Categorization of Primary Scientific Literature Based on Research Domain Criteria

Charles C Carey (charlieccarey@gmail.com)

March 31, 2016

Contents

Abstract

The Research Domains and Constructs (RDoC)[1] is a framework established by the Natiional Institute of Mental Health to help direct the future of NIMH supported research and more broadly clinical practice. A main feature of this framework is the establishment of concepts or basic functions that might bridge or cross traditional clinical boundaries and thereby reorient research towards a functional rather than a disease based approach. For example, hallucinations are features not only manifested in patients with schizophrenia but also with borderline personality disorder, brain injury, and even the non-clinical population[2]. The ability to do discovery of the existing scientific literature along RDoC constructs would be very desirable. In practice this is sometimes a difficult or frustrating experience as Pubmed searches are difficult to design or contain a lot of non-target articles for some constructs. We herein explore use of the machine learning database system DeepDive[3] coupled with simple natural language processing on abstracts to categorize the psychological primary literature according to RDoC Constructs or subconstructs.

1 Introduction

The primary literature associated with certain RDoC constructs is easy to obtain by searches of PubMed. In the best case, the National Library of Medicine has defined a Medical Subject Heading (MeSH) which closely corresponds to RDoC. An example of a great correspondence between MeSH and RDoC is 'Auditory Perception' which **is** a MeSH term that seems to accurately reflect the RDoC Constructs meaning. However, such correspondence does not exist for many RDoC constructs. Examples of RDoC constructs

[1] http://www.nimh.nih.gov/research-priorities/rdoc/index.shtml

[2] http://www.ncbi.nlm.nih.gov/pmc/articles/PMC4141312/

[3] http://deepdive.stanford.edu/

with no MeSH term, and more importantly, poor full text retrieval from pubmed searches are ‘Loss’ and ‘Agency’.

Additionally, even where the MeSH terms are useful, their is the recognition that the MeSH annotation process is a human based manual effort assisted by an automated Medical Text Indexer (MTI) program developed by NLM (for some discussion, see Huang[4]). MeSH annotations efforts, as all such labeling and indexing methods are potentially error prone, with moderate inter-human agreement. As a result, this is the subject of active research.

Given the possible non-reliability of MeSH annotation, and the fact that pubmed searches are not RDoC aware, we explore an alternative to categorization of the psychological literature with awareness of the goals of RDoC and with our own RDoC aware human annotators.

We therefore sought our own categorization of the primary literature along RDoC constructs, designing our own DeepDive applications.

2 Customizing our use of the Deepdive System.

The DeepDive system is an open source tool now developed at Stanford that combines the management of information and inference in a database. The main advantages of DeepDive over other machine learning tools is this database management along with keeping the machine learning algorithms (specially designed in a database aware manner) at a low level within the system such that users do not need to be concerned with the particular machine learning algorithms used. The database at the heart of DeepDive is the popular Postgres database. The algorithms are combinations or variations of machine learning algorithms that appear to be common in the machine based text training and learning fields.

2.1 DeepDive, Spacy and our training and testing data

The focus of DeepDive is on the user’s data rather than the machine learning algorithm. In our case, the raw data consisted of sentences or entire abstracts derived from pubmed and labeled according to the RDoC construct confirmed by our human annotators or presumed based on how clean simple pubmed searches were for a particular construct.

DeepDive includes a training step to recognize a relationshiop (to RDoC term in our case) according to rules we supply and manipulations of the

[4] http://www.ncbi.nlm.nih.gov/pmc/articles/PMC3168302/

raw data that we instruct. The most common and successful approach in machine learning text categorization is bags of words and variations of bags of words. Bags of words simply means the text is split into its constituent single words.

2.2 SpaCy to obtain bags of words.

Obtaining the bags of words requires a Natural Language Processing tool. Using the spaCy[5] library in python gives us the capability to do other variations of NLP including stemming, lemmatizing, and filtering of most common words as well as combining words according to their adjacency or a particular skip pattern designed to give more contextual information to the single words. Many other features of spaCy mauy be useful for future efforts but we are concerned here solely with bag of words variations. We used a lemmatized un-filtered bag of words approach. We also explored combining adjacent words but, when using small training and test sets discovered this had little effect.

2.3 Disadvantages and circumvention of disadvantages of DeepDive

While the DeepDive system has some advantages in its simplicity and hiding of the learning algorithms, it suffers somewhat from non-modular design and an apparent oversight of the workflow and validation that those more experienced with machine learning algorithms seek.

