, o_{1i,2}, \ldots o_{i,k}$). The high bits in the final categorical vectors (O_i in Figure 6C) are a union of all high bits from $o_{i,1}, o_{i,2}, \ldots o_{i,k}$.

Model Training

In this section, the procedure used for architecture optimization, hyperparameter tuning, model training, and model testing is described. The efficacy of the resultant models is measured using multiple metrics, such as precision, recall, the F_1 score, and the area under the receiver operating characteristics (AUROC) curve. Finally, several use cases for the SA vector within clinical settings are described. Optimization and evaluation involve the generation of training, validation, and testing sets from the original data. Because some categories do not contain a significant number of labeled/validated data, categories are split into different groups, as tabulated in Table 1, depending on the total number of categories available. Ways in which data have been split into categories in the different groups are explained in the following paragraphs.

First, nomenclature that is consistent with current artificial intelligence literature (Hastie, Friedman, & Tibshirani, 2001) is defined:

- The *test set* (or the *holdout set*) is the set used for evaluating the model. These are data that the model does not see until the model parameters and hyperparameters are trained.
- The *validation set* is used for optimizing the hyperparameters of the model. Very often, the training and validation sets are generated multiple times for a better estimation of the validation error in the form of cross-validation—something that has been followed here for categories in Groups 1 and 2.
- The *training set* is used for optimizing the parameters of a model for a given set of hyperparameters.

The generation of the training and test metrics is schematically represented in Figure 7.

Owing to the paucity of unique values in some categories, it is unfortunately not possible to generate sufficient samples that would lead to the creation of separate training, testing, and validation sets. Certain categories, such as "sleep" and "homicidal," have only 24 and 3 unique signs, respectively. These are categorized into Group 3. Others, such as "fund of knowledge," "language," and "sensorium," have 162 and 301 unique factors, respectively. However, not all of these are labeled/validated by the SME. For example, for these categories, only 136 and 99 unique signs are actually labeled/validated, respectively. These are put in Group 2. Then there are some groups for which more signs have been labeled and validated. These are put into Group 1.

Since different amounts of labeled data are available in the different groups, the train–validation–test splitting procedures for the different groups are also slightly different. These differences are explained in Figure 7.

For signs that belong to categories in Group 1, an 80–20 split for the training and test datasets has been performed. Using the training dataset, the model hyperparameters are first

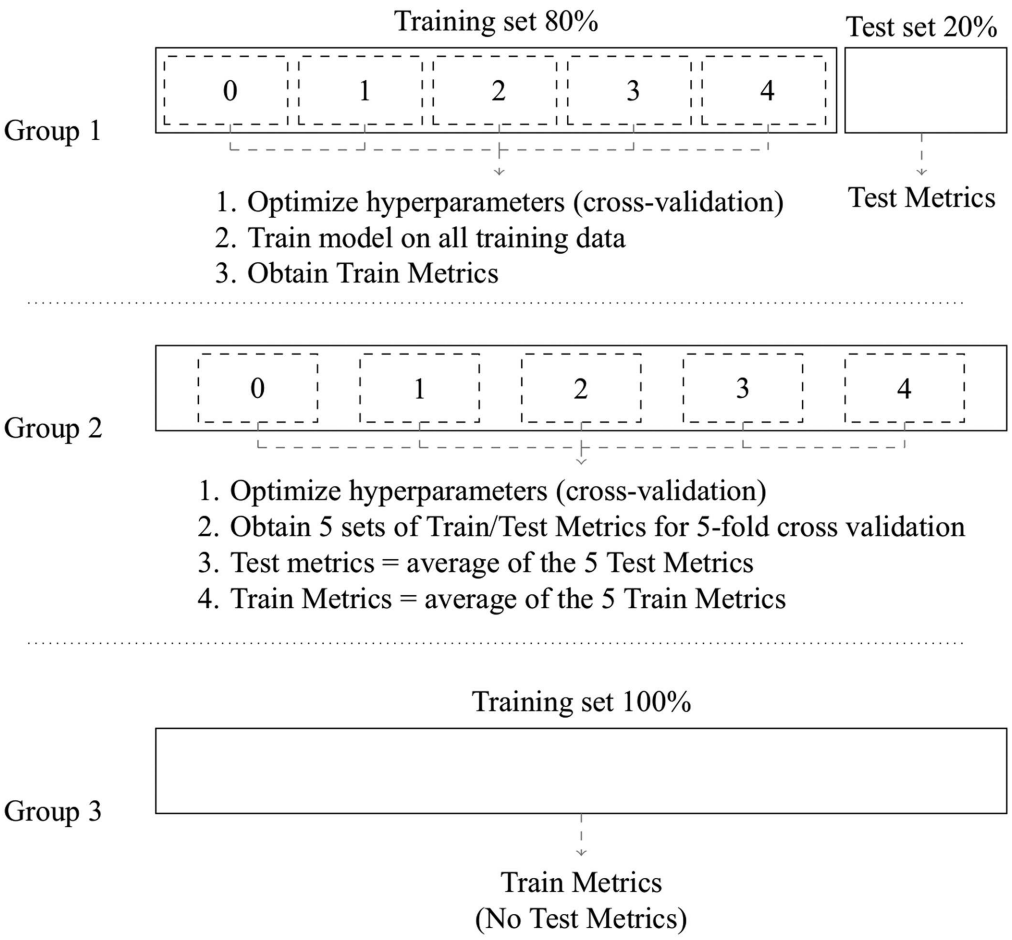

Figure 7. Schematic of the methodology used for splitting data into training, validation, and testing datasets for generating the training and testing metrics.

optimized using a fivefold CV. Once the model hyperparameters are optimized, a new model is trained on the entire training dataset using these hyperparameters. This final training produces the train metrics for Group 1. The test dataset that has been held out is subsequently used for generating the test metrics for Group 1.

Signs belonging to categories in Group 2 do not have sufficient labeled data for generating separate training and test datasets. For this reason, a separate method is employed. The entire dataset is used for training the hyperparameters using fivefold CV. Once this is done, a fivefold CV is employed for generating five sets of train and test metrics. The resultant train and test metric is the average of the five train and test metrics generated earlier.

Signs belonging to categories in Group 3 contain so few unique values that it is not meaningful to generate any realistic models. For completeness, models have been trained on all available data for categories in this group, and only train metrics have been calculated. These metrics are provided only for completeness and not for the explicit purpose of creating NLP models for prediction.

A mini-batch of sentence matrices of size (n_{batch}, N, 300), where n_{batch} is the batch size, is fed into the ith NLP unit for a multiclass (F_i classes), multilabel classification task. An output

vector o_i is generated and compared to the ground truth (a_i). Both o_i and a_i are binary vectors with a length of F_i. The NLP model is then optimized to diminish the difference between a_i and o_i. Note that the embedding size (300) is kept the same for all 27 categories, whereas the sentence length (N) and output factor size (F_i) are determined by the sizes of inputs and labels for each category.

Evaluation Metrics

Accuracy is computed to assess the generalization of the model. Accuracy (acc) of the NLP unit i is calculated using Equation 2, where eleCheck is a function on NLP unit i output ($o_i^{(j)}$) and SME classified factor(s) ($a_i^{(j)}$) for the jth sign. The function eleCheck returns 1 only when every element in $o_i^{(j)}$ matches with the elements in $a_i^{(j)}$:

$$\text{acc}\,(\Theta) \;=\; \frac{1}{m}\Sigma_{j=1}^{m}\text{eleCheck}\left(o_i^{(j)}, a_i^{(j)}\right). \tag{2}$$

Other than accuracy, the precision, recall, F_1 score, and AUROC are also calculated. These measures are calculated for every factor and subsequently averaged per category for plotting.

Hyperparameter Optimization

During training, the following hyperparameters are tuned and optimized using a fivefold CV: the learning rate, the size of the LSTM cell, numbers of neurons and layers for both the attention and dense networks, the standard deviation of the Gaussian noise in the dense network, and the type of activation function of the last attention layer. Note that each category has its own set of optimized hyperparameters, except for the activation function of the last attention layer, where "sigmoid" is chosen for all 27 categories. To optimize this hyperparameter, two different activation functions are tested, that is, sigmoid and softmax.

The attention assigned to each word in 16 unique signs is presented in Figure 8 as Hinton diagrams, wherein the length of a box represents the magnitude of attention paid to the hidden state corresponding to a word shown on the left. Generally, sigmoid tends to pay distributed and high attention (close to 1) to most of the words in a sign. This result is sensible, considering the relatively short sentence lengths within the categories (95th percentile of the sentences has a word length of six or fewer) and the fact that the resultant hidden states are semantically relevant.

