TITLE: Automated identification of RDoC construct from PubMed abstracts

TEAM: Suraj Subramanian (Pitt), William Schwarzmann (Pitt)

INTRODUCTION:

The Research Domain Criteria (RDoC) is a comprehensive framework developed by the National Institute of Mental Health (NIMH) for describing mental illness in multiple dimensions. This approach is a novel mechanism for mental health classification compared to one-dimensional methods (Sanislow et. al). One of the 2019 RDoC Task challenges is aimed at information retrieval and extraction. The competition provides a dataset of PubMed article abstracts annotated with RDoC constructs. The goal is to submit a ranked list of relevant articles to each RDoC construct. A given Average Precision evaluation metric is used to qualify each construct independently, and then averaged across constructs to compute a Mean Average Precision. Our goal is to develop an approach to this task that retrieves indexed abstracts for a given RDoC "search," providing accurate results by the challenge criteria.

METHODS:

DATA COLLECTION:

The RDoC 2019 Task challenge provides a labeled corpus of 250 PubMed abstracts and the RDoC construct it is relevant to. This is too small a dataset to train a language model, so we obtained 350K abstracts from Medline. From the 250 abstracts, we first identified the set of MeSH headings of those papers. We then retrieved all the literature from Medline that had any of these keywords in their MeSH headings. Although noisy (since not all the retrieved publications were related to the RDoC constructs, or sometimes even mental health), it is a sizable corpus to train a word embeddings hyperspace on.

We also procured definitions from the following mental health ontologies to train a thesaurus of key concepts:

- Mental Disease (https://bioportal.bioontology.org/ontologies/MFOMD/)
- Mental Functioning (https://bioportal.bioontology.org/ontologies/MF/)
- APA Thesaurus (https://bioportal.bioontology.org/ontologies/APAONTO)
- Suicidology (https://bioportal.bioontology.org/ontologies/suicideo)
- Wordnet

PREPROCESSING

Basic preprocessing steps included

- Stop word removal (using NLTK's English stopword dictionary).
- Composite words were split on hyphens and slashes. In future work, we plan to retain these as composite words often have a different significance than their constituents.
- Non-words (punctuations, numbers) are removed.
- Abbreviation expansion. We built a simple function that looks at the *n* preceding words for an *n*-length abbreviation, and adjudges them as candidates if their initials match the

abbreviation by more than 40% (arbitrarily chosen) accuracy. It performs well since in our corpus (article abstracts), abbreviations are usually explained in its first occurrence.

- Filter out verbs, prepositions, adverbs and modals to learn stronger linkages among nouns (concepts) using NLTK's Averaged Perceptron Parts of Speech Tagger.

Lemmatization might serve to reduce noise but we are unsure of the quantum of benefit. So far, we have avoided it since it could have unpredictable results on medical jargon.

VECTORIZATION

We use the gensim (v3.8.1) Word2Vec library to train hyperspaces on our corpus. We suspect a hyperspace trained only on the Medline abstracts would be excessively specific. We are currently generating a model that is initially trained on the thesaurus, and updated with the Medline abstracts. Our hope is the hyperspace trained only on the thesaurus would encode core truths about concepts in mental health, and training over the Medline corpus will provide the linkages among these entities.

Once they hyperspace(s) is obtained, we translate documents in the corpus provided in the RDoC task to the vector space. We tried multiple schemes for this:

1. **Non-weighted average**: Each word in a document is replaced by its corresponding vector. All the vectors are averaged to obtain the representation of the document.
2. **Tfidf-weighted average**: Use tf-idf weights of the words before averaging.
3. **Tfidf-PosFilt:** POS-filtering before tf-idf weighting
4. **Psymed-nonweight:** Remove words in the lower 20th percentile of tf-idf scores in the Medline corpus. Each document is non-weighted average of its tokens.

