RDoC Task at BioNLP-OST 2019: A Mental Health Informatics Task with Research Domain Criteria

Mohammad Anani[1*], **Nazmul Kazi**[1*], **Matthew Kuntz**[23] **and Indika Kahanda**[1]
[1] Gianforte School of Computing, Montana State University, MT, USA
[2] National Alliance of Mental Illness (NAMI) Montana, Helena, MT, USA
[3] Center for Mental Health Research and Recovery, Montana State University, MT, USA
mohammad.anani@student.montana.edu
{kazinazmul.hasan,matthew.kuntz,indika.kahanda}@montana.edu
[*] The authors wish it to be known that Mohammad Anani and Nazmul Kazi should be regarded as joint first authors.

Abstract

BioNLP Open Shared Tasks (BioNLP-OST) is an international competition organized to facilitate development and sharing of computational tasks of biomedical text mining and solutions to them. For BioNLP-OST 2019, we introduced a new mental health informatics task called "RDoC Task", which is composed of two subtasks: information retrieval and sentence extraction through National Institutes of Mental Health's Research Domain Criteria framework. Five and four teams around the world participated in the two tasks, respectively. According to the performance on the two tasks, we observe that there is room for improvement for text mining on brain research and mental illness.

1 Introduction and Motivation

The breadth of brain research is too expansive to be effectively curated without computational tools especially involving machine learning models. For example, a Pubmed search for "Brain" on August 12, 2019, revealed 854,612 articles[1]. More specifically, an August 12, 2019 search for the single mental illness diagnosis of "depression" revealed 530,519 articles[2]. And a search for anxiety revealed 224,305 articles[3]. It is not possible for researchers to functionally analyze all of the critical data patterns both within a single diagnosis or across diagnoses that could be revealed by those articles.

The challenge of curating brain research has been further complicated by the National Institute of Mental Health's adoption of the Research Domain Criteria (RDoC) [6]. Since 1952, the Diagnostic and Statistical Manual of Mental Disorders and International Classification of Diseases [5] (popularly known as DSM and ICD, respectively), have "reigned supreme" as the single "overarching model of psychiatric classification" [14]. That supremacy began to crumble in 2010 when the National Institute of Mental Health launched the RDoC initiative, an alternate framework to conceptually organize and direct biological research on mental disorders [1]. The RDoC initiative intends "to foster integration not only of psychological and biological measures but also of the psychological and biological constructs those measures measure" [13].

The RDoC initiative has fostered significant debate among brain health researchers. It has also created a significant categorization challenge - specifically how to curate articles completed under the DSM-ICD criteria so their data can be incorporated into the RDoC model. Brain science cannot afford to lose critical insights from the numerous articles on different sides of the categorization divide. Hence, it is vital that all existing and future biomedical literature related to brain research is correctly categorized with respect to the RDoC terminology in addition to DSM-ICD models.

However, manual curation of brain research articles using RDoC terminology by human annotators can be highly resource-consuming due to several reasons. RDoC framework is comprehensive and complex. It is made up six major *domains* of human functioning, which is further broken down to multiple *constructs* that comprise different aspects of the overall range of functions[4]. The RDoC matrix helps describe these constructs using several *units of analysis* such as molecules and circuits. On top of this, the rate of publication of biomedical literature (and by extension brain re-

[1] Pubmed search for Brain conducted on August 12, 2019
[2] Pubmed search for depression conducted on August 12, 2019
[3] Pubmed search for anxiety conducted on August 12, 2019
[4] https://www.nimh.nih.gov/research/research-funded-by-nimh/rdoc/definitions-of-the-rdoc-domains-and-constructs.shtml

Proceedings of the 5th Workshop on BioNLP Open Shared Tasks, pages 216–226
Hong Kong, China, November 4, 2019.

search related literature) is growing at an exponential rate [10]. This means that the gap between annotated versus unannotated articles will continue to grow at an alarming rate unless more efficient means of automated annotation is developed soon.

In order to invite text mining teams around the world to develop informatics models for RDoC, we introduced the RDoC Task[5] at this years' BioNLP-OST 2019 workshop[6]. RDoC task is a combination of two subtasks focusing on a subset of RDoC constructs: (a) Task 1 (RDoC-IR) - retrieving PubMed Abstracts related to RDoC constructs, and (b) Task 2 (RDoC-SE) - extracting the most relevant sentence for a given RDoC construct from a known relevant abstract. Both these tasks represent two very important steps of the typical triage process [10], which are finding the articles related to RDoC constructs and then extracting a specific snippet of information that is useful for curation or downstream tasks such as automatic text summarization [15].

There have been several shared tasks on text mining from biomedical literature and clinical notes in the last decade [19, 12] as well as a few shared tasks related to mental health topics ([4, 18, 22, 21, 30]). CLPsych 2015 Shared Task [4] focused on identifying depression and PTSD users from twitter data, while the same task from the following year (i.e. CLPsych 2016 Shared Task [18]) revolved around classifying the severity of peer support forum posts. One of the i2b2[7] challenges from 2011 focused on the sentiment analysis of suicide notes [22, 21].