In contrast to the sigmoid activation, the softmax actively highlights important words within the sentence. Consequently, attention is allocated to one or two key words within a sentence. At first glance, the key words proposed by the softmax activation appear to be less meaningful from a human perspective in many instances. Some examples are listed in Figure 8A, where key words for a diagnosis of schizophrenia, such as "delusional," "delusions," and "hallucinations," are undervalued. However, one should notice that hidden states, which are utilized to generate attention vectors, are semantically and syntactically correlated, as we discussed earlier. In other words, the amount of attention that the RNN pays to an individual hidden state is not necessarily equivalent to the attention paid to an individual input word. During the training of the RNN, hidden states generated later in a sequence tend to have richer information than earlier ones and incorporate information about previous words as well. Hence attention routinely gives importance to hidden states that lag important words in a

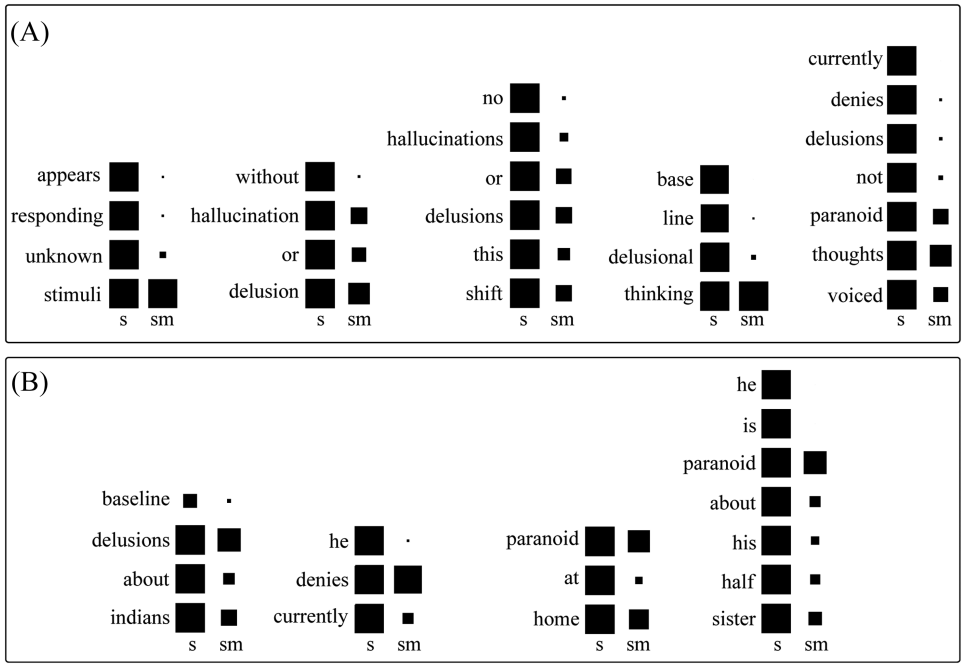

Figure 8. Hinton diagram of attentions generated using sigmoid (s) and softmax (sm) as shown in the bottom of each subdiagram. The amount of attention is illustrated as the box length. A) Signs for which softmax produces less than satisfactory attentions. B) Examples where softmax gives precise attention.

sentence. When important words are separated from one another by many unimportant words, softmax can generate precise attention vectors on some signs, as shown in Figure 8B.

In this study, sigmoid is adopted given that it achieves a slightly higher accuracy score (as discussed in the section "Evaluation Metrics") on the test set (0.71 for Category 1, while softmax produces an accuracy of 0.70 on the same category).

Comparison With Other Models

To compare the strength of the aforementioned model in extracting labeled signs from text, we have developed three separate types of models for each category: support vector machine (SVM) models, *K*-nearest neighbor (KNN) models, and naive Bayes models. All three sets of models were developed for each of the 27 categories and were trained and tested on the same data. The input to the models are generated using a binary encoded vector containing the most common 300 words within the vocabulary using a bag-of-words approach, so that the input sizes of the vectors used for both the LSTM model and the more standard classification models are the same. Each model is then trained with exactly the same data as the LSTM models with the same splits that are used for training the LSTM models. After training, the accuracy, precision, recall, F_1 score, and AUROC are calculated for the test sets in Groups 1 and 2 and training sets in Group 3. They have also been plotted in Group 3.

Results

Data for 27 categories are separated into three groups based on their sample size (refer to Table 1, columns "Validated Counts" and "Groups"). The accuracy, precision, recall, and F_1 scores have been averaged for all factors within a category.

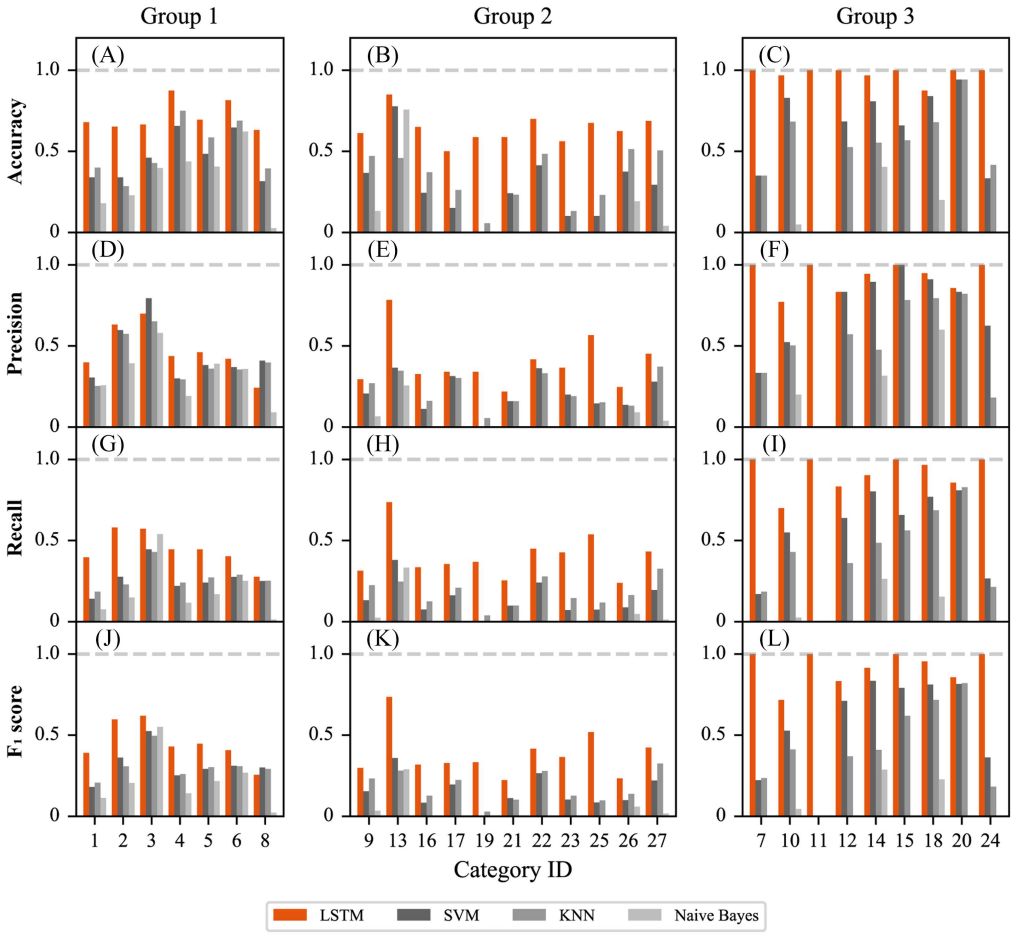

Figure 9. Classification metrics for accuracy, precision, recall, and F_1 scores plotted for all categories. As can be seen, multiple models have been trained for each category and their scores compared. For categories belonging to Group 1 and Group 2, only test metrics have been plotted, as defined at the beginning of the section "NLP Classification Model." For Group 3, only training metrics have been plotted, due to the paucity of data.

Figure 9 displays the results of the four models for all categories. As described previously, the categories are divided into three groups based on the number of labeled data present in each group. The accuracy, precision, recall, and F_1 score have been plotted for each category. Also plotted are comparative scores for baseline models of SVM, KNN, and naive Bayes.

For Groups 1 and 2, only the test metrics, as defined in the section "NLP Classification Model," have been plotted. For Group 3, owing to the unavailability of sufficient labeled/validated data, the train metrics have been plotted. As mentioned before, this is done only for completeness. As can be seen, the LSTM model performs better than the baseline models, except in the precision of Category 3 and Category 8.

Other than the metrics of accuracy, precision, recall, and F_1 score, the AUROC is also calculated for all of the 241 factors. The distribution of the results is shown in Figure 10. As can be seen, the median AUROC is approximately 0.9 for the LSTM model and less than 0.8 for the rest of the models.

