

Analysis of Risk Factor Domains in Psychosis Patient Health Records

Eben Holderness^{1,2}, Nicholas Miller^{1,2}, Philip Cawkwell¹, Kirsten Bolton¹,
James Pustejovsky², Marie Meteer² and Mei-Hua Hall¹

¹Psychosis Neurobiology Laboratory, McLean Hospital, Harvard Medical School

²Department of Computer Science, Brandeis University

{eholderness, mhall}@mclean.harvard.edu
nicholas.anthony.miller@gmail.com
{pcawkwell, kbolton}@partners.org
{jamesp, mmeteer}@cs.brandeis.edu

Abstract

Readmission after discharge from a hospital is disruptive and costly, regardless of the reason. However, it can be particularly problematic for psychiatric patients, so predicting which patients may be readmitted is critically important but also very difficult. Clinical narratives in psychiatric electronic health records (EHRs) span a wide range of topics and vocabulary; therefore, a psychiatric readmission prediction model must begin with a robust and interpretable topic extraction component. We created a data pipeline for using document vector similarity metrics to perform topic extraction on psychiatric EHR data in service of our long-term goal of creating a readmission risk classifier. We show initial results for our topic extraction model and identify additional features we will be incorporating in the future.

1 Introduction

Psychotic disorders typically emerge in late adolescence or early adulthood (Kessler et al., 2007; Thomsen, 1996) and affect approximately 2.5-4% of the population (Perälä et al., 2007; Bogren et al., 2009), making them one of the leading causes of disability worldwide (Vos et al., 2015). A substantial proportion of psychiatric inpatients are readmitted after discharge (Wiersma et al., 1998). Readmissions are disruptive for patients and families, and are a key driver of rising healthcare costs (Mangalore and Knapp, 2007; Wu et al., 2005). Reducing readmission risk is therefore a major unmet need of psychiatric care. Developing clinically implementable machine learning tools to enable accurate assessment of risk factors associated with readmission offers opportunities to inform the selection of treatment interventions and implement appropriate preventive measures.

In psychiatry, traditional strategies to study readmission risk factors rely on clinical observa-

tion and manual retrospective chart review (Olfson et al., 1999; Lorine et al., 2015). This approach, although benefitting from clinical expertise, does not scale well for large data sets, is effort-intensive, and lacks automation. An efficient, more robust, and cheaper NLP-based alternative approach has been developed and met with some success in other medical fields (Murff et al., 2011). However, this approach has seldom been applied in psychiatry because of the unique characteristics of psychiatric medical record content.

There are several challenges for topic extraction when dealing with clinical narratives in psychiatric EHRs. First, the vocabulary used is highly varied and context-sensitive. A patient may report “feeling ‘really great and excited’” – symptoms of mania – without any explicit mention of keywords that differ from everyday vocabulary. Also, many technical terms in clinical narratives are multiword expressions (MWEs) such as ‘obsessive body image’, ‘linear thinking’, ‘short attention span’, or ‘panic attack’. These phrasemes are comprised of words that in isolation do not impart much information in determining relatedness to a given topic but do in the context of the expression.

Second, the narrative structure in psychiatric clinical narratives varies considerably in how the same phenomenon can be described. Hallucinations, for example, could be described as “the patient reports auditory hallucinations,” or “the patient has been hearing voices for several months,” amongst many other possibilities.

Third, phenomena can be directly mentioned without necessarily being relevant to the patient specifically. Psychosis patient discharge summaries, for instance, can include future treatment plans (e.g. “Prevent relapse of a manic or major depressive episode.”, “Prevent recurrence of psychosis.”) containing vocabulary that at the word-level seem strongly correlated with readmission

risk. Yet at the paragraph-level these do not indicate the presence of a readmission risk factor in the patient and in fact indicate the absence of a risk factor that was formerly present.

Lastly, given the complexity of phenotypic assessment in psychiatric illnesses, patients with psychosis exhibit considerable differences in terms of illness and symptom presentation. The constellation of symptoms leads to various diagnoses and comorbidities that can change over time, including schizophrenia, schizoaffective disorder, bipolar disorder with psychosis, and substance use induced psychosis. Thus, the lexicon of words and phrases used in EHRs differs not only across diagnoses but also across patients and time.

Taken together, these factors make topic extraction a difficult task that cannot be accomplished by keyword search or other simple text-mining techniques.

To identify specific risk factors to focus on, we not only reviewed clinical literature of risk factors associated with readmission (Alvarez-Jimenez et al., 2012; Addington et al., 2010), but also considered research related to functional remission (Harvey and Bellack, 2009), forensic risk factors (Singh and Fazel, 2010), and consulted clinicians involved with this project. Seven risk factor domains – Appearance, Mood, Interpersonal, Occupation, Thought Content, Thought Process, and Substance – were chosen because they are clinically relevant, consistent with literature, replicable across data sets, explainable, and implementable in NLP algorithms.

