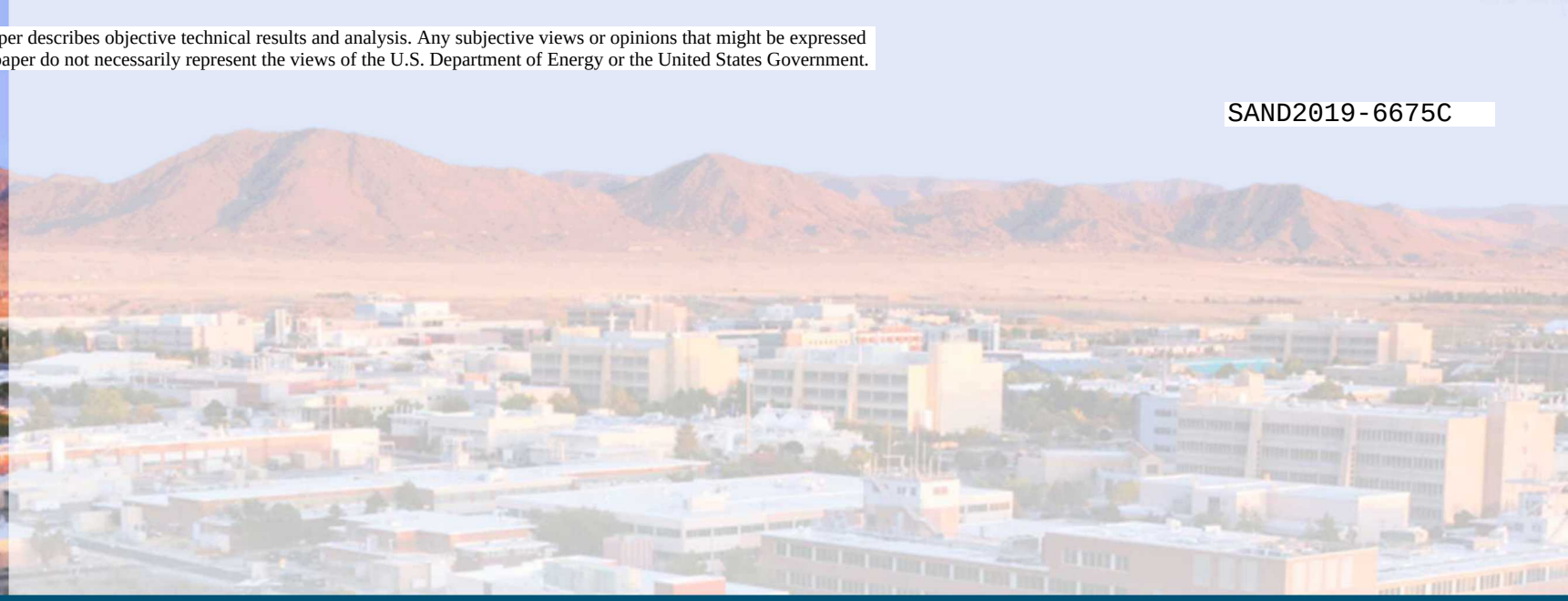
SAND2019-6675C



Automated Recognition of Dual Use Publications

Jacob Caswell[1], Christina Ting[1], Christopher Cuellar[1], Jonathan Bisila[1], Mengfei Ho[2], Brenda Wilson[2], Kelsey Cairns[1], Jerilyn Timlin[1]

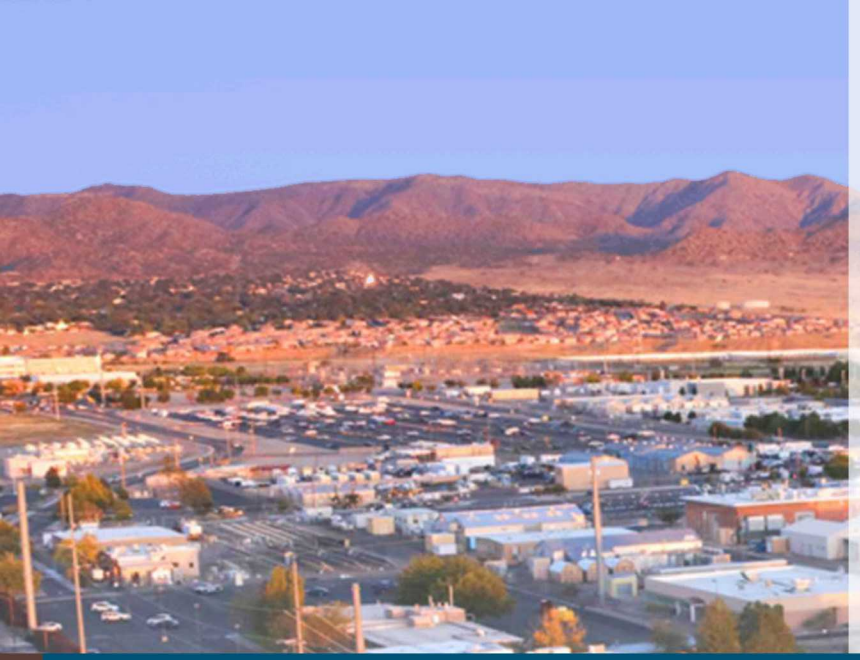
[1] Sandia National Laboratories, New Mexico

[2] University of Illinois, Urbana Champaign (UIUC)

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. _____

Motivation

- Research in biological fields has great potential to do good as well as harm.
- Unlike other fields with potential for Dual-Use Research Concerns (DURC) (e.g., nuclear engineering) the line that separates good and harm is not as well-defined.
- Journals and grants now rely on manual content review to identify material that could be misused.
- The community would benefit from automated recommendation methods to expedite the review process.
- Machine Learning (ML) models may be able to help automate the discovery of potentially DURC manuscripts.