Much of our effort was therefore aimed at overcoming some of the deficiencies of DeepDive. In particular, DeepDive, presumably in its efforts to hide some of the complexity of machine learning, offered no easy way to establish the performance of the training sets, and did not provide tools for cross-validation. Instead it simply offered up a single model and some difficult to understand plots of the performance of a heldout portion of the data used as a test set.

The ability to assess the performance of the model on the training data as well as test data, and cross validation using different subsets of training and / or test data are critical features that machine learning experts might wish to more fully understand. Thus much of our efforts in the 1st 2 months of the project were deployed to accomplish this visibility.

[5] https://spacy.io/

2.4 A newer version of DeepDive might have obviated the need for some of our effort.

While we succeeded in this effort as much as possible within their framework, we take note that in the course of our project, DeepDive released a new version with the specific goals of offering a more modular system. Having achieved our main objectives we have not explored this new DeepDive system. However, we would suggest starting with the newer system or its derivatives as they seemed to be designed specifically to addresss some of the concerns we shared above (and which we shared, without response, with some of the DeepDive developers themselves).

DeepDive version we developed against:

- deepdive version: v0.7.0-0-gf4e6dfe (07/14/2015)

DeepDive latest release includes more modularity:

- deepdive version: v0.8.0 (2/18/2016)

3 PDF data sources and management.

3.1 Pubmed searches.

We formulated various searches to identify primary literature in pubmed related to a topic. Substantial filtering was included to specify more recent results and focus on the human literature available as full text, or at least abstract as free. The searches were manually and interactively tuned in pubmed. By specifying full text availability, we had reasonable expectations we could programatically retrieve a PDF, an article format with which annotators had ease and familiarity and that facilitates manual markup by annotators.

Example searches:

- search_filter included along with many additional search terms
  - '("humans"[MeSH Terms] OR "humans"[All Fields] OR "human"[All Fields]) AND "loattrfree full text"[sb] AND ("2011/01/31"[PDAT] : "2016/01/29"[PDAT]) AND '
- Auditory Perception (An easier search)

  - ‘(“humans”[MeSH Terms] OR “humans”[All Fields] OR “human”[All Fields]) AND (“auditory perception”[All Fields] OR “auditory perception”[MeSH Terms]) AND (hasabstract[text] AND “2011/01/31”[PDat] : “2016/01/29”[PDat])’

- Agency (Required more advanced searches or many searches, a few examples.)
  - search_filter + ‘”corollary discharge”[All Fields] AND (“schizophrenia”[MeSH Terms] OR “schizophrenia”[All Fields])’
  - search_filter + ‘(Corollary[All Fields] AND “discharge”[All Fields]) OR (“corollary discharge”[All Fields] AND Paradigm[All Fields])’
  - search_filter + ‘”agency”[All Fields] AND ((“volition”[MeSH Terms] OR “volition”[All Fields]) AND (“psychology”[Subheading] OR “psychology”[All Fields] OR “psychology”[MeSH Terms]))’
  - search_filter + ‘”perception”[MeSH Terms] AND agency[All Fields]’

3.2 Obtaining and organizing PDFs based on pubmed searches.

NCBI provides a toolset called Entrez[6] for programatically obtaining information from its databases, including pubmed. The Biopython project provides an interface to Entrez, Bio.Entrez[7].

We supply our search terms to Bio.Entrez which calls Entrez to obtain lists of pubmed ids for articles matching our search criteria. Pubmed Central is the best resource for downloading full PDFs. A custom python script we wrote therefore takes the pubmed ids to build URLs to obtain PDFs based on pubmed or pubmed central ids. Using the URLs the PDFs are programatically downloaded and named according to the pubmed id. Some of the URLs fail in this process, usually due to less open full text access to certain journals, in which case the PDFs are obtained manually. Addition of other text to the PDF names is done to facilitate organization and collection of annotations according to annotator, search term or task, and upon completion of annotation, the annotators apply a label indicating the success or other status of their annotation work.

3.3 Extraction of annotations from PDFs.

When working at the whole document level, annotators’ append the task status to their PDF. For example, the suffix ‘Annotated’ confirms the article

[6] http://www.ncbi.nlm.nih.gov/books/NBK3836/

[7] http://biopython.org/DIST/docs/tutorial/Tutorial.html#htoc109

appears to be classed appropriate to the given RDoC construct we attempted to recover using our pubmed search; ‘Irrelevant’ or ‘Misclassified’ indicates the article does not belong to this RDoC construct.

Some tasks requested annotators to markup individual sentences or phrases in PDF abstracts, targeting those sentences or phrases that appeared to be relevant or identified the RDoC construct. The annotator’s used common PDF readers to highlight these relevant passages.

