SAND2019-6676C

Sandia National Laboratories

ILLINOIS

Automated Recognition of Dual Use Publications

Jacob Caswell, Christina Ting, Christopher Cuellar, Jonathan Bisila, Mengfei Ho, Brenda Wilson, Kelsey Cairns, Jerilyn Timlin

Abstract

With the growing acknowledgment that research in various biological fields could be used for harm as well as for good, publishers are increasingly concerned about the dissemination of research with dual use potential. To date, journals must rely on manual review by experts with specific knowledge of dual use concerns to identify if a submitted draft contains material that could be used maliciously. To alleviate this burden, we have taken steps towards an automated system capable of recommending documents for expert review.

Our approach is to train and test different machine learning models on publications labeled by subject matter experts from academia. Our models can detect with 90% accuracy publications containing content that is of dual use concern. Interestingly, keywords provided by our experts were not the strongest indicators of dual use. Future work includes explainability indicators to help experts resolve flagged publications and expand their knowledge of dual use topics.

Dataset

UIUC has provided 272 PubMed abstracts, manually labeled with respect to expert-deemed dual use research concern (DURC) level:

- Not of concern
- At least Potential DURC

Data format example:

	Label	Text
0	y0	(Title-Abstract from paper 0)...
1	y1	(Title-Abstract from paper 1)...
2	y1	(Title-Abstract from paper 2)...
3	y0	(Title-Abstract from paper 3)...
...		

Also provided: list of DURC-related keywords. It is believed that these keywords were used to identify DURC publications.

Methods

Feature extraction: We use a 'bag of words' technique to convert raw abstracts (text) to a matrix of features (numerical) representing normalized word occurrences. Rather than raw counts, we use the term-frequency inverse document frequency (TFIDF).

$$\mathrm{tfidf}(t, d, D) = \mathrm{tf}(t, d) \times \mathrm{idf}(t, D)$$

Frequency a term t appears in a document, d

Offset by the number of documents, d in the corpus D containing the term t

Supervised Machine Learning (ML): We use two algorithms, Linear SVC which attempts to find a line that optimally separates labelled points (Figure 1), and Random Forest, which asks an ensemble of decision trees to vote on a point's predicted label given patterns discovered in its features.