Data and Methods

Data

- UIUC has provided 272 PubMed abstracts manually labeled by a Subject Matter Expert (SME) according to the following levels of DURC:
  - Not DURC
  - Potential DURC
  - DURC
- Labels are binarized into:
  - Not DURC
  - At least potentially 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 used by SME to identify DURC publications.

Representation of data

- For analysis, the PubMed articles' titles and abstracts are converted into a numerical vector representing normalized word occurrences.
- Each entry in the vector is a word's term-frequency inverse document frequency (TFIDF)
- TFIDF highlights words that are most signature to a given document.

$$\text{tfidf}(t, d, D) = \text{tf}(t, d) * \text{idf}(t, D)$$

- Where:
  - **tf** corresponds to the frequency that term ***t*** appears in document ***d***.
  - **idf** offsets the term frequency according to how many documents in the corpus ***D*** contain the term ***t***.

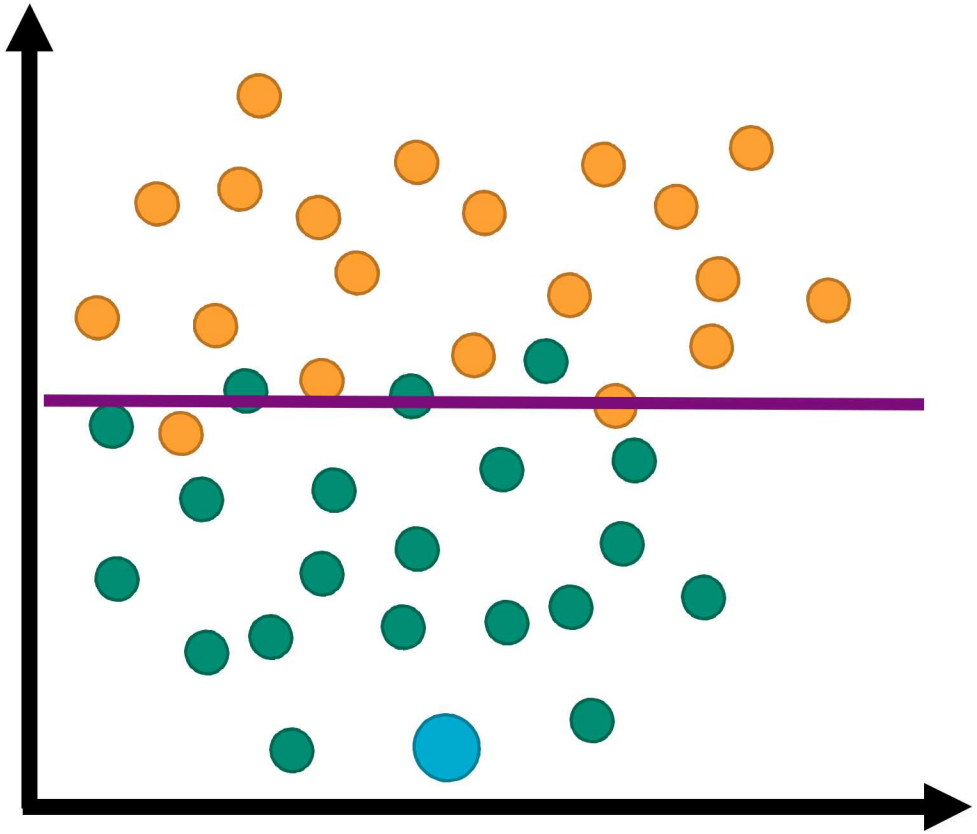
Robertson, S.E. and Spärck Jones, K. 1976. Relevance weighting of search terms. J. Amer. Soc. Inf. Sci. 27, 3, 129–146.

Data preprocessing

- We preprocess the titles and abstracts accordingly:
  - Abstracts and titles are all set to lower case.
  - Abstracts and titles are tokenized.
  - Stop words are removed.
  - Words with document frequency of >0.7 are removed.
- We compare this baseline to two additional preprocessing methods:
  1. Remove all tokens that are not keywords identified by a SME.
  2. Remove all tokens that are keywords identified by a SME.
- These methods are used as control to identify which tokens are most significant to the classifier's decision boundary.