In 2017, Uzuner et al. introduced the "The RDoC for Psychiatry" challenge, which was composed of three tracks: de-identification of mental health records [28], determination of symptom severity from a psychiatric evaluation of a patient) related to one of the RDoC domains) [9], and the use of mental health records released through the challenge for answering novel questions [32, 29, 7]. In contrast, the RDoC task is a combination of information retrieval and sentence extraction from Biomedical literature related to RDoC constructs.

To generate benchmark data for the RDoC task, three annotators were used to curate the gold-standard datasets. The registration for the RDoC Task opened in March of 2019. Over 30 teams around the world registered for the two tasks. Training data in two batches were released in the month of April. Test data, again in two batches, were released in June. The participants were asked to submit their final predictions by June 19. Eventually, 4 and 5 groups each competed in Tasks 1 and 2, respectively. The final results were made public immediately after the submission deadline.

Two (out of four) and four (out of five) teams each outperformed the baseline methods in task 1 and 2, respectively. The increase in performance over the baselines were more noticeable in task 2 suggesting that information retrieval for RDoC task may be more challenging. There was quite a lot of variation across the several RDoC constructs used for the tasks suggesting that the complexity of different constructs may hinder certain models and construct-specific methods or models may be a requirement in the future. Overall observations from the RDoC Task highlights the need for more sophisticated method development.

The rest of the paper is organized as follows. Section 2 describes the benchmark or gold-standard data preparation process, development of training and test sets, submission requirements, baseline methods used by the organizers, and the performance measures used for the evaluation. Section 3 presents and discusses the overall results for the two tasks. Finally, Section 4 summarizes the task findings as well as describes the potential future work.

2 RDoC Task setup

RDoC Task is a combination of two subtasks. Participants were allowed to choose to participate in one or both tasks. Task 1 is on retrieving PubMed Abstracts related to RDoC constructs, while Task 2 is on extracting the most relevant sentences for an RDoC construct from an already relevant abstract.

In task 1, participants are given a set of PubMed abstracts and they are required to rank abstracts according to relevance for various RDoC constructs. In task 2, participants are given a set of PubMed abstracts relevant for an RDoC construct, and they are required to extract the most relevant sentence from each abstract for the corresponding RDoC construct.

[5] https://sites.google.com/view/rdoc-task/home
[6] http://2019.bionlp-ost.org
[7] https://www.i2b2.org/

2.1 Timeline

The RDoC Task was organized in two main phases (a) *Training* phase (8 weeks, from April-June 2019), and (b) *Evaluation* phase (1 week in mid-June). At the beginning of the training phase, participants were provided with labeled data (i.e. Training data) and they were expected to develop and fine-tune their models using these known labels. At the beginning of the Evaluation phase, unlabeled data (i.e. Test data) was made available to the participants. They were required to predict labels for this data and submit the predictions to the organizers at the end of the Evaluation phase. Finally, the organizers used the (with-held) labels of the test data for evaluating the accuracy of submissions.

2.2 The benchmark preparation

For the RDoC Task, 8 RDoC constructs out of 25 total constructs from the latest version of the RDoC matrix[8] were used. The motivation was to restrict ourselves to a subset of RDoC framework for which benchmark data can be gathered within a reasonable time-frame. However, these 8 constructs completely cover two of the six domains in the RDoC framework – namely *Negative Valence Systems* and *Arousal and Regulatory* Systems as shown in Table 1.

Table 1: Subset of RDoC constructs used for this task and their domain.

Domain	Construct
Negative Valence Systems	Acute Threat (Fear) Potential Threat (Anxiety) Frustrative Nonreward Sustained Threat Loss
Arousal/Regulatory Systems	Arousal Circadian Rhythms Sleep and Wakefulness

Under the guidance of the Subject Matter Experts from the National Alliance of Mental Illness (NAMI) Montana, the RDoC task benchmark was created by using Entrez e-search utility [26] to search the PubMed database to collect abstracts related to RDoC constructs. That is, we start by using the RDoC construct name as the only keyword to retrieve relevant articles.

If such an approach does not generate the desired number of articles or is too ambiguous on its own (e.g., *Loss* construct), we have utilized terms from the *Behaviors* unit of the RDoC matrix in addition to the construct name.

For example, the The query used for *Loss* construct was "Loss""Amotivation" or "Loss""Anhedonia" or "Loss""Crying" or "Loss""Guilt" or "Loss""Rumination" or "Loss""Sadness" or "Loss""Shame" or "Loss""Withdrawal" or "Loss""Worry". This retrieves about 315 articles, whereas using only "Loss" as the sole query retrieves too many articles (approximately one million articles).