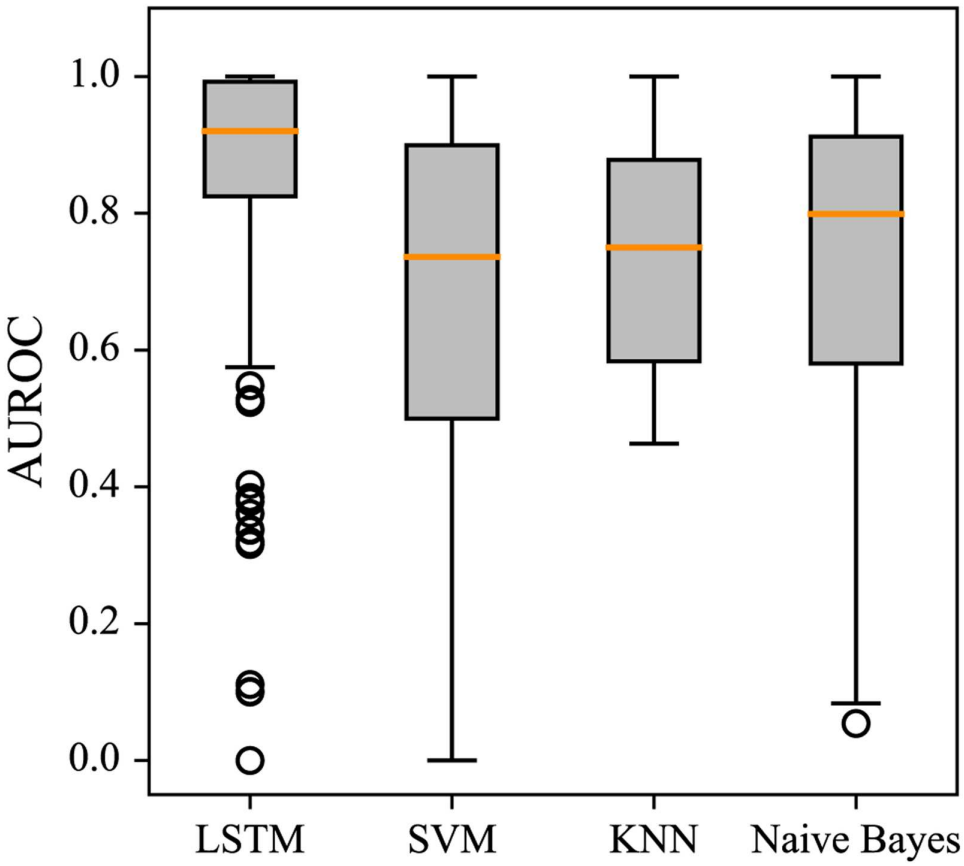

Figure 10. A summary of the area under the receiver operating characteristics (AUROC) of all of the 241-dimensional vectors plotted as box plots for the various models that have been trained. Comparisons are made between the long short-term memory (LSTM) model and the other baseline models: the support vector machine (SVM) model, the *K*-nearest neighbor (KNN) model, and the naive Bayes model.

It may be important to see whether a model generated from data in one set of hospitals is generalizable to a new hospital. In this specific instance, because of the large amount of overlap in the way in which behavior is described by clinicians, the test results appear to improve. This is explored in Appendix C of the Supporting Information.

SA VECTOR USE CASES

In this section, three use cases are described that leverage the power of the generated SAV in psychiatry.

Suicidality Scale and Hospitalization

In this section, a simple suicidality scale is developed based on the elements of SAV that will inform about the propensity of suicidality of an individual patient. The suicidality scale is calculated using the rules mentioned in Table 3. If multiple factors are present at the same visit, then the value of the scale is set to be the maximum of the set obtained for each of the indices set to 1.

Thus the suicidality score is a discrete-value score that takes one of the values 0, 1 2, 3, or 4, corresponding to whether the person has "no suicidal intent," "suicidal ideation,"

Table 3. Steps to be taken for converting the status assessment vector (SAV) to a suicidality score

No.	Steps	Value of scale
1	If all of indices 166, 223, 231, 224, 225, and 226 are set to 0	0
2	If any of the indices 166, 223, or 231 within the SAV is 1	1
3	If the index 224 is set to 1	2
4	If the index 225 is set to 1	3
5	If the index 226 is set to 1	4

"suicidal ideation with intent," "suicidal ideation with plan," or a "suicidal attempt" present in the SAV, respectively.

It is believed that a large part of hospitalizations for patients suffering from major depressive disorder (MDD) is due to either a suicidal attempt or an increased risk of suicide. However, the exact reason for hospitalization is rarely captured, which makes it difficult to directly validate the belief. However, given a suicidality scale, a hypothesis can be tested to see whether the patients tend to have a higher risk of suicide (indicated by a higher value of the suicidality score) before hospitalization as compared to other times. Here the null hypothesis to be tested is as follows:

> H_0: patients suffering from MDD do not exhibit a change in suicidal tendencies just before hospitalization.

The diagnosis of a patient is captured in the database at each visit in a tabular format. This table contains the patient ID, the ICD code (either ICD-9-CM or ICD-10-CM) as entered by the clinician, and an ID for determining the visit number, among other information required for the database. The ICD code is used for selecting patients with a particular diagnosis.

For testing this hypothesis, a cohort of patients who have been diagnosed with MDD[2] and at least one recorded hospitalization event was generated. For these patients, the day of the first hospitalization was identified. The suicidality scale for individual patients was calculated from their SA vectors 1 day (1D), 3 days (3D), 1 week (1W), 2 weeks (2W), 1 month (1M), 2 months (2M), and 6 months (6M) before and after the first hospitalization date.

The average suicidality score has been plotted in Figure 11 for all these times before and after hospitalization. As can be seen, there is an increase in the mean suicidal score 1 day before hospitalization (25% increase from 1 week before hospitalization). Also, immediately after hospitalization, there appears to be a sharp decrease in the suicidality scale (approximately the same as 1 week before).

Although there appears to be an increase in the suicidality scale just before hospitalization, it is important to establish the statistical validity of this increase. To validate the increase, two conditions are checked. In the first instance, the mean value of the suicidality scale of the patient cohort calculated at each time point is compared to the mean suicidality scale 6 months before the hospitalization event. For this, an unpaired t test is performed between

[2] Patients with MDD are identified as individuals who have been diagnosed with one or more of the following ICD codes: 296.20, 296.21, 296.22, 296.23, 296.24, 296.25, 296.26, 296.30, 296.31, 296.32, 296.33, 296.34, 296.35, 296.36, F32.0, F32.1, F32.2, F32.3, F32.4, F32.5, F32.9, F33.0, F33.1, F33.2, F33.3, F33.41, F33.42, F33.9. These include both ICD-09-CM and ICD-10-CM codes.

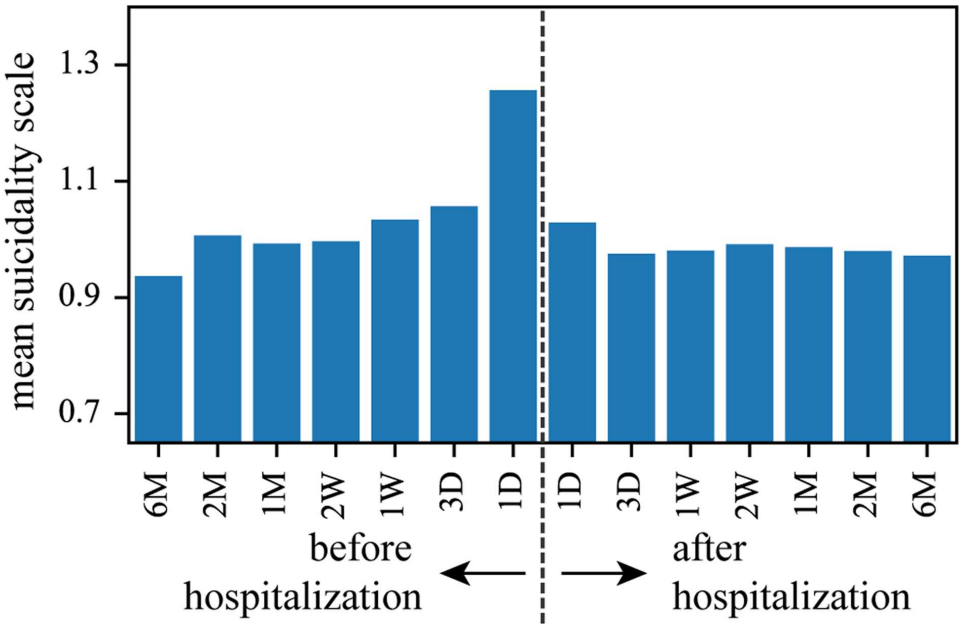

Figure 11. The mean suicidality score of individuals before and after hospitalization for users who have been diagnosed with major depressive disorder (MDD) and have at least a single hospitalization record.

the mean suicidality scale 6 months before the hospitalization event and the mean suicidality scales at all other time points. The p values resulting from this series of t tests are shown in Table 4 as "6M before." In a similar manner, another set of t tests are performed to see whether the mean suicidality scale 1 day before the hospitalization is the same as the mean suicidality scale for all the other days; p values from this set of tests are also tabulated in Table 4. The actual distributions of the scales at the same time points are compared using a Pearson's χ^2 test, and the p values are also tabulated in Table 4.