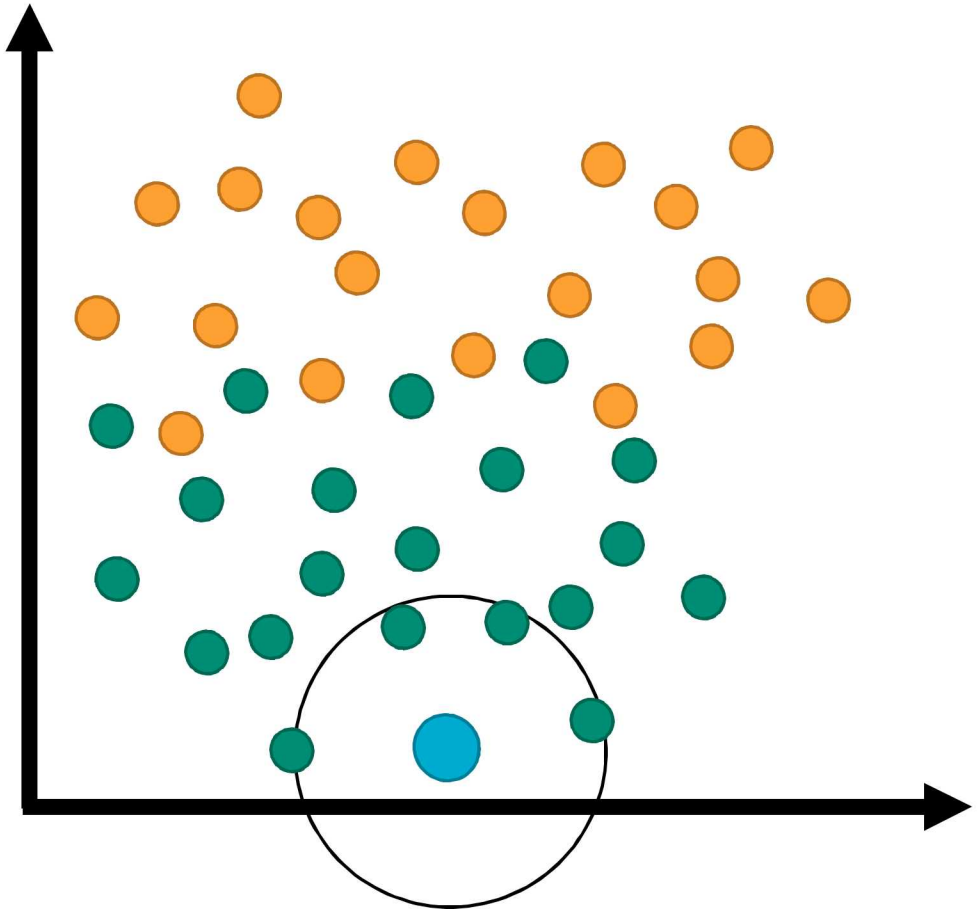
Classifiers

Classifiers – Linear Support Vector Classifier (Linear SVC)



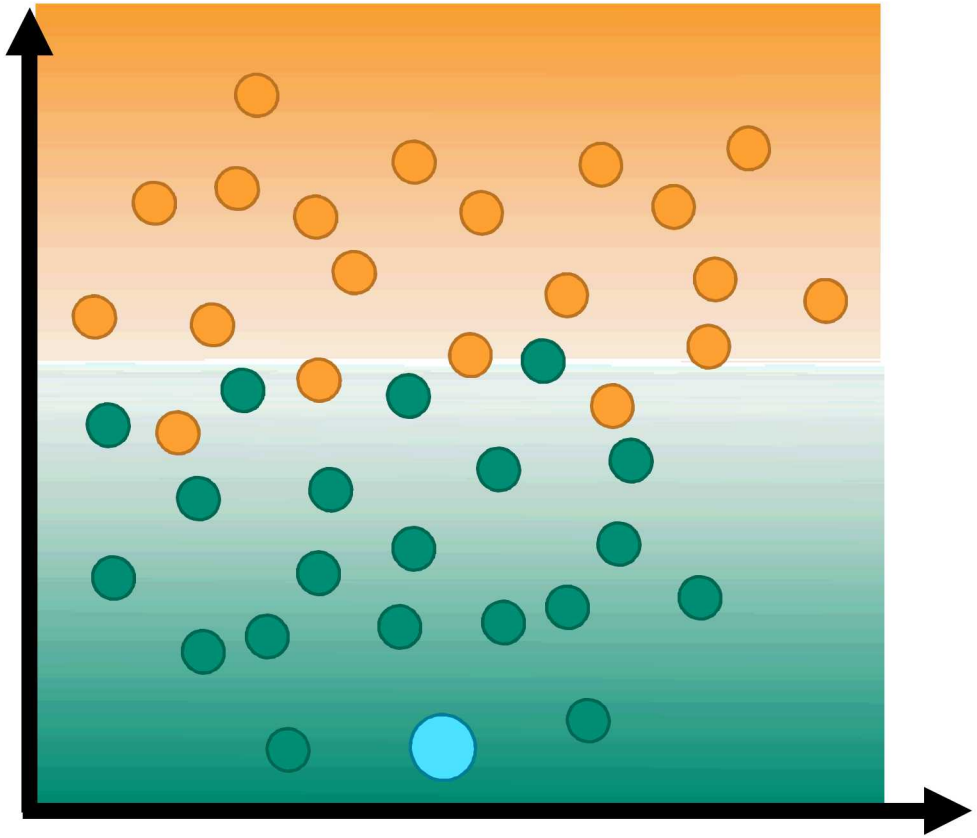
- Identify a line in the data's space that optimally separates the labels.
- For higher dimensions of data, it uses a hyperplane.
- Resistant to overfitting
- Parameters we used:
  - C=0.5
  - Non DURC weight: 0.25
  - Potential DURC weight: 0.75