In our present study, we evaluate multiple approaches to automatically identify which risk factor domains are associated with which paragraphs in psychotic patient EHRs.¹ We perform this study in support of our long-term goal of creating a readmission risk classifier that can aid clinicians in targeting individual treatment interventions and assessing patient risk of harm (e.g. suicide risk, homicidal risk). Unlike other contemporary approaches in machine learning, we intend to create a model that is clinically explainable and flexible across training data while maintaining consistent performance.

To incorporate clinical expertise in the identification of risk factor domains, we undertake an annotation project, detailed in section 3.1. We identify a test set of over 1,600 EHR paragraphs

which a team of three domain-expert clinicians annotate paragraph-by-paragraph for relevant risk factor domains. Section 3.2 describes the results of this annotation task. We then use the gold standard from the annotation project to assess the performance of multiple neural classification models trained exclusively on Term Frequency – Inverse Document Frequency (TF-IDF) vectorized EHR data, described in section 4. To further improve the performance of our model, we incorporate domain-relevant MWEs identified using all in-house data.

2 Related Work

McCoy et al. (2015) constructed a corpus of web data based on the Research Domain Criteria (RDoC)(Insel et al., 2010), and used this corpus to create a vector space document similarity model for topic extraction. They found that the ‘negative valence’ and ‘social’ RDoC domains were associated with readmission. Using web data (in this case data retrieved from the Bing API) to train a similarity model for EHR texts is problematic since it differs from the target data in both structure and content. Based on reconstruction of the procedure, we conclude that many of the informative MWEs critical to understanding the topics of paragraphs in EHRs are not captured in the web data. Additionally, RDoC is by design a generalized research construct to describe the entire spectrum of mental disorders and does not include domains that are based on observation or causes of symptoms. Important indicators within EHRs of patient health, like appearance or occupation, are not included in the RDoC constructs.

Rumshisky et al. (2016) used a corpus of EHRs from patients with a primary diagnosis of major depressive disorder to create a 75-topic LDA topic model that they then used in a readmission prediction classifier pipeline. Like with McCoy et al. (2015), the data used to train the LDA model was not ideal as the generalizability of the data was narrow, focusing on only one disorder. Their model achieved readmission prediction performance with an area under the curve of .784 compared to a baseline of .618. To perform clinical validation of the topics derived from the LDA model, they manually evaluated and annotated the topics, identifying the most informative vocabulary for the top ten topics. With their training data, they found the strongest coherence occurred

¹This study has received IRB approval.

in topics involving substance use, suicidality, and anxiety disorders. But given the unsupervised nature of the LDA clustering algorithm, the topic coherence they observed is not guaranteed across data sets.

3 Data

Our target data set consists of a corpus of discharge summaries, admission notes, individual encounter notes, and other clinical notes from 220 patients in the OnTrack™ program at McLean Hospital. OnTrack™ is an outpatient program, focusing on treating adults ages 18 to 30 who are experiencing their first episodes of psychosis. The length of time in the program varies depending on patient improvement and insurance coverage, with an average of two to three years. The program focuses primarily on early intervention via individual therapy, group therapy, medication evaluation, and medication management. See Table 1 for a demographic breakdown of the 220 patients, for which we have so far extracted approximately 240,000 total EHR paragraphs spanning from 2011 to 2014 using Meditech, the software employed by McLean for storing and organizing EHR data.

These patients are part of a larger research cohort of approximately 1,800 psychosis patients, which will allow us to connect the results of this EHR study with other ongoing research studies incorporating genetic, cognitive, neurobiological, and functional outcome data from this cohort.

We also use an additional data set for training our vector space model, comprised of EHR texts queried from the Research Patient Data Registry (RPDR), a centralized regional data repository of clinical data from all institutions in the Partners HealthCare network. These records are highly comparable in style and vocabulary to our target data set. The corpus consists of discharge summaries, encounter notes, and visit notes from approximately 30,000 patients admitted to the system’s hospitals with psychiatric diagnoses and symptoms. This breadth of data captures a wide range of clinical narratives, creating a comprehensive foundation for topic extraction.