To extract these annotations, we wrote a custom python script making use of the python pdfminer[8] package. This code is fragile and best suited for a semi-automated approach in which the results of pdfminer are examined side by side with the plain text abstract obtained from another source (medic in our case). The plain text abstract is edited to include the annotations from pdfminer. We used ‘{’ and ‘}’ to indicate the annotations in plain text. The semi-automation simply refers to our ability to script such that we open both the pdfminer extracted abstract and the medic abstract side by side, similarly processed into separate sentences by spaCy to give us a visual clue of where the ‘transfer’ or ‘editing towards’ annotation needs to occur. Our pdfminer based script simply does not extract highlights from some of our pdfs, thus requiring a fully manual transfer of the annotations to the plain text version.

3.3.1 Guidance on alternative to pdfminer for highlight extraction from PDFs.

The pdfminer python package is poorly documented and our solution using it to extract highlighted text is fragile and sometimes fails. Despite this, it was the one best Python library enabling parsing of highlights from PDFs that we could find after extensive searches. There are some wrappers to faciliate use of pdfminer that are better documented, but we were unsuccessful using these.

The best alternatives to pdfminer are likely non-python alternatives and / or commercial services. We explored several GUI based shareware or freeware packages but they often had the same issues we encountered with pdfminer. Itext[9] is a Java library that appears to be the best available tool. Elements such as annotations in the PDF can be extracted by programming in the pdfreader class in itext. Itext version < 5 has the most open licensing.

Finally, if the annotators made their annotations to plain text version of the abstracts or were instructed to copy paste from the PDFs into a plain

[8] https://github.com/euske/pdfminer

[9] http://itextpdf.com/

text file many of these issues would not exist.

3.4 Alternatives sources of pubmed article training data.

For abstract, author, title, MeSH and some other metadata, python's medic package is likely the best solution for obtaining training data.

For full text, programmatic access via XML and OAI or Entrez is feasible. But it would probably be easier to work with data obtained as bulk downloads from Pubmed central. Full text documents under open access[10]. Particularly I think it is this source that has done OCR on additional PDFs for which XML full text is not available. Additionally, it provides figures and supplementary data in some cases.

4 Some results of our Deepdive apps.

All results are for lemmatized bags of words and a simple training and prediction that an abstract belongs to a targetted category.

4.1 Using a small number of abstracts to train for our target category.

DeepDive was trained with 145 abstracts manually confirmed by annotators to reflect the RDoC construct 'Auditory Perception'. To represent non-Auditory Perception targets, we used 1000 non-target abstracts. In addition, we had a 2nd set of 1000 articles that should also be predicted as 'Auditory Perception' and we call that the 'prediction' set.

DeepDive splits the 1146 abstracts into training (75% of abstracts) and test (25% of abstracts) sets. We explore the results of the predicted model DeepDive created.

4.1.1 non-target class for training is disease

Figures 1-3 show results of training DeepDive with a small set of Auditory Perception abstracts and a larger set of non-target (disease) abstracts.

Disease is distant enough from Auditory Perception that a decent model is built. As we see later, the limiting factor appears to be the small training set or the particular attributes of this small training set such that the model is not generalized to all Auditory Perception abstracts.

[10] http://www.ncbi.nlm.nih.gov/pmc/tools/openftlist/

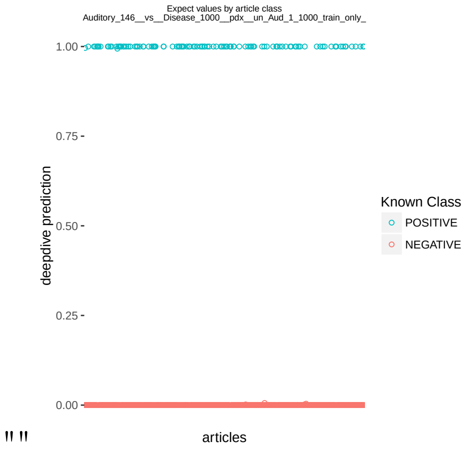


Figure 1: Training set for small set of Auditory Perception as target (blue), vs larger set of disease as non-target (orange). The training resulted in a model that cleanly split the target and non-target abstracts. There are approximate 750 orange marks on bottom access indicating they are very unlikely to be Auditory Perception. The X axis are all abstracts in the training set (as either target or non-target) in random order.