Classifiers – K-Nearest Neighbors (KNN)



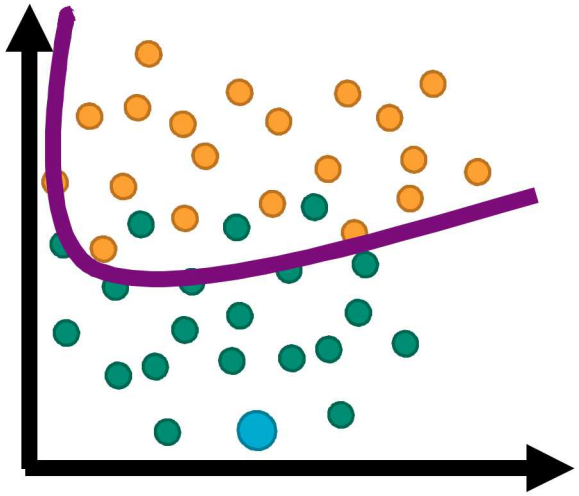
- A point's label is determined by majority label of its K-nearest neighbors.
- Doesn't need to draw a decision boundary.
- Performs best when data are not well mixed.
- Parameters we used:
  - K=15

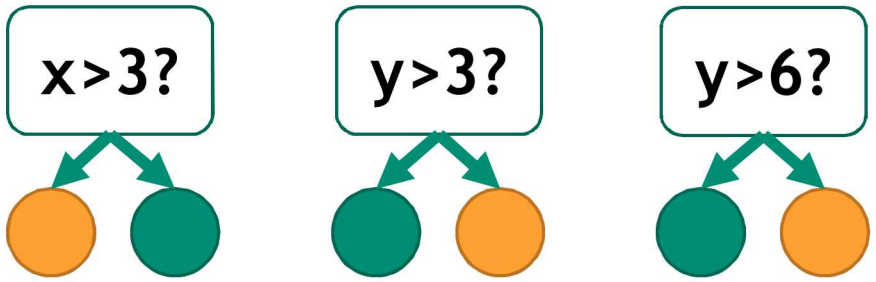
Classifiers – Naïve Bayes



- Train a probability model for label prediction using data as "evidence."
- Assumes that different features are independent (a naïve assumption).
- Also doesn't need to draw an explicit decision boundary.

Classifiers – Random Forest (RF)

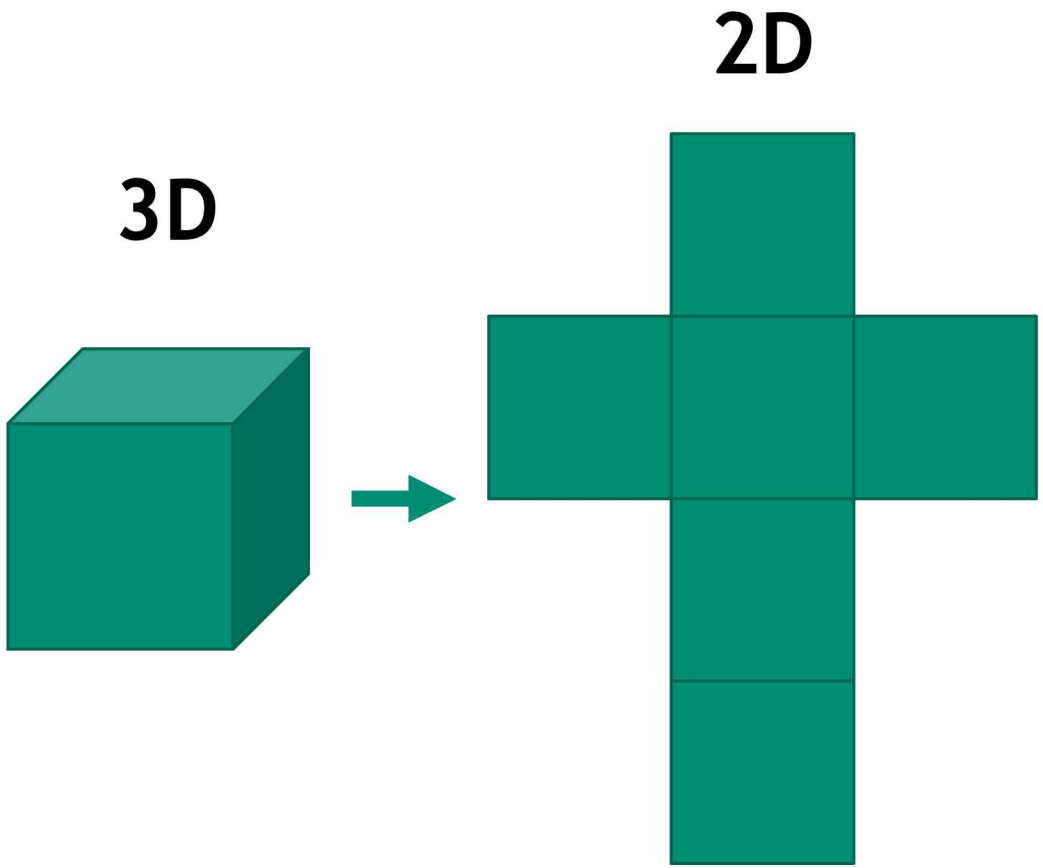




- Use many, small decision trees (a forest) that are all trained on different parts of the data.
- Majority vote for determining new datapoint labels
- Capable of more complex decision boundaries than Linear SVC, but also not susceptible to uncontrolled overfitting.
- Can struggle with class imbalance.
- Parameters we used:
  - n_estimators=100
  - max_depth=2
  - Non-DURC weight: 0.25
  - Potential DURC weight: 0.75