As can be seen, it is highly unlikely that the distribution with a higher mean suicidality scale observed 1 day before hospitalization comes from the same distribution as that of the other days.

In a similar way, it may be possible to generate the inputs for positive symptoms and negative symptoms directly from the SA vector if measurements of the Positive and Negative Syndrome Scale (PANSS; Kay, Fiszbein, & Opler, 1987) are unavailable.

Clinical Phenotyping

It is well known that multiple mental health disorders share symptomatic characteristics among themselves. For example, the negative symptoms of schizophrenia contain many overlapping symptoms with those of major depression with anhedonia (Chaturvedi, Rao, Mathai, Sarmukaddam, & Gopinath, 1985). Whether a particular symptom should be included in a diagnosis is largely determined by consensus among psychiatric practitioners.[3]

[3] For the *DSM*-5, for example, this would be the *DSM*-5 Task Force, along with the associated work groups, other review bodies, and the APA Board of Trustees.

Table 4. p Values of t tests performed to determine whether the mean suicidality score is the same as that 6 months and 1 day before hospitalization

	p Value of a t test		p Value of a χ^2 test	
Time before/after hospitalization	**6 months before**	**1 day before**	**6 months before**	**1 day before**
6 months before	–	8.60E-32**	–	2.44E-201**
2 months before	2.87E-02	1.25E-21**	1.19e-03	3.34E-157**
1 month before	6.89E-02	5.21E-26**	9.35e-02	1.25E-149**
2 weeks before	5.39E-02	2.37E-25**	7.22e-02	1.3E-147**
1 week before	1.87E-03	8.63E-19**	1.84e-04*	3.91E-90**
3 days before	2.21E-04*	2.35E-14**	1.98e-06*	1.92E-96**
1 day before	8.60E-32**	–	8.38e-50**	–
1 day after	3.67E-04*	1.06E-59**	1.45e-05*	3.03E-155**
3 days after	1.30E-01	8.63E-84**	1.51e-03	1.32E-233**
1 week after	9.01E-02	5.18E-75**	2.08e-04*	6.3E-225**
2 weeks after	3.61E-02	6.64E-66**	2.71e-04*	4.43E-201**
1 month after	5.37E-02	4.59E-66**	5.56e-04*	4.48E-190**
2 months after	9.96E-02	1.62E-66**	6.51e-04*	5.37E-207**
6 months after	1.70E-01	3.18E-67**	1.30e-01	7.06E-175**

Given these current challenges with diagnoses in mental health, it might be interesting to investigate the clinical phenotyping of patients with mental health disorders from a mathematical standpoint. One can approach this problem in a multitude of ways. In this use case, we have used principal component analysis (PCA) of the SA vector to investigate whether it may be possible to tease out differences and similarities at a syndrome level.

A cohort comprising patients who have been diagnosed with a single diagnosis of either MDD or schizophrenia[4] is generated. At each visit, along with the SA, a CGI-S (Busner & Targum, 2007) score is recorded. For each patient, the SA vectors corresponding to CGI-S values more than 3 (i.e., more than moderately ill patients with relatively severe symptoms) have been selected. Once this list of SA vectors is collected, a set of unique SA vectors is obtained from this list, for both schizophrenia and MDD. The SA vectors are transformed using PCA into their orthogonal primary components. This new space is orthogonal and continuous, unlike the space spanned by the SAV, which is discrete and binary. It is much more promising to find phenotype distributions in this continuous space rather than the original binary space spanned by the SAV.

A three-dimensional representation of the distributions of schizophrenia and MDD is shown in Figure 12. It can be seen that there are some regions where the two diagnoses are separate, whereas in others, they overlap. This shows that the SA vector may be used to develop clinical phenotypes and diagnoses with minimal overlap and can advance the diagnosis and treatment of patients with mental health disorders.

[4] Patients diagnosed with schizophrenia are those who have been diagnosed using one or more of the following ICD codes: 295.10, 295.20, 295.30, 295.60, 295.90, and F20.9.

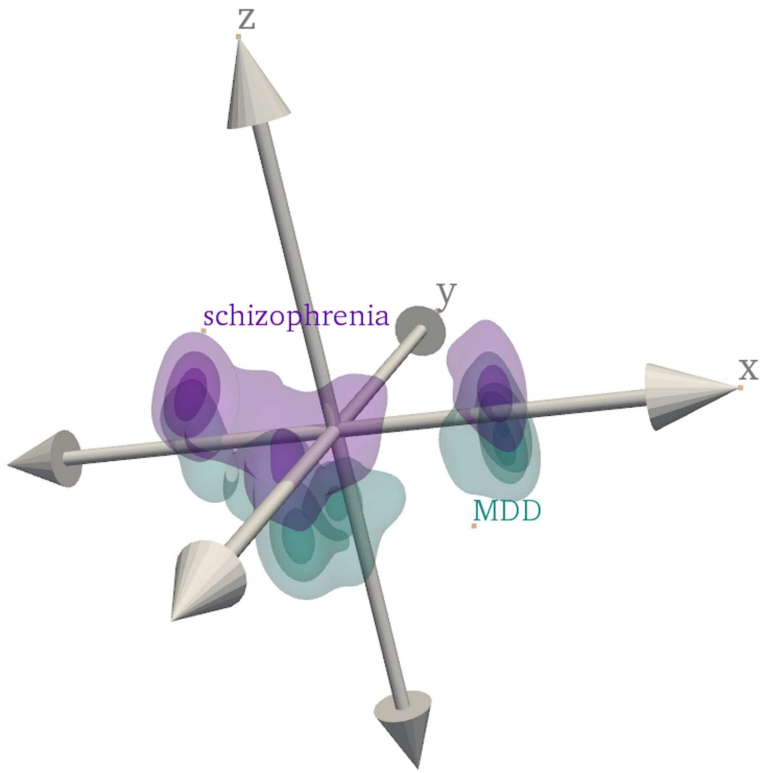

Figure 12. A 3D density plot showing the densities of status assessment (SA) vectors for patients suffering from schizophrenia and major depressive disorder (MDD) after they have been transformed with principal component analysis (PCA). In this plot, the first, second, and fourth dimensions have been plotted along the x-, y-, and z-directions, respectively.

Using the SAV as an Input to ML Model

Finally, we demonstrate a simple use case where SA vectors can be utilized to train a multiclass logistic regression model and generate longitudinal diagnoses for patients. Specifically, probabilities of a selected patient having either schizophrenia, depression, or mania[5] over multiple days are predicted using the model with the patient's longitudinal SA vector as inputs.

SAVs from approximately 200 patients each with schizophrenia, MDD, and mania are selected to train a logistic regression model. A test set is isolated from the cohort using 20% of the data:

$$\mathbf{a}(m) = m \cdot \mathbf{C} + \mathbf{b} \tag{3}$$

$$p_k = \frac{\exp\left(a_k(m)\right)}{\Sigma_{j=1}^{K} \exp(a_j(m))}. \tag{4}$$

The model is trained to minimize the cross-entropy loss between predictions ($\mathbf{p} = [p_1, p_2, p_3]$) and patient diagnoses as recorded by the clinicians. For prediction, the likelihood for each of the three diagnoses is generated based on the patient's SA vector on that day, and the diagnosis with the highest probability is used as the final prediction. Training and testing accuracies during the training process are shown in Figure 13A.

[5] Patients diagnosed with mania are those who have been diagnosed using one or more of the following ICD codes: 296.10, 296.11, 296.12, 296.13, 296.14, 296.15, 296.16, F30.1, F30.10, F30.11, F30.12, F30.13, F30.2, F30.3, F30.4, F30.8, or F30.9.