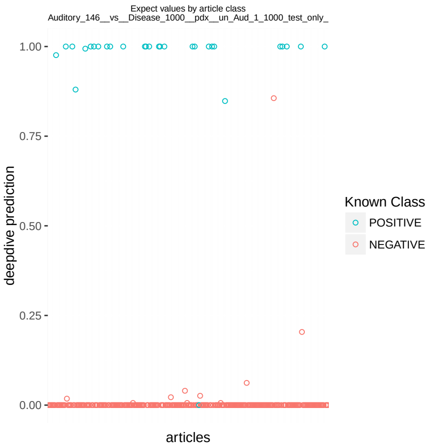
4.1.2 non-target class for training is arousal.

Figures 4-6 show resuts of training the same small set of Auditory Perception abstracts. But now the non-target abstracts are from the RDoC topic Arousal.

These topics are much closer to each other compared to Auditory Perception and Disease. The model reflects this as the training shows some abstracts include one abstract with features of each the target and non-target class. Correct categorization of the 1000 abstracts in the prediction set is poor, with many not being predicted correctly. The model is not very generalizable.

Tables 1 and 2 show the confusion matrix and statistics on the confusion matrix for this model using a small auditory perception training set.

Despite the quite good statistics (with the excption of sensitivity) on the test set, the model does not generalize well to another set of 1000 auditory perception abstracts (the prediction set). Most of the prediction set are inaccurately categorized along with the non-target classes.



" "

Figure 2: Test set for small set of Auditory Perception as target vs larger set of disease as non-target. The test set shows some miscategorized abstracts. The model may be overfit on the training set or there is something strange about some of these test abstracts.

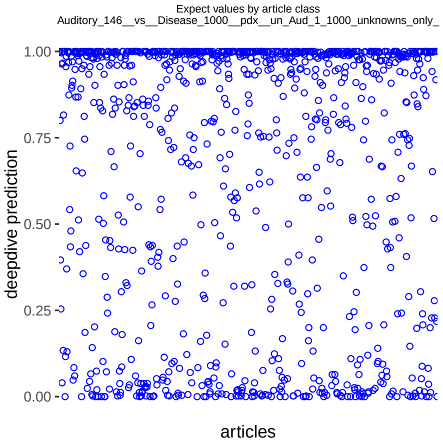
4.2 Using a larger number of abstracts to train for our target category.

DeepDive was trained with 1000 abstracts highly likely to reflect the RDoC construct ‘Auditory Perception’ (they were retrieved using the MeSH term ‘Auditory Perception’). To represent non-Auditory Perception targets, we used the same 1000 non-target abstracts as above. In addition, we had a 2nd set of 1000 articles that should also be predicted as ‘Auditory Perception’ and we call that the ‘prediction’ set.

DeepDive splits the 2000 abstracts into training (75% of abstracts) and

Table 1: The confusion matrix for the test subset of the 146 auditory and 1000 arousal abstracts. (Should total approx 1/4 of the 1146 total abstracts). See Figure 5.

	known not-auditory (i.e. arousal)	known auditory
Predicted not-auditory	251	11
Predicted Auditory	3	32



"" ""

Figure 3: Prediction set for small set of Auditory Perception as target vs larger set of disease as non-target.. Most of the 1000 abstracts in the prediction set are correctly categorized, but quite a few are missed.

test (25% of abstracts) sets. We explore the results of the predicted model DeepDive created.

4.2.1 non-target class for training is disease

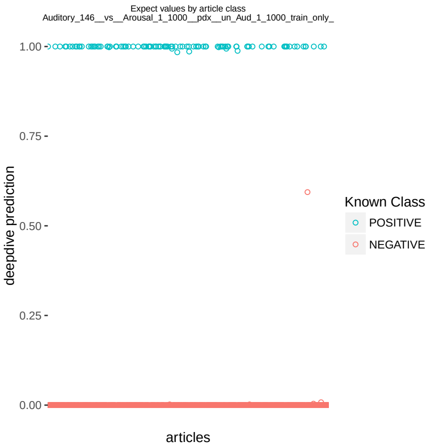
Figures 7-9 show results of training DeepDive with a larger set of Auditory Perception abstracts.

Disease is distant enough from Auditory Perception that a quite good model is built. The improvement in the model is likely due to the much larger target training set. However, we cannot entirely rule out that the training set here is better simply because it is more representative.

4.3 non-target class for training is arousal

Figures 10-12 show results for a model trained and tested on 1000 Auditory Perception abstracts as the target class and 1000 Arousal abstracts as the non-target class.