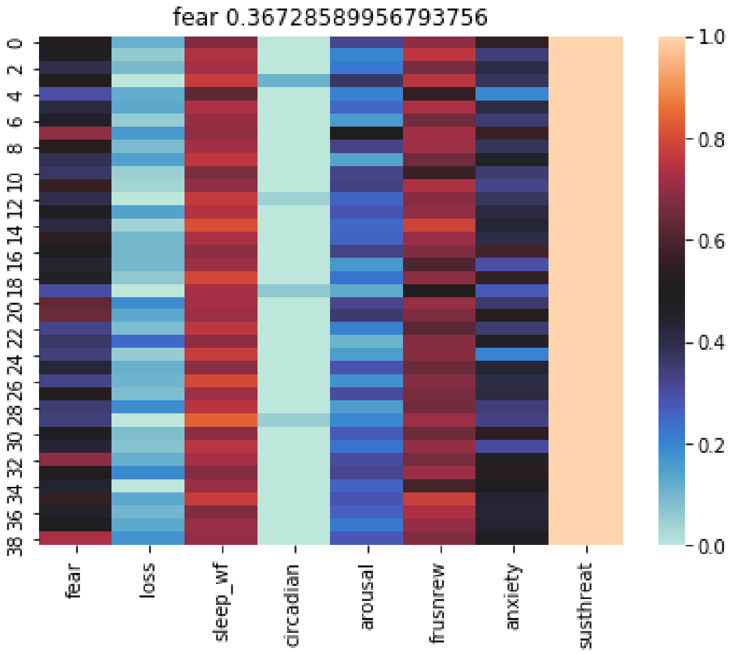
RESULTS:
We obtained the Average Precision for each RDoC, and averaged that to obtain a single-point MAP for each model.

	Non-weighted	Tfidf-Weight	Tfidf-PosFilt	Psymed-weighted	Psymed-nonweight
Fear	0.33	0.39	0.38	0.35	0.37
Loss	0.27	0.74	0.84	0.60	0.81
Arousal	0.31	0.58	0.51	0.29	0.42
Circadian Rhythm	0.44	0.28	0.41	0.96	0.99
Frustrative Nonreward	0.06	0.07	0.08	0.09	0.12
Anxiety	0.36	0.32	0.43	0.26	0.31
Sleep Wakefulness	0.36	0.52	0.42	0.24	0.23
Sustained Threat	0.23	0.21	0.23	0.23	0.22
MAP	**0.30**	**0.39**	**0.41**	**0.38**	**0.43**

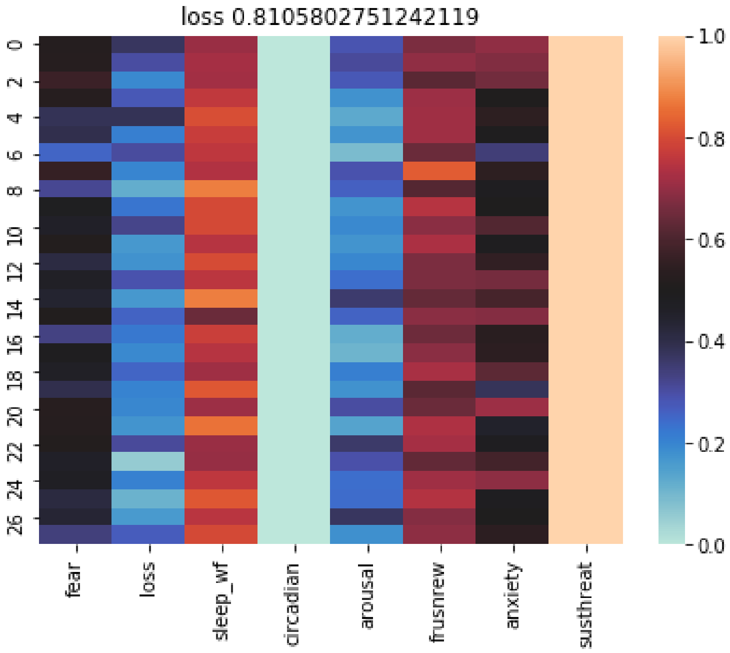
These numbers are far lower than what the competition winners have obtained (the winner has a MAP of 0.86). However, those results were obtained by training and testing on a much larger dataset of ~50k annotated abstracts that we do not have access to.

To get a better sense of what the model is doing, we provide the true annotation and similarity heatmaps of the best model. Given a document, we obtain the similarities with each RDoC. These similarities are independent of each other, so we scale these values between 0 to 1 (using MinMax scaling) to obtain a ranked similarity. We then generate heatmaps based on this scaled relevance to see which RDoC constructs are most often confounded with each other.