After using the RPDR query tool to extract EHR paragraphs from the RPDR database, we created a training corpus by categorizing the extracted para-

Mean Age (2014)	20.7
Gender (Male)	79%
Race	
Asian	6%
Black	7%
Caucasian	77%
Latino	5%
Multiracial	5%
Insurance (Public)²	5.5%
30-day Inpatient Readmission Rate	14%

Table 1: Demographic breakdown of the target cohort.

graphs according to their risk factor domain using a lexicon of 120 keywords that were identified by the clinicians involved in this project. Certain domains – particularly those involving thoughts and other abstract concepts – are often identifiable by MWEs rather than single words. The same clinicians who identified the keywords manually examined the bigrams and trigrams with the highest TF-IDF scores for each domain in the categorized paragraphs, identifying those which are conceptually related to the given domain. We then used this lexicon of 775 keyphrases to identify more relevant training paragraphs in RPDR and treat them as (non-stemmed) unigrams when generating the matrix. By converting MWEs such as ‘shortened attention span’, ‘unusual motor activity’, ‘wide-ranging affect’, or ‘linear thinking’ to non-stemmed unigrams, the TF-IDF score (and therefore the predictive value) of these terms is magnified. In total, we constructed a corpus of roughly 100,000 paragraphs consisting of 7,000,000 tokens for training our model.

3.1 Annotation Task

In order to evaluate our models, we annotated 1,654 paragraphs selected from the 240,000 paragraphs extracted from Meditech with the clinically relevant domains described in Table 2. The annotation task was completed by three licensed clinicians. All paragraphs were removed from the surrounding EHR context to ensure annotators were not influenced by the additional contextual information. Our domain classification models consider each paragraph independently and thus we designed the annotation task to mirror the information available to the models.

The annotators were instructed to label each

²The vast majority of patients in our target cohort are dependents on a parental private health insurance plan.

Domain	Description	Example Paragraph	Example Keywords
Appearance	Physical appearance, gestures, and mannerisms	“A well-appearing, clean young woman appearing her stated age, pleasant and cooperative. Eye contact was good.”	disheveled, clothing, groomed, wearing, clean
Thought Content	Suicidal/homicidal ideation, obsessions, phobias, delusions, hallucinations	“No SI ³ , No HI ⁴ , No hallucinations, Ideas of reference, Paranoid delusions”	obsession, delusion, grandiose, ideation, suicidal, paranoid
Interpersonal	Family situation, friendships, and other social relationships	“Pt. overall appears to be functioning very well despite this conflict with a romantic interest of hers.”	boyfriend, relationship, peers, family, parents, social
Mood	Feelings and overall disposition	“Pt. indicates that his mood is becoming more ‘depressed.’”	anxious, calm, depressed, labile, confused, cooperative
Occupation	School and/or employment	“Pt. followed through with decision to leave college at this point in time.”	boss, employed, job, school, class, homework, work
Thought Process	Pace and coherence of thoughts. Includes linear, goal-directed, perseverative, tangential, and flight of ideas	“Disorganized (Difficult to communicate with patient.), Paucity of thought, Thought-blocking.”	linear, tangential, prosody, blocking, goal-directed, perseverant
Substance	Drug and/or alcohol use	“Patient used marijuana once which he believes triggered the current episode.”	cocaine, marijuana, ETOH ⁵ , addiction, narcotic
Other	Any paragraph that does not fall into any of the other seven domains	“Maintain mood stabilization, prevent future episodes of mania, improve self-monitoring skills.”	—

Table 2: Annotation scheme for the domain classification task.

paragraph with one or more of the seven risk factor domains. In instances where more than one domain was applicable, annotators assigned the domains in order of prevalence within the paragraph. An eighth label, ‘Other’, was included if a paragraph was ambiguous, uninterpretable, or about a domain not included in the seven risk factor domains (e.g. non-psychiatric medical concerns and lab results). The annotations were then reviewed by a team of two clinicians who adjudicated collaboratively to create a gold standard. The gold standard and the clinician-identified keywords and MWEs have received IRB approval for release to the community. They are available as supplementary data to this paper.

3.2 Inter-Annotator Agreement

Inter-annotator agreement (IAA) was assessed using a combination of Fleiss’s Kappa (a variant of Scott’s Pi that measures pairwise agreement for annotation tasks involving more than two annotators) (Fleiss, 1971) and Cohen’s Multi-Kappa as proposed by Davies and Fleiss (1982). Table 3 shows IAA calculations for both overall agreement and agreement on the first (most important) domain only. Following adjudication, accuracy scores were calculated for each annotator by evaluating their annotations against the gold standard.

Overall agreement was generally good and

aligned almost exactly with the IAA on the first domain only. Out of the 1,654 annotated paragraphs, 671 (41%) had total agreement across all three annotators. We defined total agreement for the task as a set-theoretic complete intersection of domains for a paragraph identified by all annotators. 98% of paragraphs in total agreement involved one domain. Only 35 paragraphs had total disagreement, which we defined as a set-theoretic null intersection between the three annotators. An analysis of the 35 paragraphs with total disagreement showed that nearly 30% included the term “blunted/restricted”. In clinical terminology, these terms can be used to refer to appearance, affect, mood, or emotion. Because the paragraphs being annotated were extracted from larger clinical narratives and examined independently of any surrounding context, it was difficult for the annotators to determine the most appropriate domain. This lack of contextual information resulted in each annotator using a different ‘default’ label: Appearance, Mood, and Other. During adjudication, Other was decided as the most appropriate label unless the paragraph contained additional content that encompassed other domains, as it avoids making unnecessary assumptions.