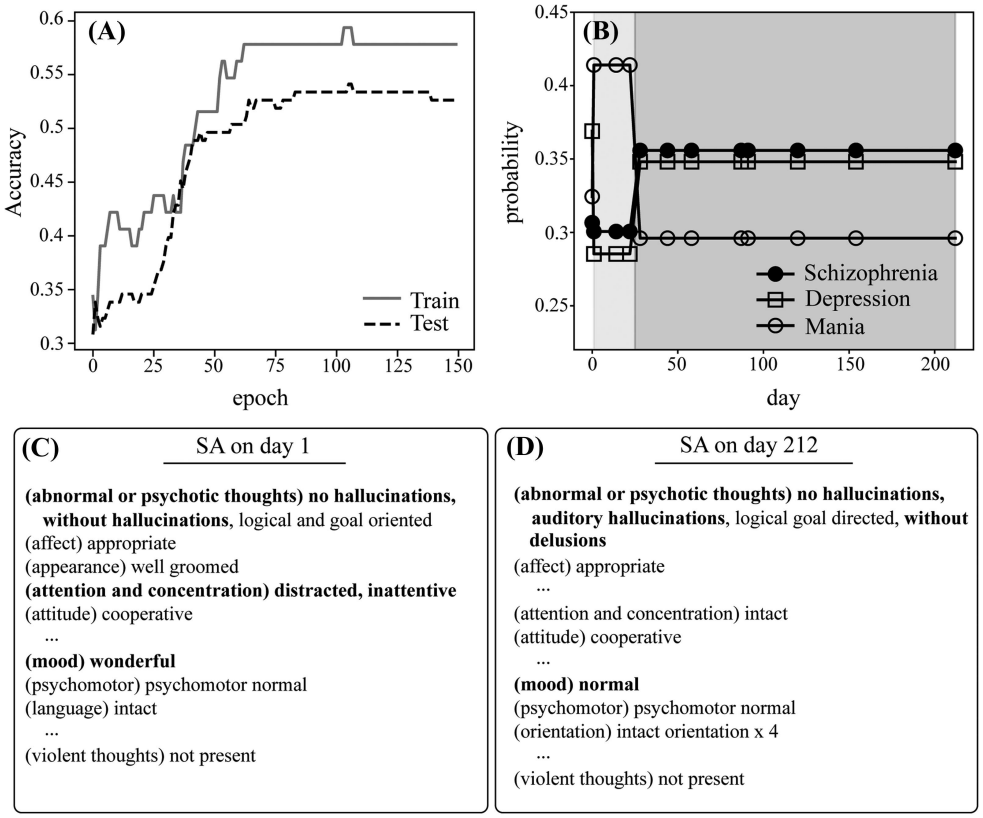

Figure 13. An overview of training a simple machine-learning model for predicting whether a person is suffering from depression, mania, or schizophrenia. A) Training and testing accuracy for the linear model while training on the data over several epochs. A longitudinal symptom profile for a selected patient with personal particulars is shown beside the plot: status assessments (SA) from B) Day 1 in the light gray "manic period" and C) Day 212 in the dark gray "schizophrenic period." Factors are prefixed, with their category names embraced in parentheses, and signs of interest are bolded.

For illustration, the longitudinal data of a patient diagnosed with schizophrenia are used. The patient is an English-speaking, Black, single male who was 43 years old at the start of his treatment. The probabilities of this person suffering from schizophrenia, depression, or mania are plotted over multiple days in Figure 13B. Two diagnostic regions can be observed: one with the highest probability of a manic diagnosis (light gray region) and the other with the highest probability of schizophrenic diagnosis (dark gray region). Despite these discrepancies, the patient was given a diagnosis of schizophrenia throughout. Selected SA signs for the manic (Day 1) and schizophrenic (Day 212) periods are presented in Figure 13C and 13D, respectively. Evident signs (shown in bold text), such as "(abnormal or psychotic thoughts) no hallucinations, without hallucinations," indicate that the patient is less likely to have schizophrenia, while "(attention and concentration) distracted, inattentive" and "(mood) wonderful" suggest symptoms of mania on Day 1. SA signs on Day 1 agree with the model prediction where a high probability of mania is presented. On the other hand, key SA signs (shown in bold text) on Day 212 lead to a diagnosis of mild schizophrenia, consistent with a moderate schizophrenia probability, as predicted by the linear model.

To ascertain whether models are able to learn meaningful clinical information from the given data, the top 10 factors increasing the likelihood of the predictions of schizophrenia,

Table 5. A summary of the status assessment (SA) signs that positively and negatively affect the prediction of diagnosis from a multiclass logistic regression model

SA IDs affecting schizophrenia		SA IDs affecting depression		SA IDs affecting mania	
Positively correlated	**Negatively correlated**	**Positively correlated**	**Negatively correlated**	**Positively correlated**	**Negatively correlated**
(ab. or psy. thou.) ex. delusions/ab. thou. (paranoia)	(ab. or psy. thou.) no issues	(ab. or psy. thou.) no issues	(ab. or psy. thou.) ex. delusions/ab. thou. (paranoia)	(mood) declined	(association) intact/no issues
(affect) blunted/restricted	(sleep) hypersomnia	(appearance) positive grooming/ hygiene	(appearance) issues with grooming/hygiene	(suicidal) suicidal ideation	(affect) no issues/appropriate affect
(ab. or psy. thou.) ex. hallucinations (visual)	(mood) declined	(sensorium) clouded; sedate	(affect) blunted/restricted	(impulse control) limited/some issues	(attitude) positive/appropriate
(homicidal) present	(affect) loss of energy/tiredness	(sleep) neutral, unable to categorize	(ab. or psy. thou.) ex. hallucinations (visual)	(affect) expansive/overly animated	(association) declined
(memory) neutral/unable to categorize	(suicidal) suicidal ideation with means	(association) declined	(association) tangential	(suicidal) suicidal ideation with means	(attention/concentration) issues due to MH
(appearance) issues with grooming/ hygiene	(language) issues with object naming	(sleep) hypersomnia	(sleep) insomnia	(mood) irritable, angry	(violent/homicidal) present
(suicidal) history of ideation	(attention/ concentration) intact/no issues	(mood) labile	(sleep) maintenance issues (can't stay asleep)	(orientation) general orientation issues	(speech) normal, no issues
(executive func.) age appropriate	(suicidal) history of self injury	(executive func.) impaired (due to other MH issue)	(fund of knowledge) above average	(mood) misc. issues	(violent/homicidal) neutral, unable to categorize
(attention/concentration) issues due to MH	(attitude) neutral/unable to categorize	(attention/concentration) intact/no issues	(sensorium) fluctuating	(fund of knowledge) positive vocabulary	(affect) blunted/restricted
(fund of knowledge) above average	(sensorium) neutral; unable to categorize	(appearance) positive behavior	(impulse control) poor/serious issues	(mood) depressed, sad, despondent	(speech) improved

Note. ab. = abnormal. ex. = experiencing. func. = functioning. MH = mental health. psy. = psychotic. thou. = thoughts.

depression, and mania are presented in Table 5. As can be seen, important positive factors contributing to the prediction of schizophrenia are "(abnormal or psychotic thoughts) experiencing delusions/abnormal thoughts (paranoia)" and "(abnormal or psychotic thoughts) experiencing hallucinations (visual)." On the other hand, negative factors indicating no schizophrenic symptom include "(abnormal or psychotic thoughts) no issue," "(mood) declined," and other factors indicative of the absence of psychosis or factors unrelated to the schizophrenic disorder. Principal factors for depression and mania are properly recognized by the model as well.

DISCUSSION

The resulting NLP model is capable of generating SA vectors from categorized free text in the SA into a numerical format. SA vectors can subsequently be used for further analytical and ML tasks for clinical insights. As demonstrated in the previous section, even a simple logistic regression model is able to reasonably reproduce the diagnosis of a patient directly from the SA vector. This NLP model provides a new method for quantifying patient assessments and for analyzing clinical outcomes.

As is evident from the results, some of the categories contain a significant number of unique signs, stemming from the unstructured nature of the input. As one would suspect, patient assessments recorded in the MSE text are meant to be descriptive. Comprehensive annotation of the MSE text for all signs requires significant human effort and may be impractical. This serves as one of the limitations of the overall dataset used and the work presented here. We have augmented the classification of signs using NLP and trained the models incrementally. There is no easy way to increase the validation for some of the categories for them to be comprehensive. This is evident in the fact that the precision, recall, and F_1 scores of several categories (especially in Group 2) are low. Since all deep learning algorithms learn largely by example, the presence of unique descriptors presents a rather difficult challenge for any learning algorithm to generalize to.

Given the evolving nature of psychiatric practice, it is entirely possible that new categories need to be added to the existing categories by clinicians. This may be achieved by user interfaces (UIs) that allow clinicians to augment categories and factors as necessary; hence, over time, new data will be generated that can be used for refining and augmenting existing models. In case the ML classifies free text into an incorrect category, it would also be possible for a clinician, given an intuitive UI, to update the classification. This will not only develop trust between clinicians and the software but also allow the software to collect new data that can be used for refining existing models. As this is typically handled at the software architecture abstraction, it is beyond the scope of the current work.

One of the drawbacks of the SA vector is that the classification results in binary encodings that lack any measurement of severity. For example, although two patients might have "impaired to poor" reasoning abilities, the degree of impairment might be significant. Associating severity to free text, even from a human perspective, is a daunting task and fundamentally subjective in nature. While adding a severity element to the categories will be highly useful, the measures might be unreliable or have significant bias and variability in them. Hence we have chosen to stay with a binary classification that might be less effective but more robust.