Comparing Table 6 to table 4, most of the stats are better when the non-target category is distant, indicating that even with quite a few more abstracts in the training data, DeepDive could likely use even more to better distinguish Auditory Perception and Arousal. Alternatively it may suggest that there is a subset of Auditory Perception articles that should not be



" "

Figure 4: Training set for small set of Auditory Perception (blue) as target, arousal as non-target (orange). The training resulted in a model that cleanly split the target and non-target abstracts. There are approximate 750 orange marks on bottom access indicating they are very unlikely to be Auditory Perception. The X axis are all abstracts in the training set (as either target or non-target) in random order.

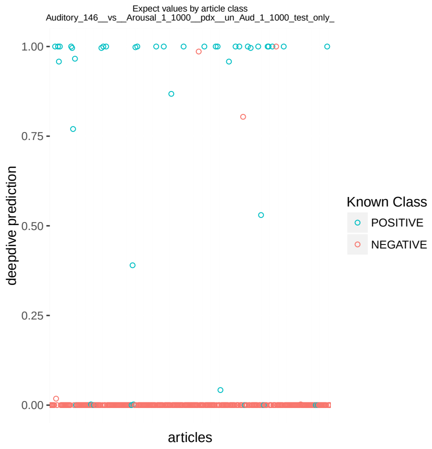
distinguished from Arousal. Inspecting the miscategorized articles and the weakly categorized articles would help distinguish these possibilities. Likewise running a mulitnomial model and including additional non-Auditory perception articles as other target classes might help.

4.4 Cross validation on DeepDive models.

DeepDive lacked the facilities to do easy cross validation. We therefore designed an approach to do this ourselves making use of the confusion matrix statistics reporting and our ability to create multiple deepdive apps simply by recycling an app template.

The standard for cross validation is K-fold cross validation in which data is split into a certain number of subsets and the model is trained and tested once with each of the small subsets removed. (So if you were doing 10-fold cross validation, the data is split into 10 chunks and the model run with 9 chunks and tested with the 10th.). That way, each piece of data is left out at least once.

With quite a bit of effort, we could have designed for this approach using for deepdive but it was far easier to make do with an alternative cross



'' ''

Figure 5: Test set for small set of Auditory Perception as target, arousal as non-target. The test set shows some miscategorized abstracts. The model may be overfit on the training set or there is something strange about some of these test abstracts. Note, as is more clear in the confusion matrix, several 'Auditory Perception' abstracts are not identified as such and are blue circles 0.00 but are difficult to see due to the number of orange circles at 0.00.

validation in which random subsets of data were sampled essentially with replacement compared to the K-fold method. While not as stringent as K-fold cross validation, it should nonetheless indicate whether the DeepDive models are excessively responding to certain subsets of the data.

The confusion matrix statistics were quite reproducible for the small Auditory Perception as target and Arousal as non-target experiment, confirming especially that our observation of the undesirable low sensitivity was reproducible.

p The sensitivity for the larger Auditory Perception was better and less variable (Figure 15) compared to the smaller Auditory Perception set (Figure 14) as seen from the cross-validation of that experiment. But internal accuracy and specificy of the model declined, despite it seeming more generalizable (Figure 6 vs. Figure 12). We would want to pursue in more detail whether this was due to the more random general selection of the larger Auditory Perception set (broader pubmed query) or the balance or unbalance in number of abstracts between target and non-target classes. Clearly additional experiments are needed to address class balance issues, to further investigate the tradeof between generalizability and model performance and manual inspection of articles to establish whether the poorer model per-

Table 2: Accuracy and other classification measures corresponding the confusion matrix in Table 1 (small set of Auditory Perception vs. Arousal as non-targets.) See Figure 5. Note sensitivity is quite low.

statistic	
Accuracy	0.952
AccuracyLower	0.922
AccuracyUpper	0.973
Sensitivity	0.744
Specificity	0.988
Pos Pred Value	0.914
Neg Pred Value	0.958
Prevalence	0.144
Detection Rate	0.107
Detection Prevalence	0.117
Balanced Accuracy	0.866

formance is due to their being more ‘Auditory Perception’ articles in the broader search that are also ‘Arousal’.

5 Discussion and Future Directions.

5.1 Potential for DeepDive to categorize or label articles with high accuracy.

We demonstrated that increasing the training set size quite a bit increased the performance of the model on external ‘prediction sets’ quite a bit. In

Table 3: Confusion matrix for the test abstracts Auditory Perception using the larger training set vs disease as non-target. Should total to approx 1/4 of the full 2000 target and non-target set of abstracts. (See Figure 8).

	known not-auditory (i.e. disease)	known auditory
Predicted not-auditory	228	12
Predicted Auditory	3	226