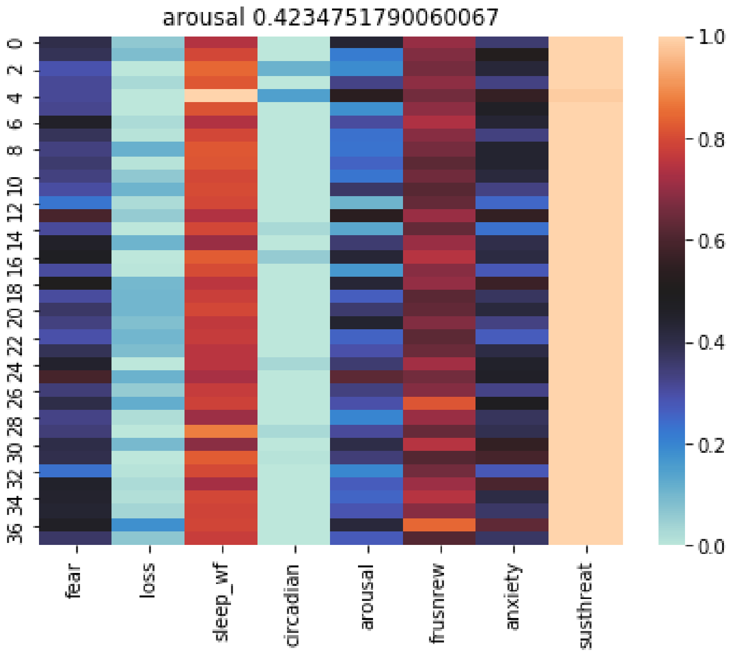
Weighting appears to have an ambiguous effect; on the Medline corpus alone, it helps performance, but on the TfIdf-filtered Medline + Psy thesaurus, performance seems to suffer. A deep-dive analysis into the hyperspaces is required to identify the reasons for this.