Other queries follow a similar format as Loss when very few (<200) or too many (>10,000) articles were retrieved with the RDoC construct name as the only keyword. 200 abstracts was the desired minimum number of abstracts per construct that we were planning to send to each annotator. So, if the initial search retrieved less articles, it was deemed too narrow for our objective, and we added terms from the *Behavior* elements belonging to that construct to retrieve more than 200 articles. For example, for the construct Frustrative Nonreward, a PubMed search with the construct name only returns 52 abstracts (retrieved on 09/30/2019)[9]. The RDoC page for Frustrative Nonreward contains one element under the Behavior unit: "physical and relational aggression"[10]. Then, using this term, the search query becomes: "Frustrative Nonreward" or "physical aggression" or "relational aggression", which returns 736 abstracts.

10,000 was a rough estimation of an excessively inclusive search term as determined by our Subject Matter Expert. In other words, the construct name on its own (construct Loss, for example) has a very general definition, resulting in retrieving a large heterogeneous set of articles. Therefore, in these situations, other more specific terms describing the construct were used to limit the scope. Upon generating a search query that retrieves a satisfactory number of articles, we sort them by relevance to the query used.

Then the above-retrieved articles were provided

[8] https://www.nimh.nih.gov/research/research-funded-by-nimh/rdoc/constructs/rdoc-matrix.shtml

[9] https://www.ncbi.nlm.nih.gov/pubmed/?term=Frustrative+Nonreward

[10] https://www.nimh.nih.gov/research/research-funded-by-nimh/rdoc/constructs/frustrative-nonreward.shtml

to three annotators for curation (an example of the annotation guidelines used is available online[11]). For each construct, they were asked to read the title and the abstract and determine whether it provides enough evidence that the abstract was related to the construct. If it was related it was annotated as "positive" (or "negative" otherwise). In addition, they were asked to identify up to 3 most relevant sentences to the abstract (i.e. the sentences that provide most evidence that the abstract is related to the said construct). The inter-annotator agreements are given in Table 2. Example annotation of an abstract is depicted in Figure 1.

While acknowledging we generated a *closed* set of articles for the information retrieval task, we emphasize that this complete process was guided by NAMI experts. They typically use keyword search for first finding relevant articles. Then they use manual curation to remove false positives. Hence, our benchmark datasets are developed using this approach. We wanted the RDoC Task to resemble how a typical curator would find information in this domain.

Table 2: Inter-annotator agreement of Task 1 and Task 2. κ_{free}: Free-Marginal Multirater Kappa [24] computed online[12]

RDoC Construct	κ_{free} Task 1	κ_{free} Task 2
Acute Threat	0.37	0.24
Potential Threat	0.45	0.27
Frustrative Nonreward	0.24	0.20
Sustained Threat	0.18	0.14
Loss	0.25	0.29
Arousal	0.64	0.35
Circadian Rhythms	0.95	0.35
Sleep & Wakefulness	0.97	0.51

We consolidated the labels from the three annotators using the majority vote (i.e. if at least 2 annotators agreed on a label, that was used as the final label for the abstract). In addition, we collected all the most relevant sentences by the three annotators (i.e. set union) as the final set of sentences. This means each abstract could have up to 9 most relevant sentences. In our dataset, at most 6 sentences were observed. This consolidated data was used to create training and test sets as described below.

We believe that the task of identifying the most relevant sentence was more challenging for the annotators than the task of identifying whether a given abstract was related to an RDoC construct or not (for the latter task, annotators were choosing between two labels while for the former, they were choosing from k sentences in the abstract). Therefore, it was possible that there would be more variability in annotations for the former task. So, we used the set union to allow for more flexibility.

2.3 Train, Test and Submission data

In the context of the RDoC task, training data refers to the labeled data sets initially provided to the participants for developing their models. Test sets refer to the unlabeled (i.e. with withheld labels) data sets for which they were asked to submit predictions. All the datasets are available online[13].

For each construct, two separate sets of articles (referred to as Set 1 and Set 2) were annotated. Data from the Set 1 and Set 2 were allocated for training and test data, respectively. Annotators were not aware of this distinction. Set 1 and Set 2 splits were randomly performed per each construct separately before annotation. Therefore, explicit stratified sampling was not applicable.

For each construct, a random subset of positive examples from Set 1 was used as the training examples for both Task 1 and 2 (negative examples were not provided). 80% of random abstracts from Set 2 were used as the test set for Task 1 (this included both positive and negative examples). The subset of positive examples in the rest of the Set 2 (i.e. 20%) was used as the test set for Task 2 (negative examples were not used).

2.3.1 Train data

As mentioned above, we provided the participants of the RDoC task with training examples for each of these 8 RDoC constructs. For task 1, the training examples are randomly selected subsets of positive abstracts for each of the RDoC constructs as shown in Table 3. For task 2, we provided up to 6 most relevant sentences for each of the abstract provided as part of Task 1 train data. In other words, the same set of PubMed IDs were used for training data of both tasks. The distribution of the training examples across the eight constructs is

[11] https://montana.box.com/s/kh0hmyn1jcj5ajvr2nibq4iwwgiv3led

[13] https://www.cs.montana.edu/rdoc-task/data/

Title: Characteristics of Physical Aggression in Children of Immigrant Mothers and Non-immigrant Mothers: A Cross-Sectional Analysis of the Survey of Young Canadians.