³Suicidal ideation

⁴Homicidal ideation

⁵Ethyl alcohol and ethanol

Labels	Fleiss's Kappa	Cohen's Multi-Kappa	Mean Accuracy
Overall	0.575	0.571	0.746
First Domain Only	0.536	0.528	0.805

Table 3: Inter-annotator agreement

Network	MLP	RBF
Input Layer		
Nodes	100	100
Dropout	0.2	0.2
Activation	ReLU ⁶	ReLU
Hidden Layer		
Nodes	100	350
Dropout	0.5	0.0
Activation	ReLU	RBF
Output Layer		
Nodes	7	7
Activation	Sigmoid	Linear
Optimizer	Adam ⁷	Adam
Loss Function	Categorical Cross Entropy	Mean Squared Error
Training Epochs	30	50
Batch Size	128	128

Table 4: Architectures of our highest-performing MLP and RBF networks.

A Fleiss's Kappa of 0.575 lies on the boundary between ‘Moderate’ and ‘Substantial’ agreement as proposed by Landis and Koch (1977). This is a promising indication that our risk factor domains are adequately defined by our present guidelines and can be employed by clinicians involved in similar work at other institutions.

The fourth column in Table 3, Mean Accuracy, was calculated by averaging the three annotator accuracies as evaluated against the gold standard. This provides us with an informative baseline of human parity on the domain classification task.

4 Topic Extraction

Figure 1 illustrates the data pipeline for generating our training and testing corpora, and applying them to our classification models.

⁶Rectified Linear Units, $f(x) = \max(0, x)$ (Nair and Hinton, 2010)

⁷Adaptive Moment Estimation (Kingma and Ba, 2014)

We use the TfidfVectorizer tool included in the scikit-learn machine learning toolkit (Pedregosa et al., 2011) to generate our TF-IDF vector space models, stemming tokens with the Porter Stemmer tool provided by the NLTK library (Bird et al., 2009), and calculating TF-IDF scores for unigrams, bigrams, and trigrams. Applying Singular Value Decomposition (SVD) to the TF-IDF matrix, we reduce the vector space to 100 dimensions, which Zhang et al. (2011) found to improve classifier performance.

Starting with the approach taken by McCoy et al. (2015), who used aggregate cosine similarity scores to compute domain similarity directly from their TF-IDF vector space model, we extend this method by training a suite of three-layer multilayer perceptron (MLP) and radial basis function (RBF) neural networks using a variety of parameters to compare performance. We employ the Keras deep learning library (Chollet et al., 2015) using a TensorFlow backend (Abadi et al.) for this task. The architectures of our highest performing MLP and RBF models are summarized in Table 4. Prototype vectors for the nodes in the hidden layer of our RBF model are selected via k-means clustering (MacQueen et al., 1967) on each domain paragraph megadocument individually. The RBF transfer function for each hidden layer node is assigned the same width, which is based off the maximum Euclidean distance between the centroids that were computed using k-means.

To prevent overfitting to the training data, we utilize a dropout rate (Srivastava et al., 2014) of 0.2 on the input layer of all models and 0.5 on the MLP hidden layer.

Since our classification problem is multiclass, multilabel, and open-world, we employ seven nodes with sigmoid activations in the output layer, one for each risk factor domain. This allows us to identify paragraphs that fall into more than one of the seven domains, as well as determine paragraphs that should be classified as Other. Unlike the traditionally used softmax activation function, which is ideal for single-label, closed-world classification tasks, sigmoid nodes output class like-

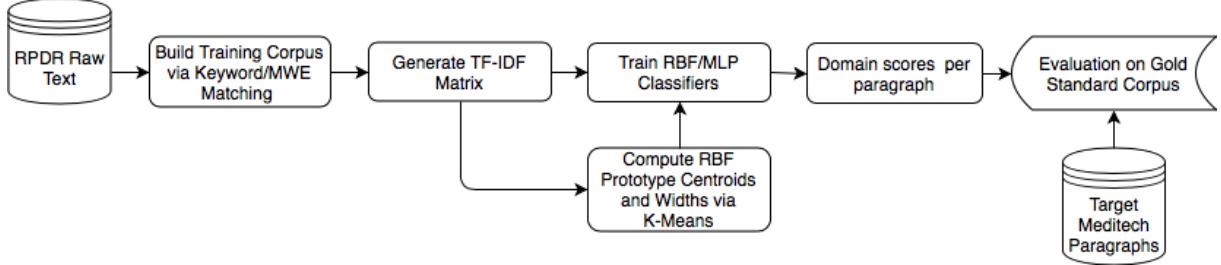


Figure 1: Data pipeline for training and evaluating our risk factor domain classifiers.

lihoods for each node independently without the normalization across all classes that occurs in softmax.