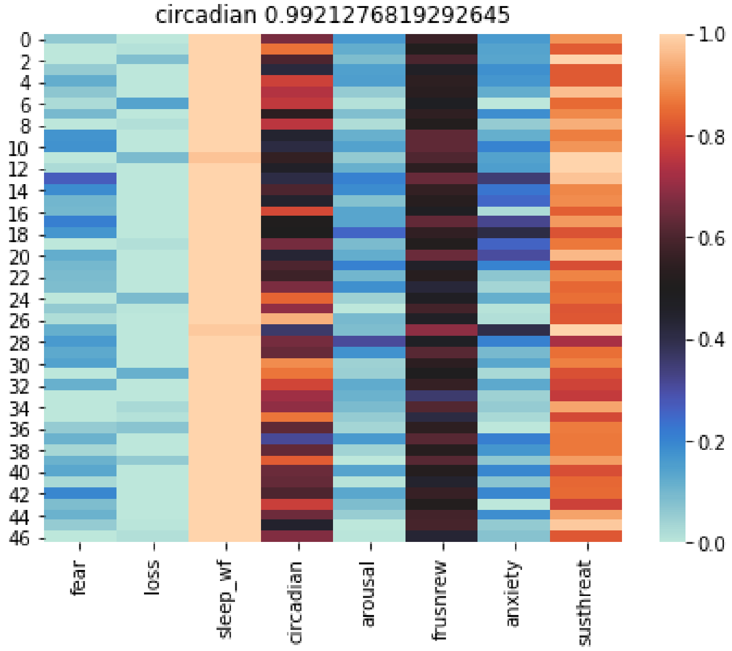


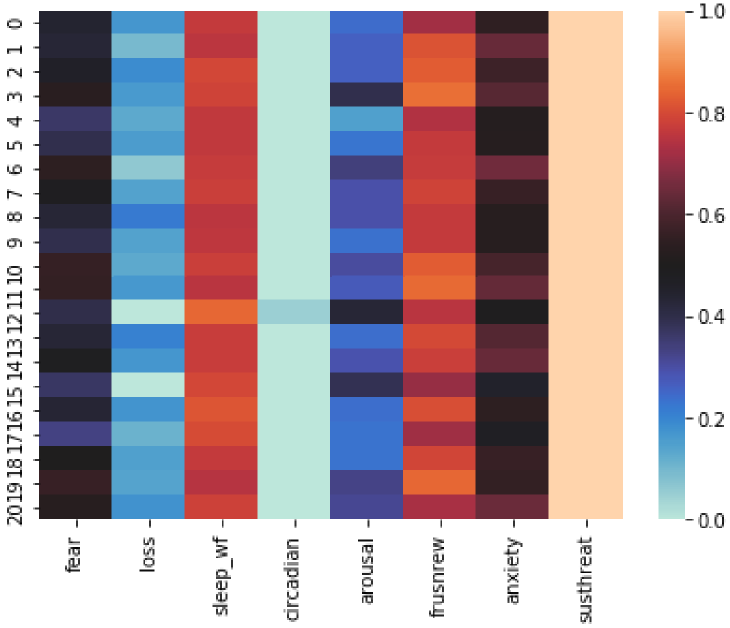
fear 0.36728589956793756



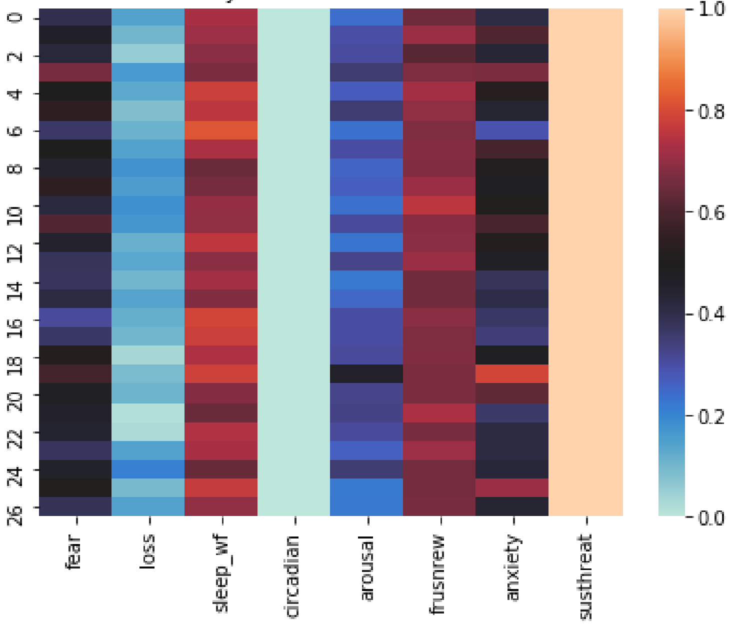
loss 0.8105802751242119


arousal 0.4234751790060067
0
2
4
6
8
10
12
14
16
18
20
22
24
26
28
30
32
34
36
fear
loss
sleep_wf
circadian
arousal
frusnrew
anxiety
susthreat
1.0
0.8
0.6
0.4
0.2
0.0

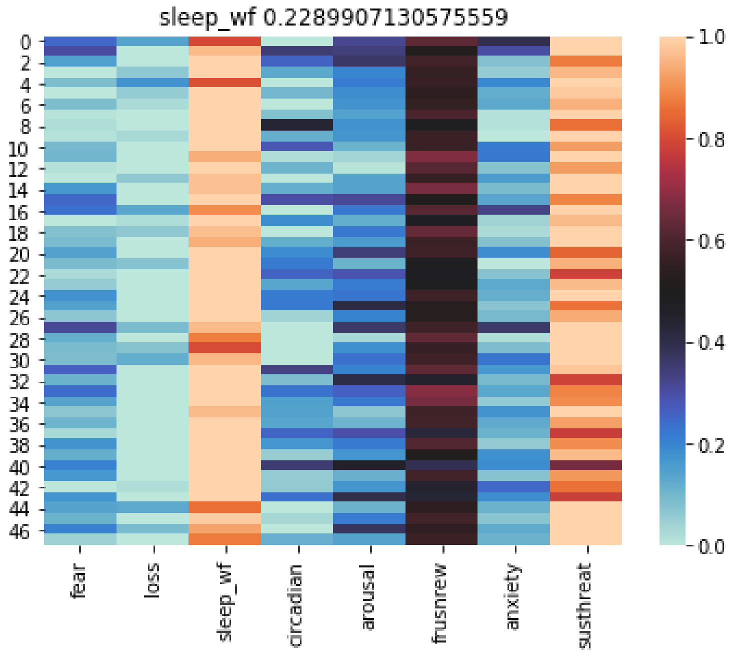

circadian 0.9921276819292645
0
2
4
6
8
10
12
14
16
18
20
22
24
26
28
30
32
34
36
38
40
42
44
46
fear
loss
sleep_wf
circadian
arousal
frusnrew
anxiety
susthreat
1.0
0.8
0.6
0.4
0.2
0.0