Abstract: Physical aggression (PA) is important to regulate as early as the preschool years in order to ensure healthy development of children. This study aims to determine the prevalence and characteristics of PA in children of immigrant and non-immigrant mothers. Bivariate and multivariable logistic regression was performed, with the outcome, PA, and covariates including maternal, child, household and neighbourhood characteristics. Twenty percent of children of non-immigrant mothers and 16% of children of immigrant mothers reported PA. The characteristics of PA differ between children of immigrant versus non-immigrant mothers therefore healthcare providers, policy makers, and researchers should be mindful to address PA in these two groups separately, and find ways to tailor current recommended coping strategies and teach children alternative ways to solve problems based on their needs.

RDoC Construct: Sustained Threat

Figure 1: An example of annotating an abstract for both Task 1 and Task 2. The abstract is annotated positive for *Sustained Threat* (Task 1; highlighted in purple) and the most relevant sentence in the abstract is identified (Task 2; highlighted in yellow).

provided in the Table 3 and the distribution of the number of most relevant sentences per construct is shown in Table 4.

Table 3: The number of training examples (positively labeled abstracts) provided for Tasks 1 and 2 across constructs.

RDoC construct	# Abstracts	%
Acute Threat (Fear)	39	14.7
Potential Threat (Anxiety)	27	10.2
Frustrative Nonreward	21	7.9
Sustained Threat	18	6.8
Loss	28	10.5
Arousal	38	14.3
Circadian Rhythms	47	17.7
Sleep and Wakefulness	48	18.1
Total	266	100.0

2.3.2 Test data

The Task 1 test set provided the participants with a random list of 999 relevant (positive) and irrelevant articles (negative) for each of the RDoC constructs (but without the actual labels). The label distribution is given in Table 5. The task 2 test set provided the participants with a list of relevant articles from which they had to extract a relevant sentence with respect to the given RDoC category. The set of abstracts used for test sets of task 1 and 2 were mutually independent for obvious reasons. The distribution of the test set for task 2 across constructs is shown in Table 6 and the distribution of the number of most relevant sentences per construct is provided in Table 4.

2.3.3 Participant Submissions

For task 1, participants were required to submit scores for each abstract in the test set. Scores should correspond to the predicted relevance of the abstract to the given construct. For task 2, participants were required to submit sentences from each abstract that is predicted as the most relevant sentence to the given construct. Submitting a score was not required.

Participants uploaded their submissions through an online web application[14]. We designed the web system to validate the content format of each submission before uploading the file(s) in the server. Upon finding a line that is not properly formatted, the system alerts the participant with an error message including the ill-formatted line number. If the file(s) are properly formatted, the system uploads the submission in the server, automatically analyzes the submission using python scripts and immediately reports the scores of two selected constructs, *Acute Threat (Fear)* and *Loss*, back to the participant.

The participants were allowed to make an un-

[14]https://www.cs.montana.edu/rdoc-task/

Table 4: Distribution of the number of most relevant (gold-standard) sentences in abstracts for each construct in the training data. #x: the percentage of abstracts with x relevant sentences.

RDoC Construct	Train Data						Test Data			
	#1	#2	#3	#4	#5	#6	#1	#2	#3	#4
Acute Threat (Fear)	0.0	15.4	35.9	35.9	10.3	2.6	15.8	31.6	42.1	10.5
Potential Threat (Anxiety)	11.1	33.3	55.6	0.0	0.0	0.0	38.2	35.3	20.6	5.9
Frustrative Nonreward	4.8	47.6	47.6	0.0	0.0	0.0	54.3	37.1	8.6	0.0
Sustained Threat	5.6	61.1	33.3	0.0	0.0	0.0	38.9	41.7	16.7	2.8
Loss	10.7	25.0	42.9	21.4	0.0	0.0	61.8	32.4	5.9	0.0
Arousal	7.9	63.2	28.9	0.0	0.0	0.0	23.1	53.8	15.4	7.7
Circadian Rhythms	2.1	51.1	46.8	0.0	0.0	0.0	20.0	40.0	26.7	13.3
Sleep and Wakefulness	10.4	62.5	27.1	0.0	0.0	0.0	26.7	36.7	30.0	6.7

Table 5: The number of abstracts in test set for task 1. Pos and %: number of positively labeled abstracts and their percentages, and Neg: number of negatively labeled abstracts.

RDoC construct	# Pos	%	# Neg
Acute Threat (Fear)	53	67.1	26
Potential Threat (Anxiety)	124	89.2	15
Frustrative Nonreward	96	66.7	48
Sustained Threat	82	56.2	64
Loss	90	65.2	48
Arousal	97	89.8	11
Circadian Rhythms	123	100.0	0
Sleep and Wakefulness	121	99.2	1
Total	786	78.7	213

Table 6: The number of abstracts and their percentages in test set for task 2.