We find that the risk factor domains vary in the degree of homogeneity of language used, and as such certain domains produce higher similarity scores, on average, than others. To account for this, we calculate threshold similarity scores for each domain using the formula $\min = \text{avg(sim)} + \alpha * \sigma(\text{sim})$, where σ is standard deviation and α is a constant, which we set to 0.78 for our MLP model and 1.2 for our RBF model through trial-and-error. Employing a generalized formula as opposed to manually identifying threshold similarity scores for each domain has the advantage of flexibility in regards to the target data, which may vary in average similarity scores depending on its similarity to the training data. If a paragraph does not meet threshold on any domain, it is classified as Other.

5 Results and Discussion

Table 5 shows the performance of our models on classifying the paragraphs in our gold standard. To assess relative performance of feature representations, we also include performance metrics of our models without MWEs. Because this is a multilabel classification task we use macro-averaging to compute precision, recall, and F1 scores for each paragraph in the testing set. In identifying domains individually, our models achieved the highest per-domain scores on Substance ($F1 \approx 0.8$) and the lowest scores on Interpersonal and Mood ($F1 \approx 0.5$). We observe a consistency in per-domain performance rankings between our MLP and RBF models.

The wide variance in per-domain performance is due to a number of factors. Most notably, the training examples we extracted from RPDR – while very comparable to our target OnTrack™

	Precision	Recall	F1
Aggregate Cosine Similarity Scores	0.602	0.563	0.574
MLP Baseline (No MWEs)	0.611	0.567	0.579
RBF Baseline (No MWEs)	0.603	0.618	0.606
MLP (w/ MWEs)	0.717	0.666	0.681
Appearance	0.886	0.414	0.564
Interpersonal	0.548	0.453	0.496
Mood	0.691	0.430	0.530
Occupation	0.826	0.461	0.592
Substance	0.920	0.703	0.797
Thought Content	0.926	0.590	0.721
Thought Process	0.654	0.617	0.635
Other	0.632	0.798	0.710
RBF (w/ MWEs)	0.684	0.630	0.645
Appearance	0.670	0.490	0.566
Interpersonal	0.410	0.493	0.448
Mood	0.655	0.399	0.496
Occupation	0.720	0.501	0.598
Substance	0.866	0.730	0.792
Thought Content	0.892	0.547	0.678
Thought Process	0.569	0.691	0.624
Other	0.651	0.650	0.651

Table 5: Overall and domain-specific Precision, Recall, and F1 scores for our models. The first row computes similarity directly from the TF-IDF matrix, as in (McCoy et al., 2015). All other rows are classifier outputs.

data – may not have an adequate variety of content and range of vocabulary. Although using keyword and MWE matching to create our training corpus has the advantage of being significantly less labor intensive than manually labeling every paragraph in the corpus, it is likely that the homogeneity of language used in the training para-

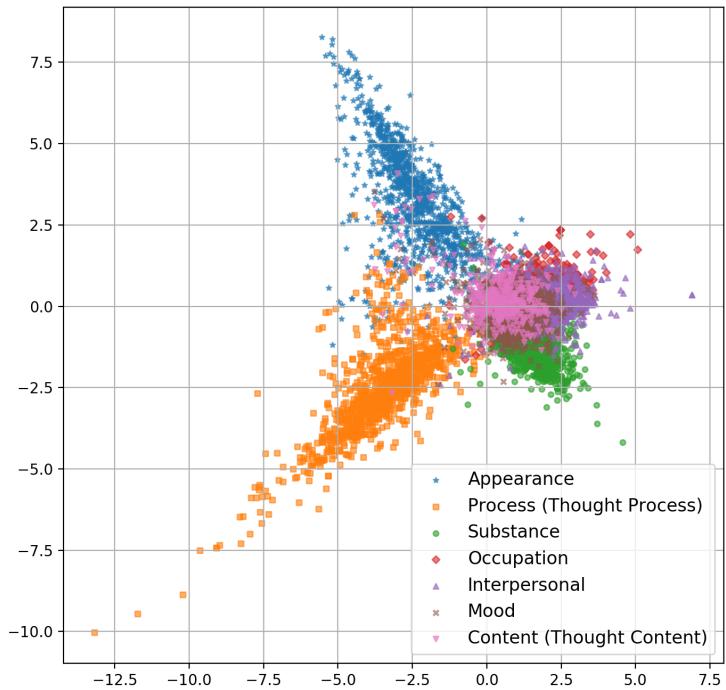


Figure 2: 2-component linear discriminant analysis of the RPDR training data.

graphs is higher than it would be otherwise. Additionally, all of the paragraphs in the training data are assigned exactly one risk factor domain even if they actually involve multiple risk factor domains, making the clustering behavior of the paragraphs more difficult to define. Figure 2 illustrates the distribution of paragraphs in vector space using 2-component Linear Discriminant Analysis (LDA) (Johnson and Wichern, 2004).