Dimensionality Reduction

Dimensionality Reduction – Spectral Embedding

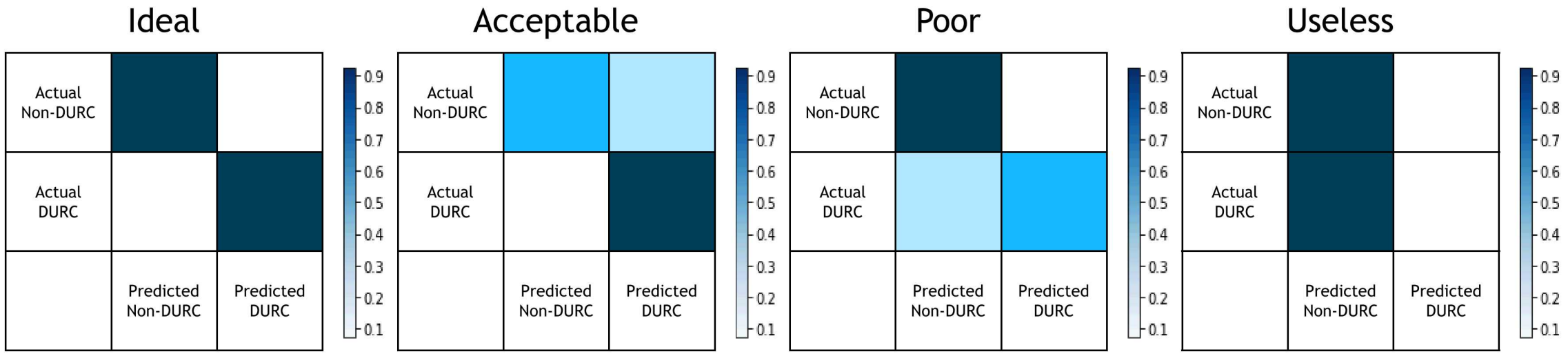


- Allows us to map high dimensional data into a 2D space.
- Preserves local distances and is nonlinear.
- May yield better classification performance while being more computationally performant.