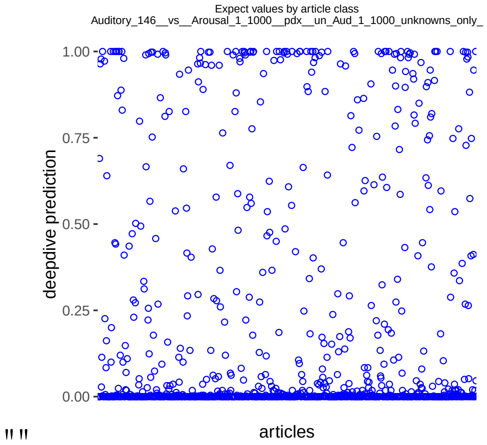


Figure 6: Prediction set for Auditory Perception with Arousal as the non-target set. Most of the 1000 abstracts are incorrectly classified.

other words, we obtained models that were much better generalized with larger data sets. We note this did not always show up best in the statistics in the confusion matrix, pointing to the need to ensure one has a good representative sample of articles for any RDoC construct in order to avoid model overfitting.

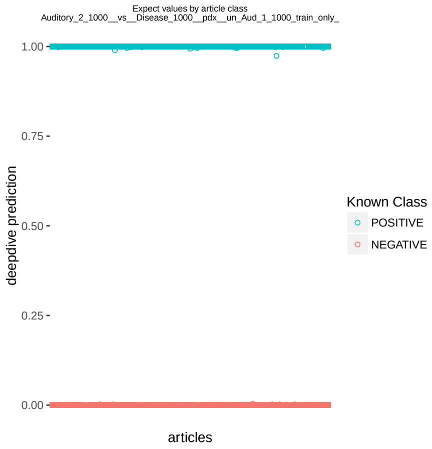
Therefore, we suggest for future development that continued tests of non-target and target data sizes and composition be explored and a variety of 'prediction sets' be used to test the models.

We would recommend at least 1000 training articles for each RDoC construct, if available.

5.2 Use of DeepDive vs other machine learning tools.

DeepDive presented specific challenges when we tried to employ common machine learning techniques to investigate the quality of its models. In some machine learning environments such ability to inspect and test variants of the model are extremely easy so this was a disappointment.

DeepDive continues to be under active development and we hope that they improve the system towards better exposing the statistics of model performance and enabling additional cross validation approaches. As the training sets grow in size, computational needs increase. It would be of interest to determine if deepdive becomes easier to recommend vs other tools under those circumstances.



" "

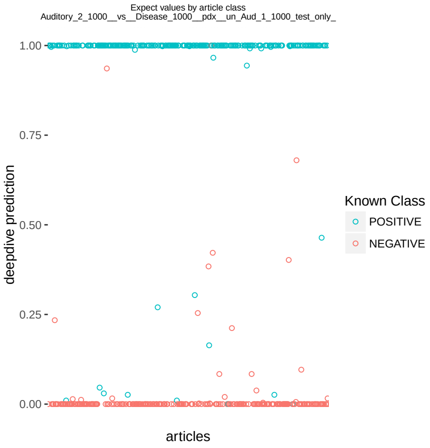
Figure 7: Training set for larger Auditory Perception target set (blue), disease as non-target (orange). The training resulted in a model that cleanly split the target and non-target abstracts. There are approximate 750 orange marks on bottom access indicating they are very unlikely to be Auditory Perception. The X axis are all abstracts in the training set (as either target or non-target) in random order.

If tools other than DeepDive were selected for future development, is is of interest to note that as far as we understand the algorithms used by DeepDive, these same algorithms of suitable substitutes are available in many other frameworks. For instance Gibbs Sampling, and alternatives to working with markov random are available in other toolsets.

5.3 Further feature development.

We only scratched the surface with respect to features to use for RDoC categorization. With larger datasets, we believe biwords and n-grams may become more useful features for inclusion in our apps. In addition, given the tight relationship between certain MeSH terms and RDoC constructs, MeSH terms should be added as features. Other groups have used a author-abstract model in which authorship is an additional tell-tale feature that identifies the subject domain.

Finally, we have focused primarily on bags of words from full abstracts, yet a large amount of literature is available as full text. An exploration of model performance using full text would be desirable, especially if it could reduce the number of articles that one must score.



" "

Figure 8: Test set for larger Auditory Perception target set, disease as non-target. The test set shows pretty good categorization of abstracts. Compare to figure 5 and note the lack of disease abstracts on the top, indicating most of the non-target set is correctly not identified as highly likely to be about Auditory Perception. Also note there are many more Auditory Perception abstracts compared to figure 5 so we shouldn't be too alarmed at the number of positive abstracts in the mid-portions of the y-axis.