Despite prior research indicating that similar classification tasks to ours are more effectively performed by RBF networks (Scheirer et al., 2014; Jain et al., 2014; Bendale and Boult, 2015), we find that a MLP network performs marginally better with significantly less preprocessing (i.e. k-means and width calculations) involved. We can see in Figure 2 that Thought Process, Appearance, Substance, and – to a certain extent – Occupation clearly occupy specific regions, whereas Interpersonal, Mood, and Thought Content occupy the same noisy region where multiple domains overlap. Given that similarity is computed using Euclidean distance in an RBF network, it is difficult to accurately classify paragraphs that fall in

regions occupied by multiple risk factor domain clusters since prototype centroids from the risk factor domains will overlap and be less differentiable. This is confirmed by the results in Table 5, where the differences in performance between the RBF and MLP models are more pronounced in the three overlapping domains (0.496 vs 0.448 for Interpersonal, 0.530 vs 0.496 for Mood, and 0.721 vs 0.678 for Thought Content) compared to the non-overlapping domains (0.564 vs 0.566 for Appearance, 0.592 vs 0.598 for Occupation, 0.797 vs 0.792 for Substance, and 0.635 vs 0.624 for Thought Process). We also observe a similarity in the words and phrases with the highest TF-IDF scores across the overlapping domains: many of the Thought Content words and phrases with the highest TF-IDF scores involve interpersonal relations (e.g. ‘fear surrounding daughter’, ‘father’, ‘family history’, ‘familial conflict’) and there is a high degree of similarity between high-scoring words for Mood (e.g. ‘meets anxiety criteria’, ‘cope with mania’, ‘ocd’⁸) and Thought Content (e.g. ‘mania’, ‘feels anxious’, ‘feels exhausted’).

⁸Obsessive-compulsive disorder

MWEs play a large role in correctly identifying risk factor domains. Factoring them into our models increased classification performance by 15%, a marked improvement over our baseline model. This aligns with our expectations that MWEs comprised of a quotidian vocabulary hold much more clinical significance than when the words in the expressions are treated independently.

Threshold similarity scores also play a large role in determining the precision and recall of our models: higher thresholds lead to a smaller number of false positives and a greater number of false negatives for each risk factor domain. Conversely, more paragraphs are incorrectly classified as Other when thresholds are set higher. Since our classifier will be used in future work as an early step in a data analysis pipeline for determining readmission risk, misclassifying a paragraph with an incorrect risk factor domain at this stage can lead to greater inaccuracies at later stages. Paragraphs misclassified as Other, however, will be discarded from the data pipeline. Therefore, we intentionally set a conservative threshold where only the most confidently labeled paragraphs are assigned membership in a particular domain.

6 Future Work and Conclusion

To achieve our goal of creating a framework for a readmission risk classifier, the present study performed necessary evaluation steps by updating and adding to our model iteratively. In the first stage of the project, we focused on collecting the data necessary for training and testing, and on the domain classification annotation task. At the same time, we began creating the tools necessary for automatically extracting domain relevance scores at the paragraph and document level from patient EHRs using several forms of vectorization and topic modeling. In future versions of our risk factor domain classification model we will explore increasing robustness through sequence modeling that considers more contextual information.

Our current feature set for training a machine learning classifier is relatively small, consisting of paragraph domain scores, bag-of-words, length of stay, and number of previous admissions, but we intend to factor in many additional features that extend beyond the scope of the present study. These include a deeper analysis of clinical narratives in EHRs: our next task will be to extend our EHR data pipeline by distinguishing between

clinically positive and negative phenomena within each risk factor domain. This will involve a series of annotation tasks that will allow us to generate lexicon-based and corpus-based sentiment analysis tools. We can then use these clinical sentiment scores to generate a gradient of patient improvement or deterioration over time.