RDoC construct	# Abstracts	%
Acute Threat (Fear)	19	7.8
Potential Threat (Anxiety)	34	13.9
Frustrative Nonreward	35	14.3
Sustained Threat	36	14.8
Loss	34	13.9
Arousal	26	10.7
Circadian Rhythms	30	12.3
Sleep and Wakefulness	30	12.3
Total	244	100.0

limited number of submissions and the scores from past submissions were discarded upon a new submission. This meant they could re-submit until they achieved a satisfactory performance for the above two constructs. The performance scores for all the constructs were made available immediately after the submission deadline. The older scores were only discarded for the purposes of the final evaluation. However, these scores are retained for potential future research.

2.4 Baseline methods

We used TF-IDF [23] with smooth IDF weights and cosine similarity [27] to calculate the similarity score for each document against a query and used these scores to rank the documents by relevance. Regardless of the task, we used the corresponding construct name concatenated with its definition as the query string. We used the definitions of constructs as defined by the National Institute of Mental Health listed online[15].

For task 1, each document is the title concatenated with the corresponding abstract and the similarity scores are used to rank the articles for each construct. For task 2, documents are the sentences of the abstracts and the top-ranked sentence per abstract was returned based on the similarity scores. All the baseline models were implemented using the Scikit-learn Python library [20]. No preprocessing techniques were applied to the abstract text. In addition to the above TFIDF-based baseline, we also used BM25 [25] as a baseline. But due to its comparatively lower performance on both tasks 1 and 2, BM25 values are not reported in this paper.

[15] https://www.nimh.nih.gov/research/research-funded-by-nimh/rdoc/definitions-of-the-rdoc-domains-and-constructs.shtml

2.5 Metrics used for evaluation

For task 1, we use Mean Average Precision (MAP) [16] as the performance measure because it is one of the most frequently used measures for IR [31, 8, 11]. First, we compute the Average Precision (AP) for each construct independently and macro-average across the constructs to compute the Mean Average Precision. For task 2, due to the non-applicability of utilizing popular standard measures such as precision and recall [3], we define the *Accuracy* as the percentage of abstracts with correctly predicted most relevant sentence. If at least one of the gold-standard sentences match the predicted sentence, it is counted as 1 and 0 otherwise (therefore, note that this measure is not the same as the typical accuracy measure used in Natural Language Processing and Machine Learning. We average across constructs to get the *Macro Average Accuracy*.

It should be pointed out that, technically, there is no "negative" class for the task 2 (in the traditional sense used for predictive models). Participants are given abstracts already known to be relevant to a construct. They are asked to submit just one sentence that they think is the most relevant (or that helps them the most for finding the relevance between the given abstract and the construct). Hence the participants are unable to gain undue advantages due to any class imbalances even though the above-defined performance measure may closely resemble the typical "Accuracy". Also, since we did not collect confidence scores for task 2, we did not compute threshold independent measures such as AUROC (area under the ROC curve).

3 Results and Discussion

Inter-annotator agreements for many of the constructs in both tasks 1 and 2 are relatively low (see table 2). According to the annotators, there were several reasons why information retrieval and sentence extraction with RDoC was reasonably challenging. The very generalized nature of the RDoC constructs, as well as ambiguity in the language stating the purpose/hypothesis/results of the experiment, made it difficult to find the relevance of a given abstract to an RDoC construct. The way the abstracts were written, made it seem such that it could be potentially tied to/or not, to various RDoC sentences.

Annotators reported that they had difficulties with the 'Sustained Threat' and the 'Frustrative Non-Reward' constructs. For example, some annotators felt that every abstract that they read was related to Frustrative Non-Reward construct because many of the abstracts specifically studied the relational and physical aggressive behaviors. Although a lot of the studies tested these behaviors, it was challenging to figure out if they were "directly" related to Frustrative Non-Reward or not. For instance, several studies comparatively tested relational and physical aggression between genders (2 behaviors of Frustrative Non-Reward), but the abstracts didn't explicitly mention "withdrawal or prevention" of a reward (the definition). Therefore, when annotating, if they've felt that the research would benefit or help further understand Frustrative Non-Reward and its associated behaviors, they've annotated it as related (this included environmental, social, and biological factors influencing relational and physical aggression).

Over thirty teams registered to participate in at least one of the RDoC tasks. Eventually, 5 teams submitted their predictions; four teams submitted for both tasks and one team for only task 1. In the following analysis, we will be using the unique team identifiers (assigned during the task registration[16]) for referring to the 5 teams. Note that these team identifiers bear no significance other than identifying different teams.

3.1 Task 1: Information Retrieval

Four teams submitted their predictions for this task and their scores are reported in Table 7. Bold entries indicate the highest score for the corresponding construct. Although included in Table 7, we excluded the two constructs, *Circadian Rhythms* and *Sleep and Wakefulness*, from the final analysis since these constructs contain one and zero negative articles, respectively, leading to perfect performance (see Table 5). Team 30 achieved the highest mean average precision (0.86) among all teams. Though Team 10 achieved the second-highest mean average precision (0.85) that is very close to the highest, we found a statistically significant difference between the scores of these two teams (paired t-test, p=0.005, $\alpha = 0.05$). Team 30 achieved the highest scores for *Frustrative Nonreward*, *Loss* and *Potential Threat (Anxiety)* whereas Team 10 achieved the highest scores for the other three constructs. Though it seems the

[16] https://sites.google.com/view/rdoc-task/registration

scores achieved by the Team 10 and 30 is close to the baseline, we found these scores to be statistically significantly higher from the baseline for both Team 10 (paired t-test, p=0.022) and Team 30 (paired t-test, p=0.043) using $\alpha = 0.05$.