The current article has used pretrained word vectors from Google's word2vec algorithms along with the LSTM model with attention for the purposes of classification. Although this is a fairly recent innovation, it is important to know that there are many other alternatives in each aspect of the modeling. For example, word embeddings have been generated using Google's

pretrained word vectors. Alternatives to these are Global Vectors for Word Representation (GloVe; Thomas et al., 2011) from Stanford and fastText (Bojanowski, Grave, Joulin, & Mikolov, 2017) from Facebook's AI Research Lab, among others. All of these have been shown to be marginally better than the others in different studies. However, they are typically all equally viable as a source of pretrained word embeddings under most circumstances. It is also possible to generate word embeddings directly from the given text. Unfortunately, most of the sentences are only a few words long, and embeddings resulting from such sentences typically result in subpar word embeddings and hence have not been attempted.

Pretrained embeddings capture the meanings of words in the most common contexts in which they have been trained. In this particular case, Google's word2vec model has been trained on text obtained from Google News. The context of some of the medical terms within this pretrained vector may not correspond to its colloquial meaning. For example, the words "flooded" or "blunted" in common parlance have a very different meaning than they do in psychiatry. The best way of improving this is to train new word vectors using clinicians' notes. Owing to our current inability to reliably anonymize clinicians' notes, it is difficult to obtain a good dataset on which this can be done. This shall be attempted in the future.

A potential problem of using pretrained word vectors is that it is unable to do word-sense disambiguation. For example, the word "flooded" might be mentioned in two different contexts and thus mean two different things in the same document. Since the meaning of the word "flooded" is different in the two instances, the model ideally should be able to identify to which word it is referring, based on the semantics of the sentence. However, this semantic information is lost in pretrained word vectors. A more recent development in word embeddings is Embeddings From Language Models (Gardner et al., 2018), which generates word embeddings from a multilayer model. Currently, the state of the art in language models are generalized language models, such as GPT (Radford, Narasimhan, Salimans, & Sutskever, 2018), BERT (Devlin et al., 2018; Vaswani et al., 2017), GPT-2 (Devlin et al., 2018; Vaswani et al., 2017), and XLNet (Yang et al., 2019). These modern versions of NLP appear to be much better than RNNs, and in future versions of this implementation, we may attempt to incorporate these language models into the system.

Another possible place where the model might be improved is in named entity recognition (NER; Song, Jo, Park, Kim, & Kim, 2018) and coreference resolution (Stylianou & Vlahavas, 2019). Most of the expressions within the SA typically refer to observations about the patient. Since the SA is free text, it is possible that some clinicians might wish to note down observations about a relative of the patient within the same note. After going through the free text that is present within the SA, we have yet to find such a description. Hence NER and coreference resolution have not been performed on this current dataset. However, if clinician's notes are to be used for NLP and information extraction, then this step would become fundamentally important.

Similarly, unlike typical aspects of natural language, hypothetical sentences are generally not present in the SA, which happens to be a dispassionate record of observations about the patient. Again, if it is important to convert clinicians' notes into a SA vector, it would be important to identify hypothetical sentences and disregard them while generating the SA vector.

The following provides a brief overview of the process that might be used for extracting the same type of information from unstructured notes. Unstructured notes contain a multitude of information about the patient apart from the SA. This includes, and is not limited to, the mental state of the patient that we call the SA; medications, including their doses and

regimens; recordings of stressful and benevolent events in the patient's life; recommendations for other forms of psychotherapy, such as cognitive behavioral therapy; records of recent hospitalizations; concomitant recordings of caregivers; records of scores of psychiatric rating scales, such as the CGI, the Global Assessment of Functioning (GAF), the Montgomery–Åsberg Depression Rating Scale, the PANSS, the Work and Social Adjustment Scale, and the Patient Health Questionnaire–9; and medication side effects and their effects on patients' lives. A detailed exposition of how all metrics might be extracted from clinicians' free text is well beyond the scope of this work. However, it might be possible to extract SA information from within clinicians' notes with the following algorithm:

1. Convert all of the clinicians' notes into sentences.
2. Generate a group of sentences that contain SA information within them and another that does not contain SA information. Furthermore, it would be helpful if the SA information were further classified into different categories.
3. Train a Siamese network (Li et al., 2020) to group sentences into different SA categories.
4. For sentences in each category, request a SME to subcategorize each sentence into its subcomponents (factors).
5. Finally, generate a set of models that will be able to classify which parts of an SAV are associated with the current sentence.

A new clinician's note would subsequently be processed by the following steps:

1. Perform sentence tokenization.
2. Assign sentences to probabilistic categories through the pretrained Siamese network (along with a "NOT SA" class for the sentences that do not have SA information within them).
3. Sentences that do belong to a particular class will subsequently be passed through the pertinent NLP classifier to obtain the NLP-based subclasses.

Again, extracting information from clinicians' notes is a significantly more complex problem than extracting information from categorized free text.

Other than certain factors, such as "history of suicidal ideation," most SAs are assumed to be temporally bound to the time of the visit. It may be that the clinician, while describing the patient, is describing the behavior of the patient at a different time. This has not been explicitly incorporated into the current model and can be the basis for future improvements.

It is instructive to compare our classification method, which follows naturally from text generated by clinicians through the EHR in a practice setting, with classifications present within the RDoC (Cuthbert, 2014) in a deliberative manner. The RDoC divides neurobiological domains into the following major groups: negative valence systems, positive valence systems, cognitive systems, systems for social processes, and arousal/regulatory systems. It aims to elicit and identify physical mechanisms and systems associated with behavior. On the other hand, the SA is a summary of patient behavior.

It is interesting to see that some of the behaviors within the currently classified SA vector can typically be classified into one of the major systems defined within the RDoC. For example, within the system defined by cognitive systems in the RDoC, it is possible to incorporate large portions of the SA vector belonging to the categories "attention and concentration," "level of consciousness," "reasoning," "abnormal or psychotic thoughts," "sensorium," "memory," "language," "speech," "executive functioning," "impulse control," and "psychomotor." Similarly,

Table 6. A summary of systems that are available in the Research Domain Criterion (RDoC) and not present in the status assessment (SA) vector

Category	Subcategory
Systems for social process	"imitation, theory of the mind"; "social dominance"; "facial expression identification"; "attachment/separation fear"
Arousal/regulatory systems	"arousal and regulation"; "resting state activity"
Negative valence systems	"frustrative nonreward"
Positive valence systems	"reward learning"; "habit"

other categories within the SA vector can be incorporated into other categories within the RDoC. Hence many categories and factors of the current SA are present within the RDoC. These specifically belong to the categories shown in Table 6.

On the other hand, certain behaviors within the SA vector are not amenable to incorporation within the categories of the RDoC. Examples of such categories are "homicidal," "suicidal," and "violent thoughts."

In the section "Suicidality Scale and Hospitalization," the suicidality of a patient has been calculated from the SAV. At the end of that section, it was also mentioned that it is possible to create similar scores for the positive and negative symptoms of schizophrenia. In fact, that is a general method that can be used to study how particular aspects of a patient's symptoms are doing over time. In clinical psychiatry, treatment response, remission, and resistance are very important topics (Elkis & Buckley, 2016; Pandarakalam, 2018), and a variety of psychometric scales have been developed (Aboraya et al., 2018) to study them. However, very few are used in real-world settings (Hatfield & Ogles, 2007), where clinicians prefer to track patients through unstructured notes. Without needing to change clinical practice, the current work can convert the notes into SA vectors that can then be mapped to the appropriate scales. This may provide a quantitative way of measuring symptom severity of patients over time. This is especially important when studying the comparative efficacy of different treatment strategies from real-world data. The SAV can also be directly used to help enroll patients in clinical trials, based on the prevalence of inclusion criteria and the absence of exclusion criteria defined for a particular study.

Using methods similar to that shown in the section "Suicidality Scale and Hospitalization," it should be possible to prompt clinicians to use more structured data entry. For example, if a score for mania is generated, a patient presently diagnosed with depression might show signs of negative symptoms of schizophrenia. This can be used for prompting the clinician to perform a rating on the negative symptom rating scale.

The section "Using the SAV as an Input to ML Model" describes a use case wherein one is able to compute the probability that a patient is suffering from mania, depression, or schizophrenia. It is to be noted that the symptoms and SAs overlap significantly across many mental health disorders. Since most of psychiatric practice is based on syndromes, and syndromes overlap to various degrees, diagnosis of a patient becomes rather difficult when the patient's symptoms match multiple syndromes and cannot be classified into a single diagnosis based on the *DSM*-5 categories. For example, the criteria for MDD are based on a person having five out of eight symptoms, as described in the *DSM*-5. Challenges with diagnosis are exacerbated in cases where such syndrome-based classification becomes loosely defined, as in the case of schizophrenia. This is where the SAV can make a significant contribution.

Following the same principles of the use case in the section "Clinical Phenotyping," we can envision a day when the classification of mental health disorders is done in a way to minimize the overlap of symptoms based on the differences in the underlying pathology of the disease. Improvement not only in the classification and identification of diagnosis but also in the subcategorization of patients with the same diagnosis can help guide more effective treatment strategies. As described before, the NLP-based generation of SAV from clinician notes, SA, and so on can enable better targeting of treatment in the real world without requiring significant changes in clinical practice.