We will also take into account structured data that have been collected on the target cohort throughout the course of this study such as brain based electrophysiological (EEG) biomarkers, structural brain anatomy from MRI scans (gray matter volume, cortical thickness, cortical surface-area), social and role functioning assessments, personality assessment (NEO-FFI⁹), and various symptom scales (PANSS¹⁰, MADRS¹¹, YMRS¹²). For each feature we consider adding, we will evaluate the performance of the classifier with and without the feature to determine its contribution as a predictor of readmission.

7 Acknowledgments

This work was supported by a grant from the National Institute of Mental Health (grant no. 5R01MH109687 to Mei-Hua Hall). We would also like to thank the LOUHI 2018 Workshop reviewers for their constructive and helpful comments.

References

- Martín Abadi, Paul Barham, Jianmin Chen, Zhifeng Chen, Andy Davis, Jeffrey Dean, Matthieu Devin, Sanjay Ghemawat, Geoffrey Irving, Michael Isard, et al. Tensorflow: a system for large-scale machine learning.
- Donald Emile Addington, Cindy Beck, JianLi Wang, Beverly Adams, Cathy Pryce, Haifeng Zhu, Jian Kang, and Emily McKenzie. 2010. Predictors of admission in first-episode psychosis: developing a risk adjustment model for service comparisons. *Psychiatric Services*, 61(5):483–488.
- Mario Alvarez-Jimenez, A Priebe, SE Hetrick, Sarah Bendall, Eoin Killackey, AG Parker, PD McGorry, and JF Gleeson. 2012. Risk factors for relapse following treatment for first episode psychosis: a systematic review and meta-analysis of longitudinal studies. *Schizophrenia Research*, 139(1-3):116–128.
- ⁹NEO Five-Factor Inventory (Costa and McCrae, 2010)
- ¹⁰Positive and Negative Syndrome Scale (Kay et al., 1987)
- ¹¹Montgomery-Asperg Depression Rating Scale (Montgomery and Åsberg, 1979)
- ¹²Young Mania Rating Scale (Young et al., 1978)

- Abhijit Bendale and Terrance Boult. 2015. Towards open world recognition. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 1893–1902.
- Steven Bird, Ewan Klein, and Edward Loper. 2009. *Natural language processing with Python: analyzing text with the natural language toolkit.* "O'Reilly Media, Inc.".
- Mats Bogren, Cecilia Mattisson, Per-Erik Isberg, and Per Nettelbladt. 2009. How common are psychotic and bipolar disorders? a 50-year follow-up of the lundby population. *Nordic journal of psychiatry*, 63(4):336–346.
- François Chollet et al. 2015. Keras. <https://keras.io>.
- PT Costa and Robert R McCrae. 2010. The neo personality inventory: 3. *Odessa, FL: Psychological assessment resources.*
- Mark Davies and Joseph L Fleiss. 1982. Measuring agreement for multinomial data. *Biometrics*, pages 1047–1051.
- Joseph L Fleiss. 1971. Measuring nominal scale agreement among many raters. *Psychological bulletin*, 76(5):378.
- Philip D Harvey and Alan S Bellack. 2009. Toward a terminology for functional recovery in schizophrenia: is functional remission a viable concept? *Schizophrenia Bulletin*, 35(2):300–306.
- Thomas Insel, Bruce Cuthbert, Marjorie Garvey, Robert Heinssen, Daniel S Pine, Kevin Quinn, Charles Sanislow, and Philip Wang. 2010. Research domain criteria (rdoc): toward a new classification framework for research on mental disorders.
- Lalit P Jain, Walter J Scheirer, and Terrance E Boult. 2014. Multi-class open set recognition using probability of inclusion. In *European Conference on Computer Vision*, pages 393–409. Springer.
- Richard A Johnson and Dean W Wichern. 2004. Multivariate analysis. *Encyclopedia of Statistical Sciences*, 8.
- Stanley R Kay, Abraham Fiszbein, and Lewis A Opler. 1987. The positive and negative syndrome scale (panss) for schizophrenia. *Schizophrenia bulletin*, 13(2):261–276.
- Ronald C Kessler, G Paul Amminger, Sergio Aguilar-Gaxiola, Jordi Alonso, Sing Lee, and T Bedirhan Ustun. 2007. Age of onset of mental disorders: a review of recent literature. *Current opinion in psychiatry*, 20(4):359.
- Diederik P Kingma and Jimmy Ba. 2014. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*.
- J Richard Landis and Gary G Koch. 1977. The measurement of observer agreement for categorical data. *biometrics*, pages 159–174.
- Kim Lorine, Haig Goenjian, Soeun Kim, Alan M Steinberg, Kendall Schmidt, and Armen K Goenjian. 2015. Risk factors associated with psychiatric readmission. *The Journal of nervous and mental disease*, 203(6):425–430.
- James MacQueen et al. 1967. Some methods for classification and analysis of multivariate observations.
- Roshni Mangalore and Martin Knapp. 2007. Cost of schizophrenia in england. *The journal of mental health policy and economics*, 10(1):23–41.
- Thomas H McCoy, Victor M Castro, Hannah R Rosenfield, Andrew Cagan, Isaac S Kohane, and Roy H Perlis. 2015. A clinical perspective on the relevance of research domain criteria in electronic health records. *American Journal of Psychiatry*, 172(4):316–320.
- Stuart A Montgomery and MARIE Åsberg. 1979. A new depression scale designed to be sensitive to change. *The British journal of psychiatry*, 134(4):382–389.
- Harvey J Murff, Fern FitzHenry, Michael E Matheny, Nancy Gentry, Kristen L Kotter, Kimberly Crimin, Robert S Dittus, Amy K Rosen, Peter L Elkin, Steven H Brown, et al. 2011. Automated identification of postoperative complications within an electronic medical record using natural language processing. *Jama*, 306(8):848–855.
- Vinod Nair and Geoffrey E Hinton. 2010. Rectified linear units improve restricted boltzmann machines. In *Proceedings of the 27th international conference on machine learning (ICML-10)*, pages 807–814.
- Mark Olfson, David Mechanic, Carol A Boyer, Stephen Hansell, James Walkup, and Peter J Weiden. 1999. Assessing clinical predictions of early rehospitalization in schizophrenia. *The Journal of nervous and mental disease*, 187(12):721–729.
- Fabian Pedregosa, Gaël Varoquaux, Alexandre Gramfort, Vincent Michel, Bertrand Thirion, Olivier Grisel, Mathieu Blondel, Peter Prettenhofer, Ron Weiss, Vincent Dubourg, et al. 2011. Scikit-learn: Machine learning in python. *Journal of machine learning research*, 12(Oct):2825–2830.
- Jonna Perälä, Jaana Suvisaari, Samuli I Saarni, Kimmo Kuoppasalmi, Erkki Isometsä, Sami Pirkola, Timo Partonen, Annamari Tuulio-Henriksson, Jukka Hintikka, Tuula Kieseppä, et al. 2007. Lifetime prevalence of psychotic and bipolar i disorders in a general population. *Archives of general psychiatry*, 64(1):19–28.
- A Rumshisky, M Ghassemi, T Naumann, P Szolovits, VM Castro, TH McCoy, and RH Perlis. 2016. Predicting early psychiatric readmission with natural