The last column in Table 7 reports the average score for the corresponding construct. It is seemingly easier to rank the relevant articles for *Arousal* and *Potential Threat (Anxiety)* whereas it is moderately difficult for *Sustained Threat*. Sustained Threat being more challenging for IR may be explained by the fact that the annotators also found it to be the most challenging construct for task 1 annotation.

3.2 Task 2: Sentence Extraction

Five teams submitted their predictions for this task and their scores are reported in Table 8. Bold entries indicate the highest score for the corresponding construct. Team 30 again achieved the highest macro average accuracy (0.58) among all the teams and the highest score for five out of eight constructs. Team 7 achieved the highest score for the rest of the three constructs with significant improvement over Team 30. Construct-wise highest scores of *Sustained Threat*, *Arousal* and *Circadian Rhythms*, achieved by either Team 7 or Team 30, are higher by about 0.27 compared to the baseline performance. In addition, the highest scores for other constructs are also higher by more than 0.17 compared to the baseline performance.

Frustrative Nonreward has the lowest average score (0.31) among all the constructs. Moreover, its highest score (0.43) is also the lowest among all the highest scores. So, extracting the most relevant sentences for *Frustrative Nonreward* is seemingly more difficult compared to the other constructs.

Typically, participating teams performed relatively better on shorter abstracts (see Table 9), which is intuitive due to that fact the models have a higher chance of finding the most similar sentences for shorter abstracts. Similarly, they performed well for abstracts with more gold-standard sentences (see Table 10). This is also intuitive because when there are more gold-standard sentences, there is a higher chance of matching one of them.

4 Conclusion and Future work

We introduced a novel mental health informatics task called RDoC task at this years BioNLP-OST 2019 workshop. RDoC task is a combination of two subtasks on information retrieval and sentence extraction using the RDoC framework. Originally, over 30 teams registered, highlighting a significant interested in mental health informatics and/ or RDoC. Eventually, four and five teams participated in the information retrieval and sentence extraction tasks, respectively.

Overall results show that the top-performing team was able to easily outperform the baseline models for most of the constructs. On the other hand, the baseline methods outperform at least one system (often more). This is surprising given that the baseline models are not sophisticated. One reason could be that the baseline methods do not utilize training data, while the participating methods may have been overfitted to the training data. Another reason could be, these simple baselines perform better than (most likely more complex) participating models due to working with shorter documents (i.e. abstracts). If the full texts were made available, models primarily depended on TFIDF may struggle to achieve good performance. Regardless, this calls for more sophisticated methods for both tasks because any other sophisticated method (such as Lucene [17] or MetaMap [2]) used a baseline may have outperformed even more participating teams.

The publicly made available gold-standard data should serve as a valuable resource for the brain research/ mental health and RDoC researchers and curators going forward. In the future iterations of the RDoC task, we would like to incorporate either all available or a well-representative set of RDoC constructs covering all domains. We plan to improve the quality of benchmark data using "reconciliation" instead of "majority voting" as well as using improved search that uses MeSH and/ or other vocabularies.

And equally important aspect would be to explore information extraction tasks such as extracting various entities under different RDoC units of analysis, which is likely more useful for the curators. This would also mean an exploration of incorporating full text in addition to abstracts will be required due to the abundance of entities existing in the full articles compared to just the abstract. Last but not least, exploring clever ways to maintain the enthusiasm of the registered teams would be highly valuable to the overall success of the future iterations of the RDoC task .

Table 7: Performance of retrieving PubMed Abstracts related to the corresponding RDoC construct (Task 1). Four teams participated (T10, T21, T22, and T30). IQR: inter-quartile range. Bolded scores are the highest across all teams per the construct.

RDoC construct	Baseline	T10	T21	T22	T30	Avg	IQR
Acute Threat (Fear)	0.74	**0.89**	0.83	0.67	0.85	0.81	0.17
Potential Threat (Anxiety)	0.90	0.87	0.89	0.81	**0.94**	0.88	0.10
Frustrative Nonreward	0.70	0.69	0.67	0.61	**0.73**	0.68	0.10
Sustained Threat	0.64	**0.64**	**0.64**	0.41	0.63	0.58	0.18
Loss	0.77	0.74	0.71	0.61	**0.78**	0.71	0.14
Arousal	0.95	**0.93**	0.91	0.84	0.92	0.90	0.07
Circadian Rhythms	1.00	1.00	1.00	1.00	1.00	1.00	0.00
Sleep and Wakefulness	1.00	1.00	1.00	0.98	1.00	1.00	0.02
Mean Average Precision	0.84	0.85	0.83	0.74	**0.86**	–	–

Table 8: Performance of extracting the most relevant sentence from each abstract related to the corresponding RDoC construct (Task 2). Five teams participated (T7, T10, T21, T22, and T30). IQR: inter-quartile range.