CONCLUSION

In this article, a deep learning–based NLP has been used for generating a binary vector representation of the symptoms and functional and emotional states of a psychiatric patient, given the SA of the person. This is the first step in generating enriched longitudinal data of the symptoms and the mental state of patients from EHRs containing records of patients with mental health disorders. This will allow quantitative comparison of outcomes of patients with mental health disorders and, it is hoped, lead to more effective treatment strategies.

AUTHOR CONTRIBUTIONS

Sankha Subhra Mukherjee: Conceptualization: Lead; Writing – original draft: Lead; Writing – review & editing: Lead. Jiawei Yu: Formal analysis: Lead; Software: Lead; Writing – original draft: Supporting; Writing – review & editing: Supporting. Yida Won: Data curation: Lead. Mary Jane McClay: Investigation: Lead. Lu Wang: Investigation: Supporting. Augustus John Rush: Supervision: Supporting; Validation: Lead. Joydeep Sarkar: Project administration: Lead; Resources: Lead.

FUNDING INFORMATION

A. John Rush has received consulting fees from Compass Inc., Curbstone Consultant LLC, Emmes Corp., Holmusk, Johnson and Johnson (Janssen), Liva-Nova, Neurocrine Biosciences Inc., Otsuka-US, Sunovion; speaking fees from Liva-Nova, Johnson and Johnson (Janssen); and royalties from Guilford Press and the University of Texas Southwestern Medical Center, Dallas, TX (for the Inventory of Depressive Symptoms and its derivatives). He is also named co-inventor on two patents: U.S. Patent No. 7,795,033: Methods to Predict the Outcome of Treatment with Antidepressant Medication, Inventors: McMahon FJ, Laje G, Manji H, Rush AJ, Paddock S, Wilson AS; and U.S. Patent No. 7,906,283: Methods to Identify Patients at Risk of Developing Adverse Events During Treatment with Antidepressant Medication, Inventors: McMahon FJ, Laje G, Manji H, Rush AJ, Paddock S.

REFERENCES

Aboraya, A., Nasrallah, H. A., Elswick, D. E., Ahmed, E., Estephan, N., Aboraya, D., . . . Dohar, S. (2018). Measurement-based care in psychiatry—past, present, and future. *Innovations in Clinical Neuroscience*, *15*(11–12), 13–26. Retrieved from https://www.ncbi.nlm.nih.gov/pubmed/30834167 , **PMID:** 30834167, **PMCID:** PMC6380611

Alsentzer, E., Murphy, J. R., Boag, W., Weng, W.-H., Jin, D., Naumann, T., & McDermott, M. B. A. (2019). *Publicly available clinical BERT embeddings*. (arXiv 1904:03323). Retrieved from https://arxiv.org/abs/1904.03323 , **DOI:** https://doi.org/10.18653/v1/W19-1909

American Psychiatric Association. (2013). *Diagnostic and statistical manual of mental disorders* (5th ed.). Washington, DC: Author. **DOI:** https://doi.org/10.1176/appi.books.9780890425596

Bengio, Y., Ducharme, R., Vincent, P., & Janvin, C. (2003). A neural probabilistic language model. *Journal of Machine Learning Research*, *3*, 1137–1155.

Beyer, J., Kuchibhatla, M., Gersing, K., & Krishnan, K. R. (2005). Medical comorbidity in a bipolar outpatient clinical population. *Neuropsychopharmacology*, *30*, 401–404. **DOI:** https://doi.org/10.1038/sj.npp.1300608, **PMID:** 15536492

Bojanowski, P., Grave, E., Joulin, A., & Mikolov, T. (2017). *Enriching word vectors with subword information*. (arXiv 1607.04606). Retrieved from https://arxiv.org/abs/1607.04606

Busner, J., & Targum, S. D. (2007). The clinical global impressions scale: Applying a research tool in clinical practice. *Psychiatry (Edgmont)*, *4*(7), 28–37. Retrieved from https://www.ncbi.nlm.nih.gov/pubmed/20526405, **PMID:** 20526405, **PMCID:** PMC2880930

Chaturvedi, S. K., Rao, G. P., Mathai, P. J., Sarmukaddam, S., & Gopinath, P. S. (1985). Negative symptoms in schizophrenia and depression. *Indian Journal of Psychiatry*, *27*, 237–241. Retrieved from https://www.ncbi.nlm.nih.gov/pubmed/21927110 , **PMID:** 21927110, **PMCID:** PMC3011124

Chen, D., & Manning, C. (2014). *A fast and accurate dependency parser using neural networks*. Doha, Qatar: Association for Computational Linguistics. **DOI:** https://doi.org/10.3115/v1/D14-1082

Chen, Q., Zhu, X., Ling, Z.-H., Wei, S., & Jiang, H. (2016). *Enhancing and combining sequential and tree {LSTM} for natural language inference* (CoRR abs/1609.0). Retrieved from http://arxiv.org/abs/1609.06038

Chiu, J. P. C., & Nichols, E. (2016). Named entity recognition with bidirectional LSTM-CNNs. *Transactions of the Association for Computational Linguistics*, *4*, 357–370. **DOI:** https://doi.org/10.1162/tacl_a_00104

Cuthbert, B. N. (2014). The RDoc framework: Facilitating transition from ICD/DSM to dimensional approaches that integrate neuroscience and psychopathology. *World Psychiatry*, *13*(1), 28–35. **DOI:** https://doi.org/10.1002/wps.20087, **PMID:** 24497240, **PMCID:** PMC3918011

Cuthbert, B. N., & Insel, T. R. (2013). Toward the future of psychiatric diagnosis: The seven pillars of RDoc. *BMC Medicine*, *11*, Article 126. **DOI:** https://doi.org/10.1186/1741-7015-11-126, **PMID:** 23672542, **PMCID:** PMC3653747

Cuthbert, B. N., & Kozak, M. J. (2013). Constructing constructs for psychopathology: The NIMH research domain criteria. *Journal of Abnormal Psychology*, *122*, 928–937. **DOI:** https://doi.org/10.1037/a0034028, **PMID:** 24016027

Devlin, J., Chang, M.-W., Lee, K., & Toutanova, K. (2018). *{BERT:} Pre-training of deep bidirectional transformers for language understanding* (CoRR abs/1810.04805). Retrieved from http://arxiv.org/abs/1810.04805

Dyer, C., Ballesteros, M., Ling, W., Matthews, A., & Smith, N. A. (2015). *Transition-based dependency parsing with stack long short-term memory BT*. Retrieved from https://www.aclweb.org/anthology/P15-1033/, **DOI:** https://doi.org/10.3115/v1/P15-1033

Elkis, H., & Buckley, P. F. (2016). Treatment-resistant schizophrenia. *Psychiatric Clinics of North America*, *39*, 239–265. **DOI:** https://doi.org/10.1016/j.psc.2016.01.006, **PMID:** 27216902

Gardner, M., Grus, J., Neumann, M., Tafjord, O., Dasigi, P., Liu, N. F., . . . Zettlemoyer, L. (2018). *AllenNLP: {A} deep semantic natural language processing platform* (CoRR abs/1803.07640). Retrieved from http://arxiv.org/abs/1803.07640, **DOI:** https://doi.org/10.18653/v1/W18-2501, **PMCID:** PMC5753512

Guy, W. (1976). *ECDEU assessment manual for psychopharmacology*. Rockville, MD: U.S. Department of Health Education, and Welfare. **DOI:** https://doi.org/10.1037/e591322011-001

Hastie, T., Friedman, J., & Tibshirani, R. (2001). Model assessment and selection. In *The elements of statistical learning* (p. 222) New York, NY: Springer. **DOI:** https://doi.org/10.1007/978-0-387-21606-5_7

Hatfield, D. R., & Ogles, B. M. (2007). Why some clinicians use outcome measures and others do not. *Administration and Policy in Mental Health*, *34*, 283–291. **DOI:** https://doi.org/10.1007/s10488-006-0110-y, **PMID:** 17211715

Health Insurance Portability and Accountability Act of 1996, Pub. L. No. 104-191 (1996). https://www.govinfo.gov/content/pkg/PLAW-104publ191/pdf/PLAW-104publ191.pdf

Huang, Z., Xu, W., & Yu, K. (2015). *Bidirectional {LSTM-CRF} models for sequence tagging* (CoRR abs/1508.01991). Retrieved from http://arxiv.org/abs/1508.01991

Huys, Q. J. M., Maia, T. V., & Frank, M. J. (2016). Computational psychiatry as a bridge from neuroscience to clinical applications. *Nature Neuroscience*, *19*, 404–413. **DOI:** https://doi.org/10.1038/nn.4238, **PMID:** 26906507, **PMCID:** PMC5443409