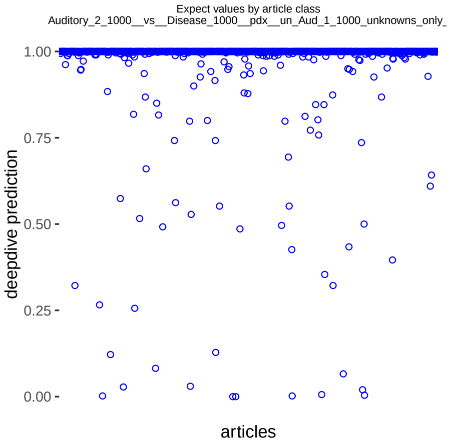
5.4 Issues in creating the confusion matrices.

To create the binary classification for the confusion matrix, we split data as predicted Auditory Perception and predicted non-Auditory perception at 0.50. This is quite arbitrary. After all, in the case of Figure 5, several of these abstracts seem to show some characteristic of Auditory Perception.

It would be wise to examine the abstracts in the middle portions of the distribution to confirm they share features of both arousal and Auditory Perception. In fact several of them do share text we would expect to be common to either RDoC category.

Depending on our goals with categorization, we may prefer to keep track of the degree of the prediction rather than convert it into a binary decision. We do expect that many of the RDoC constructs share a great number of characteristics and that discovering the features underlying these overlaps can be a fruitful area for exploration.

With more advanced models, we could do multinomial labeling, such that we get a prediction of 0.7 Auditory perception, 0.3 Arousal 0.1 Sleep. This may be more useful than a simplistic decision to assign such an abstract only



""

Figure 9: Prediction set larger for Auditory Perception target set, disease as non-target. Most of the 1000 abstracts in the prediction set are correctly categorized. Only a few abstracts show no 'Auditory Perception' characteristics, and a few more show partial expectation that they are Auditory Perception. In this case we should manually confirm that these abstracts near 0.00 really are primarily about auditory perception as opposed to primarily about disease or another topic.

to the class with the highest prediction.

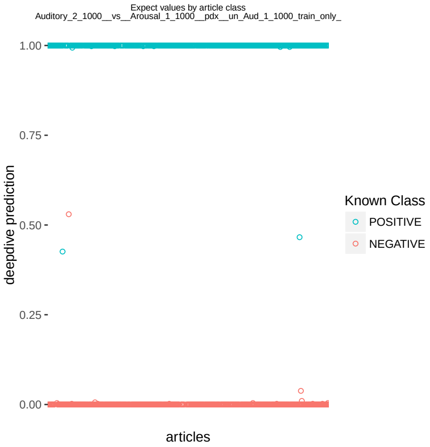
— end of text figures and tables only from hereon —

Table 4: Accuracy and other classification measures corresponding the confusion matrix in Table 3 (larger set of Auditory Perception vs. disease as non-targets.) (See Figure 8).

statistic	
Accuracy	0.968
AccuracyLower	0.947
AccuracyUpper	0.981
Sensitivity	0.949
Specificity	0.987
Pos Pred Value	0.986
Neg Pred Value	0.95
Prevalence	0.507
Detection Rate	0.481
Detection Prevalence	0.488
Balanced Accuracy	0.968

Table 5: Confusion matrix for the test abstracts Auditory Perception using the larger training set vs arousal as non-target. Should total to approx 1/4 of the full 2000 target and non-target set of abstracts. (See Figure 11).

	known not-auditory	known auditory
	(i.e. disease)	
Predicted not-auditory	220	30
Predicted Auditory	33	215

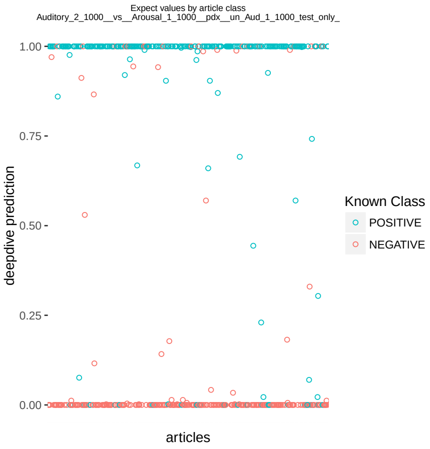


" "

Figure 10: Training set for larger Auditory Perception target set (blue), arousal as non-target (orange). The training resulted in a model that cleanly split the target and non-target abstracts. There are approximate 750 orange marks on bottom access indicating they are very unlikely to be Auditory Perception. The X axis are all abstracts in the training set (as either target or non-target) in random order.