Performance / Validation

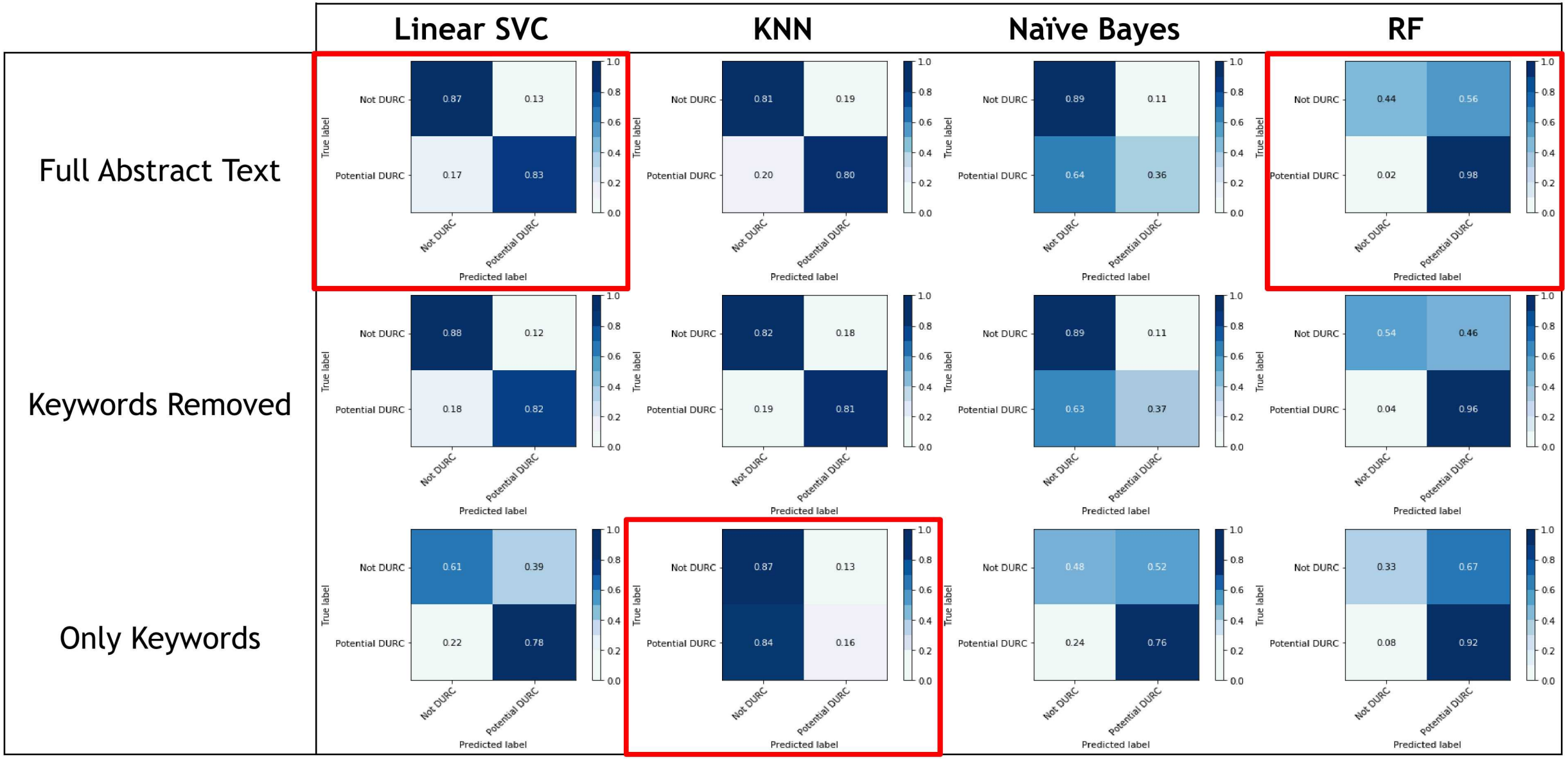
Performance Measurement

- Classifier performance is averaged over stratified 5-fold cross validation.
- Performance is summarized using a Confusion Matrix.
- We wish to find as many DURC papers as possible, constrained by also reducing the number of false positives a manual reviewer would have to read.
- Ideally, each class accuracy should be maximized.
- Pragmatically, we favor DURC class accuracy over (and empirically, at the expense of) non-DURC class accuracy.
- DURC labelled data were weighed more heavily than non-DURC for Linear SVC and Random Forest to emphasize recall.