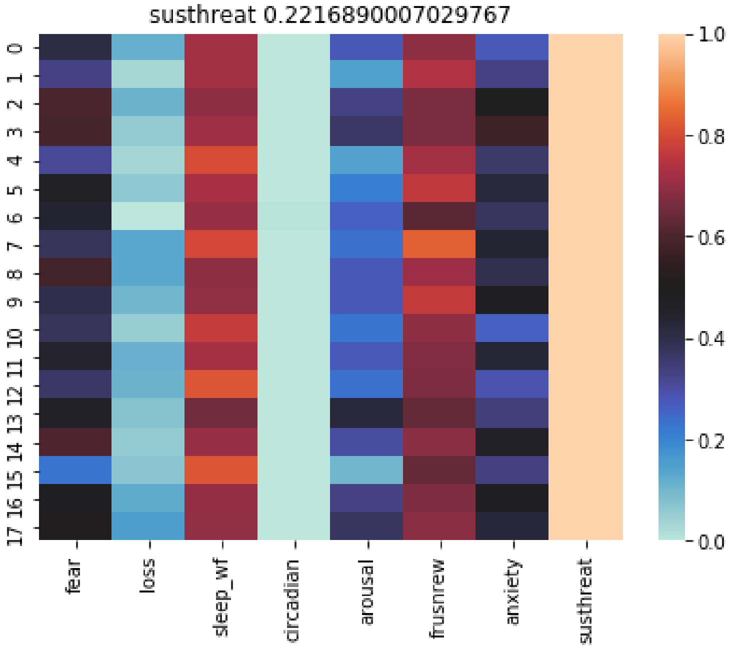
frusnrew 0.12392656787290188



anxiety 0.30526421513796526



sleep_wf 0.2289907130575559



susthreat 0.2216890007029767

LIMITATIONS & NEXT STEPS:

While the system we have right now is not a very robust information retrieval system, this general approach seems to show some promise.

A colloquialism in machine learning is, “Garbage In, Garbage Out.” While our input data is far from garbage, for our model to achieve desired performance, it is pertinent to craft this input to allow our model to focus on what is important. Medical literature is different from free language corpora in phrasing and jargon. This can trip up statistical models that are hitherto used to free text in more general contexts. For example, the POS tagger sometimes tags ‘tryptophan’ as an adjective instead of a noun.

Yet another challenge is the ambiguity of certain phrases; the intended meaning is identified from the context, and even then requires a certain degree of subject matter expertise. Such aspects of the corpora restrict a naive and simplistic application of ML algorithms, and require some craftiness.

Even within the 350K corpus obtained from Medline, we suspect there are abstracts that are not contributing positively for our purposes. While those concepts may be medically significant, it might not be informative for our specific tasks. This noise may have crept in as articles which contain one or more keywords of interest, but only in a peripheral manner. Further work will need to be done in limiting noise in the corpus.

We are using the RDoC constructs as our queries. The NIMH has built the RDoC as a multidimensional representation of certain mental health issues. However in our current system, we have only represented them based on the textual definitions provided. This might have inadequate specificity, leading to conflation and retrieval of ‘non-relevant’ documents. For example, the definition of ‘frustrated nonreward’ contains terms that figure highly in ‘sustained threat’ and ‘arousal’. In an attempt to address this, we made the use of psych- ontologies and Wordnet. However, given the nature of the Word2Vec algorithm, it does not learn conceptual representations from text; all it does is learn co-occurrence patterns. In future work, we will be looking at methods that generate embeddings from ontological relations.

We have identified the next steps that we believe will improve performance:

- Increase the number of training iterations (epochs). Since the corpora consists of short abstracts, it would help the model to go over it a higher number of times.
- Use Doc2Vec to represent documents and definitions.
- Supplement the hyperspace with ontology-embeddings.
- Gain access to the 50k labelled corpus

WORKS CITED:

Sanislow CA, Ferrante M, Pacheco J, Rudorfer MV, Morris SE. Advancing Translational Research Using NIMH Research Domain Criteria and Computational Methods. Neuron. 2019 Mar 6;101(5):779-782. doi: 10.1016/j.neuron.2019.02.024. PubMed PMID: 30844398.

Karpathy, Andrej, Joulin A, Fei-Fei L. "Deep Fragment Embeddings for Bidirectional Image Sentence Mapping." ArXiv, 22 June 2014, arXiv:1406.5679v1 [cs.CV].

Mikolov, T., Sutskever, I., Chen, K., Corrado, G. S., & Dean, J. (2013). Distributed representations of words and phrases and their compositionality. In Advances in neural information processing systems (pp. 3111-3119).

Schütze, H., Manning, C. D., & Raghavan, P. (2008, June). Introduction to information retrieval. In Proceedings of the international communication of association for computing machinery conference (p. 260).