RDoC construct	Baseline	T7	T10	T21	T22	T30	Avg	IQR
Acute Threat (Fear)	0.53	0.58	0.68	0.37	0.47	**0.74**	0.57	0.29
Potential Threat (Anxiety)	0.41	0.41	0.32	0.15	0.38	**0.59**	0.37	0.27
Frustrative Nonreward	0.23	**0.43**	0.34	0.11	0.29	0.37	0.31	0.20
Sustained Threat	0.19	**0.47**	0.36	0.14	**0.47**	0.42	0.37	0.22
Loss	0.53	0.26	0.56	0.26	0.62	**0.74**	0.49	0.42
Arousal	0.46	0.46	0.62	0.12	0.42	**0.73**	0.47	0.41
Circadian Rhythms	0.43	**0.70**	0.47	0.10	0.60	0.47	0.47	0.37
Sleep and Wakefulness	0.43	0.33	0.50	0.17	0.57	**0.60**	0.43	0.34
Macro Average Accuracy	0.40	0.46	0.48	0.18	0.48	**0.58**	–	–

Table 9: Variation of Accuracy over various size of abstract. #*m*-*n*: abstracts with *m* to *n* sentences.

RDoC construct	#3-8	#9-14	#15-20
Acute Threat	0.60	0.64	0.40
Potential Threat	0.47	0.39	–
Frustrative Nonreward	0.28	0.25	0.50
Sustained Threat	0.39	0.32	0.40
Loss	0.62	0.60	0.31
Arousal	0.53	0.39	–
Circadian Rhythms	0.38	0.54	0.00
Sleep & Wakefulness	0.58	0.42	–

Table 10: Variation of Accuracy over the number of most relevant (gold-standard) sentences in abstracts. #x: abstracts with x relevant (gold-standard) sentences.

RDoC construct	#1	#2	#3	#4
Acute Threat	0.29	0.60	0.69	0.80
Potential Threat	0.31	0.47	0.57	0.75
Frustrative Nonreward	0.18	0.35	0.54	–
Sustained Threat	0.29	0.42	0.37	0.25
Loss	0.56	0.65	0.50	–
Arousal	0.47	0.45	0.65	0.40
Circadian Rhythms	0.17	0.41	0.60	0.61
Sleep & Wakefulness	0.23	0.56	0.67	0.80

Acknowledgments

This work was partially funded by The Center for Mental Health Research and Recovery (CMHRR) at Montana State University (MSU). We would like to thank Robell Basset, Lenin Lewis, Ninoo De Silva, and Hannah Reiser (from the Department of Psychology, MSU), and Soumilee Chaudhuri (from the Department of Cell Biology & Neuroscience, MSU) for assisting the curation process.

References

[1] Dean Carcone and Anthony C Ruocco. Six years of research on the National Institute of Mental Healths Research Domain Criteria (RDoC) initiative: a systematic review. *Frontiers in cellular neuroscience*, 11:46, 2017.

[2] K Bretonnel Cohen, Tom Christiansen, and Lawrence E Hunter. Metamap is a superior baseline to a standard document retrieval engine for the task of finding patient cohorts in clinical free text. In *TREC*. Citeseer, 2011.

[3] Kevin Bretonnel Cohen and Dina Demner-Fushman. *Biomedical natural language processing*, volume 11. John Benjamins Publishing Company, 2014.

[4] Glen Coppersmith, Mark Dredze, Craig Harman, Kristy Hollingshead, and Margaret Mitchell. Clpsych 2015 shared task: Depression and ptsd on twitter. In *Proceedings of the 2nd Workshop on Computational Linguistics and Clinical Psychology: From Linguistic Signal to Clinical Reality*, pages 31–39, 2015.

[5] Bruce N Cuthbert. The rdoc framework: facilitating transition from icd/dsm to dimensional approaches that integrate neuroscience and psychopathology. *World Psychiatry*, 13(1):28–35, 2014.

[6] Bruce N Cuthbert and Thomas R Insel. Toward the future of psychiatric diagnosis: the seven pillars of rdoc. *BMC medicine*, 11(1):126, 2013.

[7] Hong-Jie Dai, Emily Chia-Yu Su, Mohy Uddin, Jitendra Jonnagaddala, Chi-Shin Wu, and Shabbir Syed-Abdul. Exploring associations of clinical and social parameters with violent behaviors among psychiatric patients. *Journal of biomedical informatics*, 75:S149–S159, 2017.

[8] Daniel Dopp, Adam Morrone, and Indika Kahanda. KinDER: A biocuration tool for extracting kinase knowledge from biomedical literature. *Proceedings of the BioCreative VI Workshop*, Oct 2017.