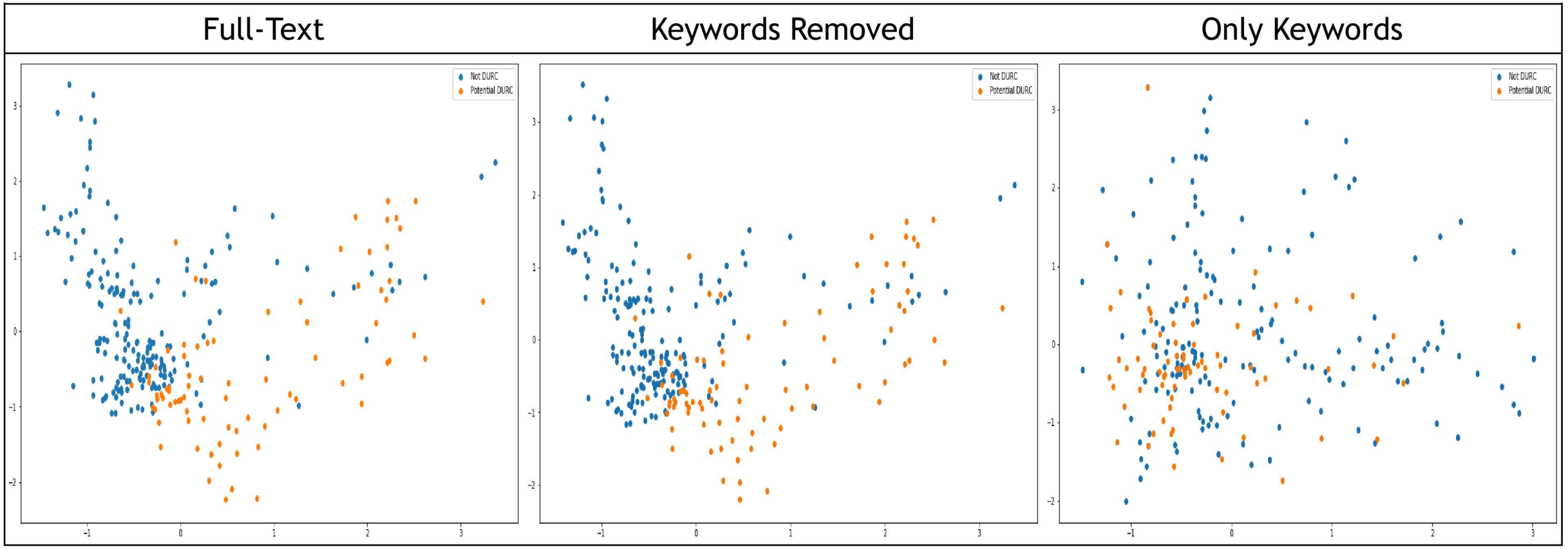


Results

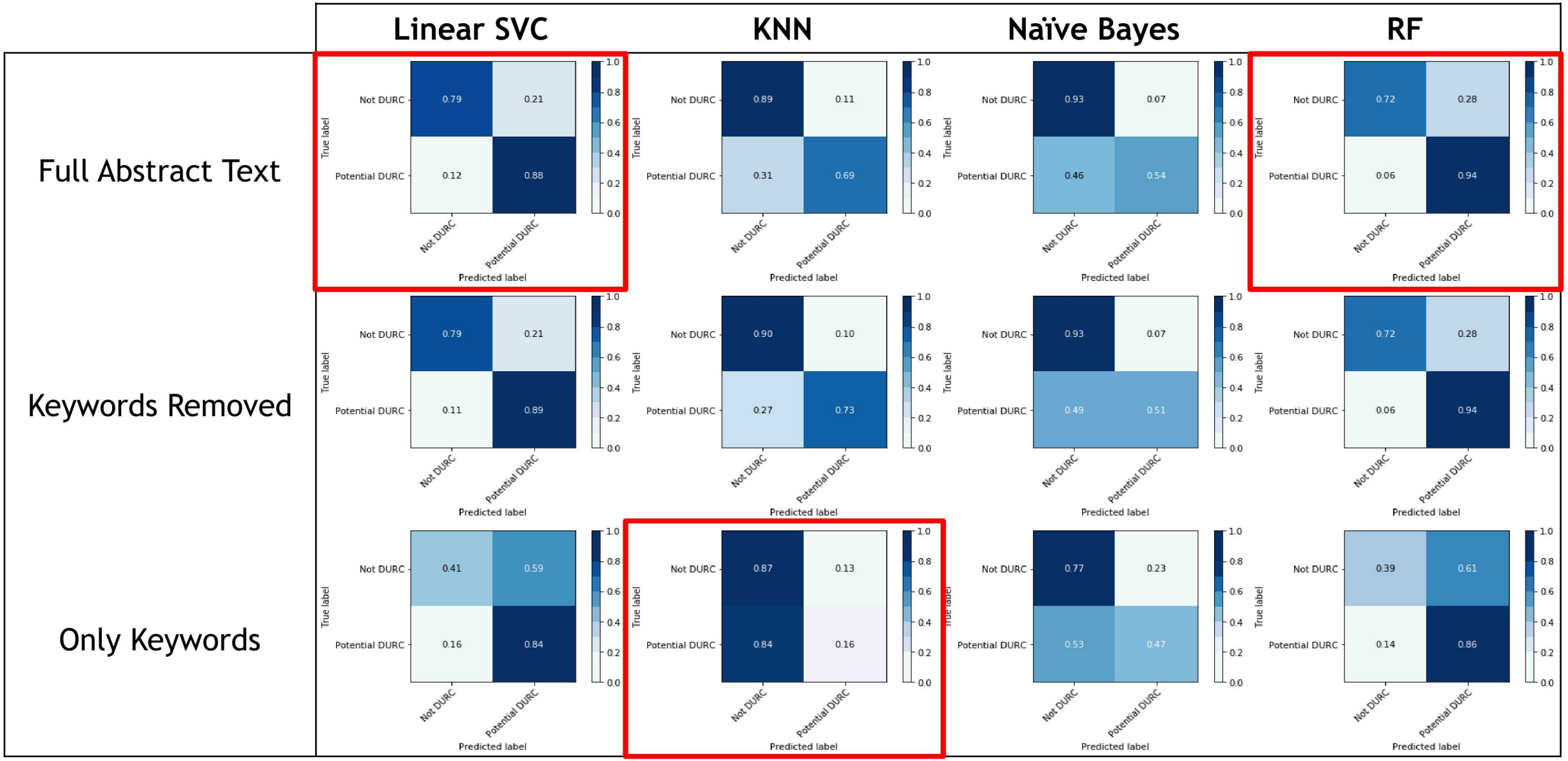
Classification Performance on Full-Dimensionality Data