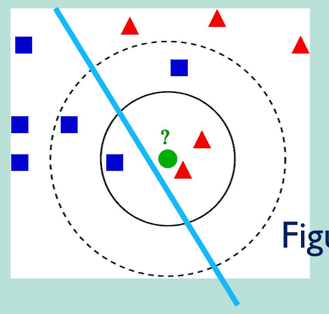
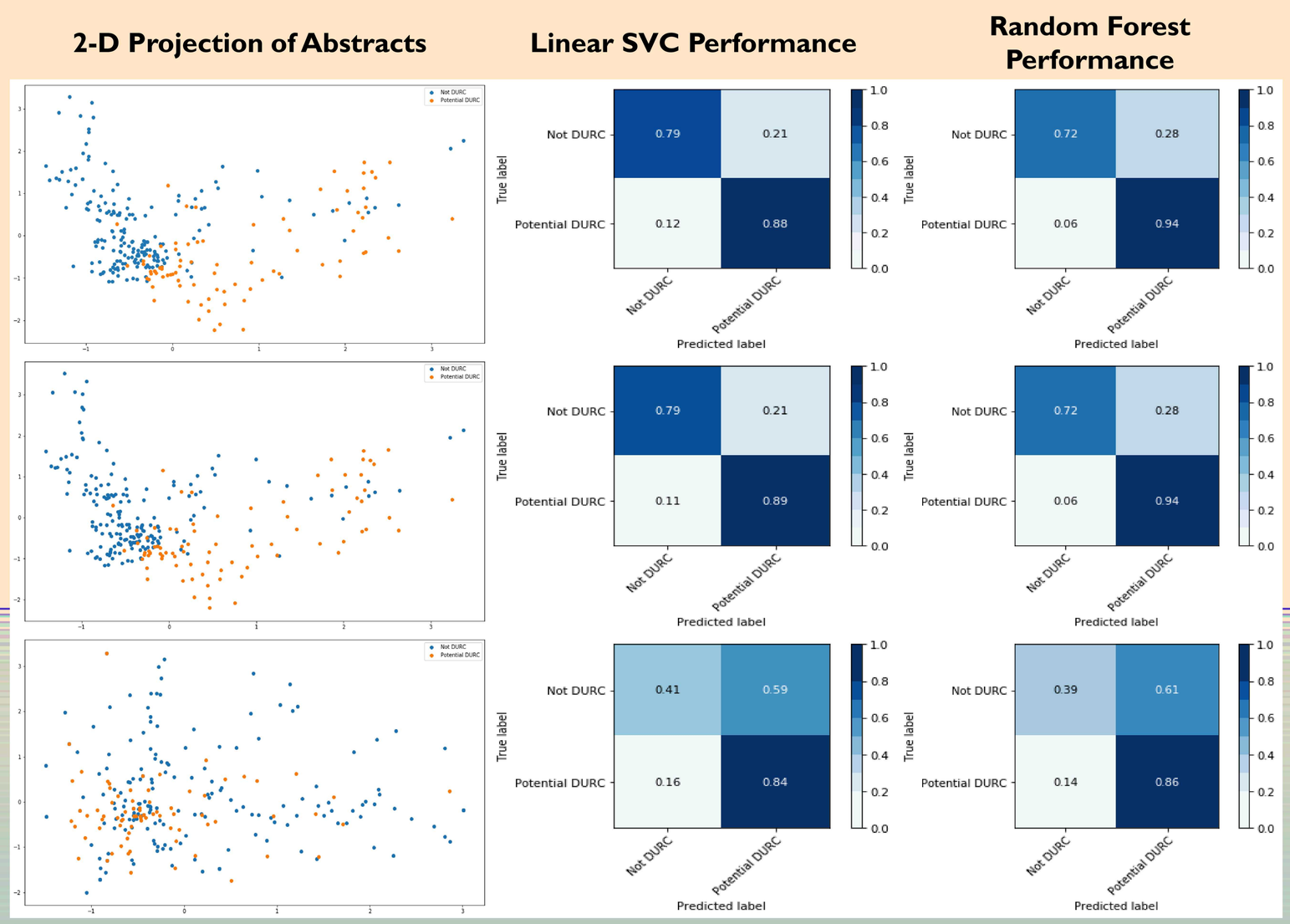


Figure 1

Dimensionality Reduction: It is difficult to visualize the space of our data, where the dimensionality of our vectors is the number of unique words in our abstracts. It is of interest to reduce the dimensionality to 2d for visualization purposes. There are many ways to reduce the dimensionality; we chose the spectral embedding approach for finding a low dimensional representation.

Results



Observations:

- Though complete separability is not achievable, we can achieve ~90% detection of DURC in our dataset.
- This detection rate, however, comes at the cost of ~¼ false positive rate.
- Non-keywords are necessary both for maximizing classification rate, and significantly reducing false positives.

Conclusions

- We have obtained a dataset of PubMed abstracts, manually labeled by their potential for dual use research concern.
- We have shown that machine learning algorithms may be used to classify documents based on 'bag-of-words' features indicative of dual use.
- Our models automatically detected DURC documents with 90% accuracy in a 5-fold validation test.
- Our analysis indicates that, for our data, non-keyword text is essential for classifying DURC level.
- Therefore, ongoing work includes identifying dual use indicators that can help experts resolve flagged publications and expand their knowledge of dual use topics.
- Future work will focus on measuring generalizability and transferability of our model to larger data sets.

U.S. DEPARTMENT OF ENERGY
NNSA National Nuclear Security Administration

Sandia National Laboratories is a multimission laboratory managed and operated by National Technology & Engineering Solutions of Sandia, LLC, a wholly owned subsidiary of Honeywell International Inc., for the U.S. Department of Energy's National Nuclear Security Administration under contract DE-NA0003525.
SAND No. _____

Sandia National Laboratories