language processing of narrative discharge summaries. *Translational psychiatry*, 6(10):e921.

Walter J Scheirer, Lalit P Jain, and Terrance E Boult.
2014. Probability models for open set recognition.
IEEE transactions on pattern analysis and machine intelligence, 36(11):2317–2324.

Jay P Singh and Seena Fazel. 2010. Forensic risk assessment: A metareview. *Criminal Justice and Behavior*, 37(9):965–988.

Nitish Srivastava, Geoffrey Hinton, Alex Krizhevsky, Ilya Sutskever, and Ruslan Salakhutdinov. 2014. Dropout: a simple way to prevent neural networks from overfitting. *The Journal of Machine Learning Research*, 15(1):1929–1958.

PH Thomsen. 1996. Schizophrenia with childhood and adolescent onseta nationwide register-based study. *Acta Psychiatrica Scandinavica*, 94(3):187–193.

Theo Vos, Ryan M Barber, Brad Bell, Amelia Bertozzi-Villa, Stan Biryukov, Ian Bolliger, Fiona Charlson, Adrian Davis, Louisa Degenhardt, Daniel Dicker, et al. 2015. Global, regional, and national incidence, prevalence, and years lived with disability for 301 acute and chronic diseases and injuries in 188 countries, 1990–2013: a systematic analysis for the global burden of disease study 2013. *The Lancet*, 386(9995):743–800.

Durk Wiersma, Fokko J Nienhuis, Cees J Slooff, and Robert Giel. 1998. Natural course of schizophrenic disorders: a 15-year followup of a dutch incidence cohort. *Schizophrenia bulletin*, 24(1):75–85.

Eric Q Wu, Howard G Birnbaum, Lizheng Shi, Daniel E Ball, Ronald C Kessler, Matthew Moulis, and Jyoti Aggarwal. 2005. The economic burden of schizophrenia in the united states in 2002. *Journal of Clinical Psychiatry*, 66(9):1122–1129.

RC Young, JT Biggs, VE Ziegler, and DA Meyer. 1978. A rating scale for mania: reliability, validity and sensitivity. *The British Journal of Psychiatry*, 133(5):429–435.

Wen Zhang, Taketoshi Yoshida, and Xijin Tang. 2011. A comparative study of tf* idf, lsi and multi-words for text classification. *Expert Systems with Applications*, 38(3):2758–2765.