Jackson, R. G., Patel, R., Jayatilleke, N., Kolliakou, A., Ball, M., Gorrell, G., . . . Stewart, R. (2017). Natural language processing to extract symptoms of severe mental illness from clinical text: The Clinical Record Interactive Search Comprehensive Data Extraction (CRIS-CODE) project. *BMJ Open*, *7*(1), e012012. **DOI:** https://doi.org/10.1136/bmjopen-2016-012012, **PMID:** 28096249, **PMCID:** PMC5253558

Jing, J. (2012). *Information extraction from text* (Vol. 10). New York, NY: Springer. **DOI:** https://doi.org/10.1007/978-1-4614-3223-4_2

Johnson, A. E. W., Pollard, T. J., Shen, L., Lehman, L., Feng, M., Ghassemi, M., . . . Mark, R. G. (2016). MIMIC-III, a freely accessible critical care database. *Scientific Data*, *3*, Article 160035. **DOI:** https://doi.org/10.1038/sdata.2016.35, **PMID:** 27219127, **PMCID:** PMC4878278

Kay, S. R., Fiszbein, A., & Opler, L. A. (1987). The Positive and Negative Syndrome Scale (PANSS) for schizophrenia. *Schizophrenia Bulletin*, *13*, 261–276. **DOI:** https://doi.org/10.1093/schbul/13.2.261, **PMID:** 3616518

Koehn, P., Och, F. J., & Marcu, D. (2003). Statistical phrase-based translation. In *Proceedings of the 2003 Human Language Technology Conference of the North American Chapter of the Association for Computational Linguistics* (pp. 54–58). Edmonton, AB: Association for Computational Linguistics. **DOI:** https://doi.org/10.21236/ADA461156

Koita, J., Riggio, S., & Jagoda, A. (2010). The Mental Status Examination in emergency practice. *Emergency Medicine Clinics of North America*, *28*, 439–451. **DOI:** https://doi.org/10.1016/j.emc.2010.03.008, **PMID:** 20709237

Lee, J., Yoon, W., Kim, S., Kim, D., Kim, S., So, C.-H., & Kang, J. (2019). *BioBERT: A pre-trained biomedical language representation model for biomedical text mining* (arXiv 1901:08746). Retrieved from https://arxiv.org/abs/1901.08746, **DOI:** https://doi.org/10.1093/bioinformatics/btz682, **PMID:** 31501885

Li, Z., Lin, H., Zheng, W., Tadesse, M. M., Yang, Z., & Wang, J. (2020). Interactive self-attentive siamese network for biomedical sentence similarity. *IEEE Access*, *8*, 84093–84104. **DOI:** https://doi.org/10.1109/ACCESS.2020.2985685

Lipton, Z. C., Berkowitz, J., & Elkan, C. (2015). *A critical review of recurrent neural networks for sequence learning* (arXiv 1506.00019). Retrieved from https://arxiv.org/abs/1506.00019

Liu, X., Shen, Y., Duh, K., & Gao, J. (2017). *Stochastic answer networks for machine reading comprehension* (CoRR abs/1712.0). Retrieved from http://arxiv.org/abs/1712.03556

Marshall, M. (2020). The hidden links between mental disorders. *Nature*, *581*, 19–21. **DOI:** https://doi.org/10.1038/d41586-020-00922-8, **PMID:** 32372044

Martin, D. C. (1990). The mental status examination. In H. K. Walker, W. D. Hall, & J. W. Hurst, (Eds.), *Clinical methods: The history, physical, and laboratory examinations* (3rd ed., pp. 924–929). Boston, MA: Butterworth.

Martín Abadi, A. A., Barham, P., Brevdo, E., Chen, Z., Citro, C., Corrado, G. S., . . . Zheng, X. (2016). *Tensorflow: Large-scale machine learning on heterogeneous distributed systems* (arXiv 1603:04467). Retrieved from https://arxiv.org/abs/1603.04467

Mikolov, T., Sutskever, I., Chen, K., Corrado, G., & Dean, J. (2013). *Distributed representations of words and phrases and their compositionality* (CoRR abs/1310.4546). Retrieved from https://papers.nips.cc/paper/2013/file/9aa42b31882ec039965f3c4923ce901b-Paper.pdf

Pandarakalam, J. P. (2018). Challenges of treatment-resistant depression. *Psychiatria Danubina*, *30*, 273–284. **DOI:** https://doi.org/10.24869/psyd.2018.273, **PMID:** 30267518

Passos, A., Kumar, V., & McCallum, A. (2014). *Lexicon infused phrase embeddings for named entity resolution* (arXiv 1404.5367). Retrieved from https://arxiv.org/abs/1404.5367, **DOI:** https://doi.org/10.3115/v1/W14-1609

Pham, T., Tran, T., Phung, D., & Venkatesh, S. (2017). Predicting healthcare trajectories from medical records: A deep learning approach. *Journal of Biomedical Informatics*, *69*, 218–229. **DOI:** https://doi.org/10.1016/j.jbi.2017.04.001, **PMID:** 28410981

Radford, A., Narasimhan, K., Salimans, T., & Sutskever, I. (2018). *Improving language understanding by generative pre-training*. Retrieved from https://www.semanticscholar.org/paper/Improving-Language-Understanding-by-Generative-Radford/cd18800a0fe0b668a1cc19f2ec95b5003d0a5035

Radford, A., Wu, J., Child, R., Luan, D., Dario, A., & Sutskever, I. (2019). *Language models are unsupervised multitask learners*. Retrieved from https://d4mucfpksywv.cloudfront.net/better-language-models/language-models.pdf

Rumshisky, A., Ghassemi, M., Naumann, T., Szolovits, P., Castro, V. M., McCoy, T. H., & Perlis, R. H. (2016). Predicting early psychiatric readmission with natural language processing of narrative discharge summaries. *Translational Psychiatry*, *6*(10), e921. **DOI:** https://doi.org/10.1038/tp.2015.182, **PMID:** 27754482, **PMCID:** PMC5315537

Song, H. J., Jo, B. C., Park, C. Y., Kim, J. D., & Kim, Y. S. (2018). Comparison of named entity recognition methodologies in biomedical documents. *Biomedical Engineering Online*, *17*(Suppl. 2), 158. **DOI:** https://doi.org/10.1186/s12938-018-0573-6, **PMID:** 30396340, **PMCID:** PMC6219049

Stylianou, N., & Vlahavas, I. (2019). *A neural entity coreference resolution review* (arXiv 1910:09329). Retrieved from https://arxiv.org/abs/1910.09329

Thomas, D. M., Martin, C. K., Heymsfield, S., Redman, L. M., Schoeller, D. A., & Levine, J. A. (2011). A simple model predicting individual weight change in humans. *Journal of Biological Dynamics*, *5*, 579–599. **DOI:** https://doi.org/10.1080/17513758.2010.508541, **PMID:** 24707319, **PMCID:** PMC3975626

Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., . . . Polosukhin, I. (2017). *Attention is all you need* (CoRR abs/1706.03762). Retrieved from http://arxiv.org/abs/1706.03762

World Health Organization. (1993). *The ICD-10 classification of mental and behavioural disorders: Diagnostic criteria for research*. Geneva, Switzerland: World Health Organization.

Wray, N. R., Ripke, S., Mattheisen, M., Trzaskowski, M., Byrne, E. M., Abdellaoui, A., . . . Major Depressive Disorder Working Group of the Psychiatric Genomics. (2018). Genome-wide association analyses identify 44 risk variants and refine the genetic architecture of major depression. *Nature Genetics*, *50*, 668–681. **DOI:** https://doi.org/10.1038/s41588-018-0090-3, **PMID:** 29700475, **PMCID:** PMC5934326

Wunnava, S., Qin, X., Kakar, T., Rundensteiner, E. A., & Kong, X. (2018). *Bidirectional LSTM-CRF for adverse drug event tagging in electronic health records*. Paper presented at the 1st International Workshop on Medication and Adverse Drug Event Detection Massachusetts.

Xu, B., Shi, X., Zhao, Z., & Zheng, W. (2018). Leveraging biomedical resources in Bi-LSTM for drug–drug interaction extraction. *IEEE Access*, *6*, 33432–33439. **DOI:** https://doi.org/10.1109/ACCESS.2018.2845840

Yang, Z., Dai, Z., Yang, Y., Carbonell, J. G., Salakhutdinov, R., & Le, Q. V. (2019). *XLNet: Generalized autoregressive pretraining for language understanding* (CoRR abs/1906.08237). Retrieved from http://arxiv.org/abs/1906.08237

Zhou, H., Zhang, Y., Cheng, C., Huang, S., Dai, X., & Chen, J. (2017). A neural probabilistic structured-prediction method for transition-based natural language processing. *Journal of Artificial Intelligence Research*, *58*, 703–729. **DOI:** https://doi.org/10.1613/jair.5259

Zhu, M., Zhang, Y., Chen, W., Zhang, M., & Zhu, J. (2013). *Fast and accurate shift-reduce constituent parsing*. Paper presented at the annual meeting of the Association for Computational Linguistics, Sofia, Bulgaria.