Table 6: Accuracy and other classification measures corresponding the confusion matrix in Table 5 (larger set of Auditory Perception vs. Arousal as non-targets.) (See Figure 11).

statistic	
Accuracy	0.873
AccuracyLower	0.841
AccuracyUpper	0.901
Sensitivity	0.877
Specificity	0.869
Pos Pred Value	0.866
Neg Pred Value	0.88
Prevalence	0.491
Detection Rate	0.431
Detection Prevalence	0.497
Balanced Accuracy	0.873



" "

Figure 11: Test set for larger Auditory Perception target set, arousal as non-target. The test set shows pretty good categorization of abstracts. Compare to figure 5 and note the lack of disease abstracts on the top, indicating most of the non-target set is correctly not identified as highly likely to be about Auditory Perception. Also note there are many more Auditory Perception abstracts compared to figure 5 so we shouldn't be too alarmed at the number o positive abstracts in the mid-portions of the y-axis.

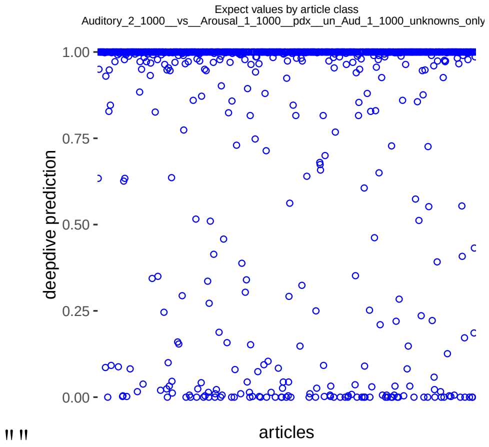


Figure 12: Prediction set for larger Auditory Perception target set, arousal as non-target. Most of the 1000 abstracts in the prediction set are correctly categorized. Only a few abstracts show no 'Auditory Perception' characteristics, and a few more show partial expectation that they are Auditory Perception. In this case we should manually confirm that these abstracts near 0.00 really are primarily about auditory perception as opposed to primarily about arousal or another topic.

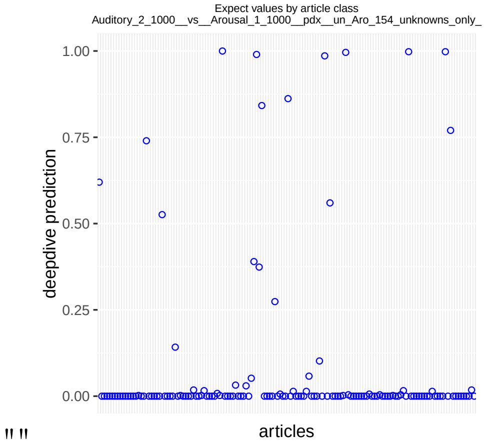
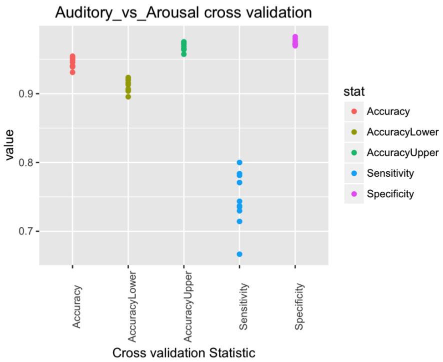
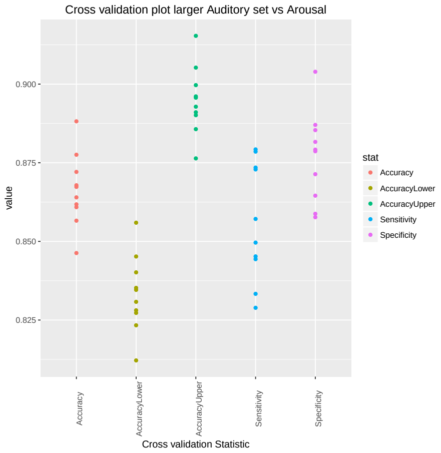


Figure 13: Prediction set for larger Auditory Perception target set, arousal as non-target. Most of the 154 abstracts in the prediction set are correctly NOT categorized as Auditory Perception, instead they are in the non-target class which is correct as the non-target class is arousal.



" "

Figure 14: Cross validation of the small Auditory Perception as target vs. Arousal as non-target. Note the reproducible poor sensitivity.



" "

Figure 15: Cross validation of the larger Auditory Perception as target vs. Arousal as non-target. Note sensitivity is better vs. the smaller Auditory Perception results (Figure 11). However specificity and overall accuracy declined. Note the narrower y-axis range vs. Figure 14.