[9] Michele Filannino, Amber Stubbs, and Özlem Uzuner. Symptom severity prediction from neuropsychiatric clinical records: Overview of 2016 cegs n-grid shared tasks track 2. *Journal of biomedical informatics*, 75:S62–S70, 2017.

[10] International Society for Biocuration. Biocuration: Distilling data into knowledge. *PLOS Biology*, 16(4):1–8, 04 2018.

[11] Julien Gobeill, Pascale Gaudet, Daniel Dopp, Adam Morrone, Indika Kahanda, Yi-Yu Hsu, Chih-Hsuan Wei, Zhiyong Lu, and Patrick Ruch. Overview of the biocreative vi text-mining services for kinome curation track. *Database*, 2018(1):bay104, 2018.

[12] Chung-Chi Huang and Zhiyong Lu. Community challenges in biomedical text mining over 10 years: success, failure and the future. *Briefings in bioinformatics*, 17(1):132–144, 2015.

[13] Jessica I Lake, Cindy M Yee, and Gregory A Miller. Misunderstanding rdoc. *Zeitschrift für Psychologie*, 2017.

[14] Scott O Lilienfeld and Michael T Treadway. Clashing diagnostic approaches: Dsm-icd versus rdoc. *Annual review of clinical psychology*, 12:435–463, 2016.

[15] Inderjeet Mani. *Advances in automatic text summarization*. MIT press, 1999.

[16] Christopher D. Manning, Prabhakar Raghavan, and Hinrich Schtze. *Evaluation in information retrieval*, page 139161. Cambridge University Press, 2008.

[17] Michael McCandless, Erik Hatcher, and Otis Gospodnetic. *Lucene in action: covers Apache Lucene 3.0*. Manning Publications Co., 2010.

[18] David N Milne, Glen Pink, Ben Hachey, and Rafael A Calvo. Clpsych 2016 shared task: Triaging content in online peer-support forums. In *Proceedings of the Third Workshop on Computational Linguistics and Clinical Psychology*, pages 118–127, 2016.

[19] Malvina Nissim, Lasha Abzianidze, Kilian Evang, Rob van der Goot, Hessel Haagsma, Barbara Plank, and Martijn Wieling. Sharing is caring: The future of shared tasks. *Computational Linguistics*, 43(4):897–904, 2017.

[20] F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, et al. Scikit-learn: Machine learning in Python. *Journal of Machine Learning Research*, 12:2825–2830, 2011.

[21] John P Pestian, Pawel Matykiewicz, and Michelle Linn-Gust. What's in a note: construction of a suicide note corpus. *Biomedical informatics insights*, 5:BII–S10213, 2012.

[22] John P Pestian, Pawel Matykiewicz, Michelle Linn-Gust, Brett South, Ozlem Uzuner, Jan Wiebe, K Bretonnel Cohen, John Hurdle, and Christopher Brew. Sentiment analysis of suicide notes: A shared task. *Biomedical informatics insights*, 5:BII–S9042, 2012.

[23] Anand Rajaraman and Jeffrey David Ullman. *Data Mining*, page 117. Cambridge University Press, 2011.

[24] Justus J Randolph. Free-marginal multirater kappa (multirater k [free]): An alternative to fleiss' fixed-marginal multirater kappa. *Online submission*, 2005.

[25] Stephen Robertson and Hugo Zaragoza. The probabilistic relevance framework: Bm25 and beyond. *Found. Trends Inf. Retr.*, 3(4):333–389, April 2009.

[26] Eric W. Sayers, Tanya Barrett, Dennis A. Benson, et al. Database resources of the National Center for Biotechnology Information. *Nucleic Acids Research*, 38(suppl_1):D5–D16, 1 2010.

[27] Amit Singhal et al. Modern information retrieval: A brief overview. *IEEE Data Eng. Bull.*, 24(4):35–43, 2001.

[28] Amber Stubbs, Michele Filannino, and Özlem Uzuner. De-identification of psychiatric intake records: Overview of 2016 cegs n-grid shared tasks track 1. *Journal of biomedical informatics*, 75:S4–S18, 2017.

[29] Tung Tran and Ramakanth Kavuluru. Predicting mental conditions based on history of present illness in psychiatric notes with deep neural networks. *Journal of biomedical informatics*, 75:S138–S148, 2017.

[30] Özlem Uzuner, Amber Stubbs, and Michele Filannino. A natural language processing challenge for clinical records: Research domains criteria (RDoC) for psychiatry. *Journal of biomedical informatics*, 75:S1–S3, 2017.

[31] Ellen M Voorhees, Donna K Harman, et al. *TREC: Experiment and evaluation in information retrieval*, volume 63. MIT press Cambridge, 2005.

[32] Yaoyun Zhang, Olivia Zhang, Yonghui Wu, Hee-Jin Lee, Jun Xu, Hua Xu, and Kirk Roberts. Psychiatric symptom recognition without labeled data using distributional representations of phrases and on-line knowledge. *Journal of biomedical informatics*, 75:S129–S137, 2017.