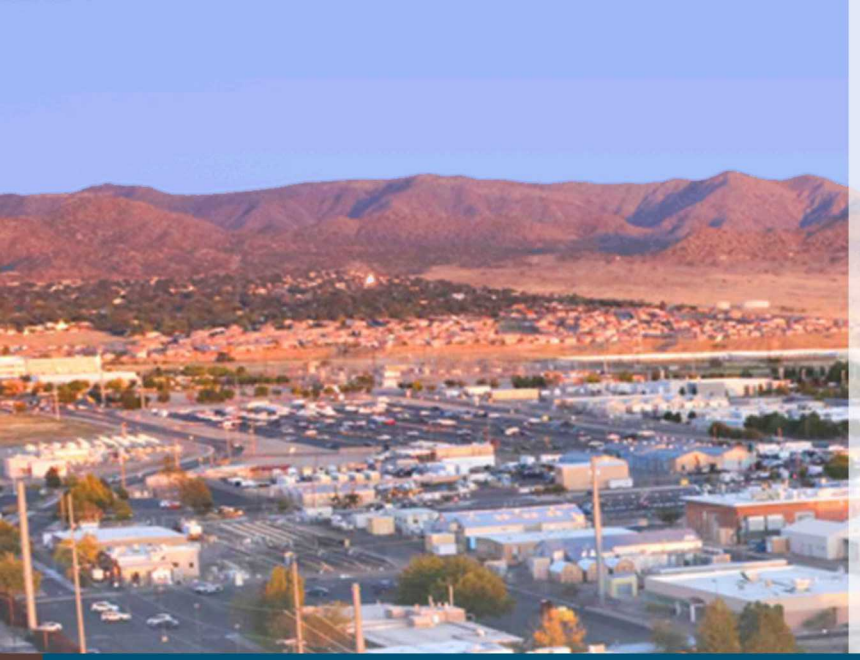


Spectral Embedding Dimensionality Reduction



Classification Performance on Spectral Embedding-Reduced Data





Conclusions

Principal Conclusions

- Both Linear SVC and Random Forest were able to achieve ~90% detection rates of DURC in our data set.
- KNN performed best on full dimensionality data, but suffered when only using keywords.
- Naïve Bayes performed worse than the other classifiers, and favored the more common class (non-DURC). It may be that its naïve assumption does not hold on this data.
- Given that DURC and non-DURC papers were not completely separable when projected, the experiments that had the highest detection rate tended to have a higher false-positive rate.
- Dimensionality reduction improved the performance of Random Forest considerably, significantly reducing the rate of false positives.
- After dimensionality reduction, false positive rates were less than 30%, significantly down-selecting the number of documents required for manual curation.
- Contextual words (Non-keywords) were critical for classification performance.

Future Work

- Our work was done on a dataset of limited size. Thus, additional work will be needed to ensure the performance generalizes to larger and more varied datasets.
- More work must be done to better explain what factors contributed to the high performance of classification in our dataset.

Questions?

Additional Slides

Provided Keywords

highly
specificity
optimize
wide
better
irreversible
cytopathogenicity
screen
mosaic
translocate
strategic
pathogenicity
cytotoxicity
increase
functional
different
immunogenicity
regulate
pathogen
versatile
combine
penetrate

optimal
amplify
robust
robustness
efficiency
functionality
catalytic activity
deliver
combinatorial
potentiate
superior
improve
effective
bioagent
multivalent
solubility
more
induce
elevated
selectivity
switch
elevate

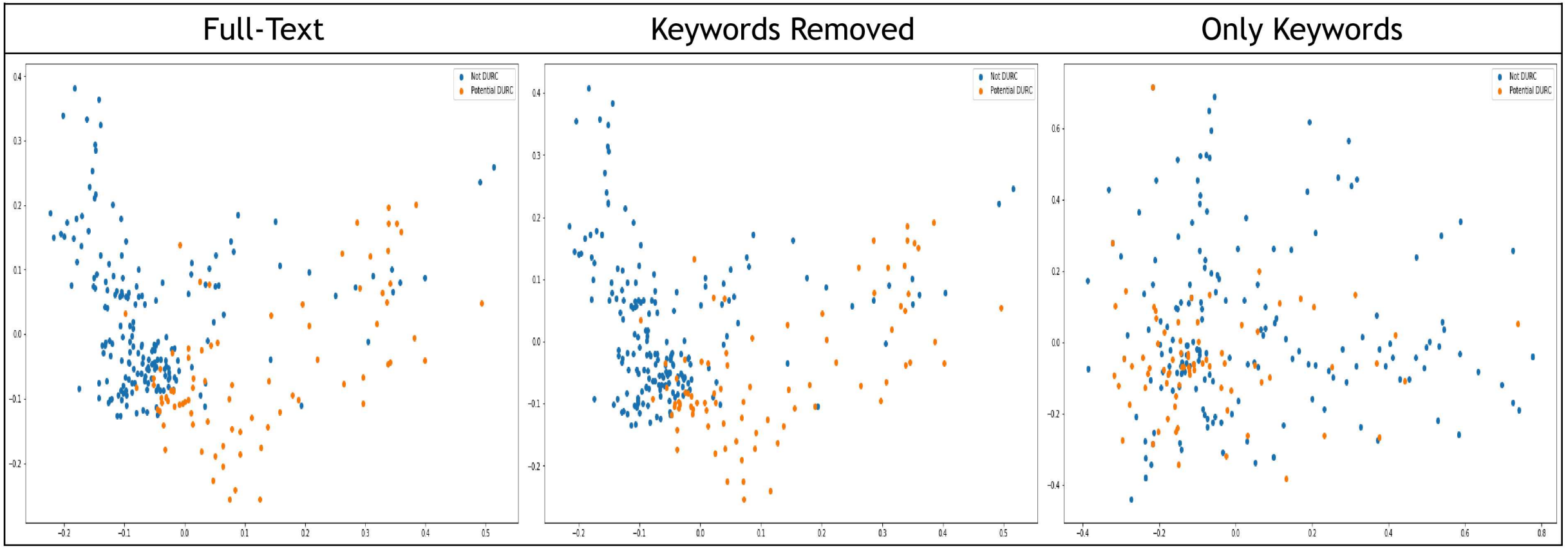
refine
diversity
introduce
novel
toxicity
virulence factor
resistance
potentiation
genetic
cytotoxic
biological activity
enhance
recombine
exogenous
host specificity
toxic
virulent
efficacy
virulence
chimeric
upregulate
stable

efficient
stability
new
biosynthesize
change
maximize
genetically
lethality
exchange
engineer
broad
secrete
edit
modular
design
swap
diversify
synthesize
resilient
alter
bacteria
selective

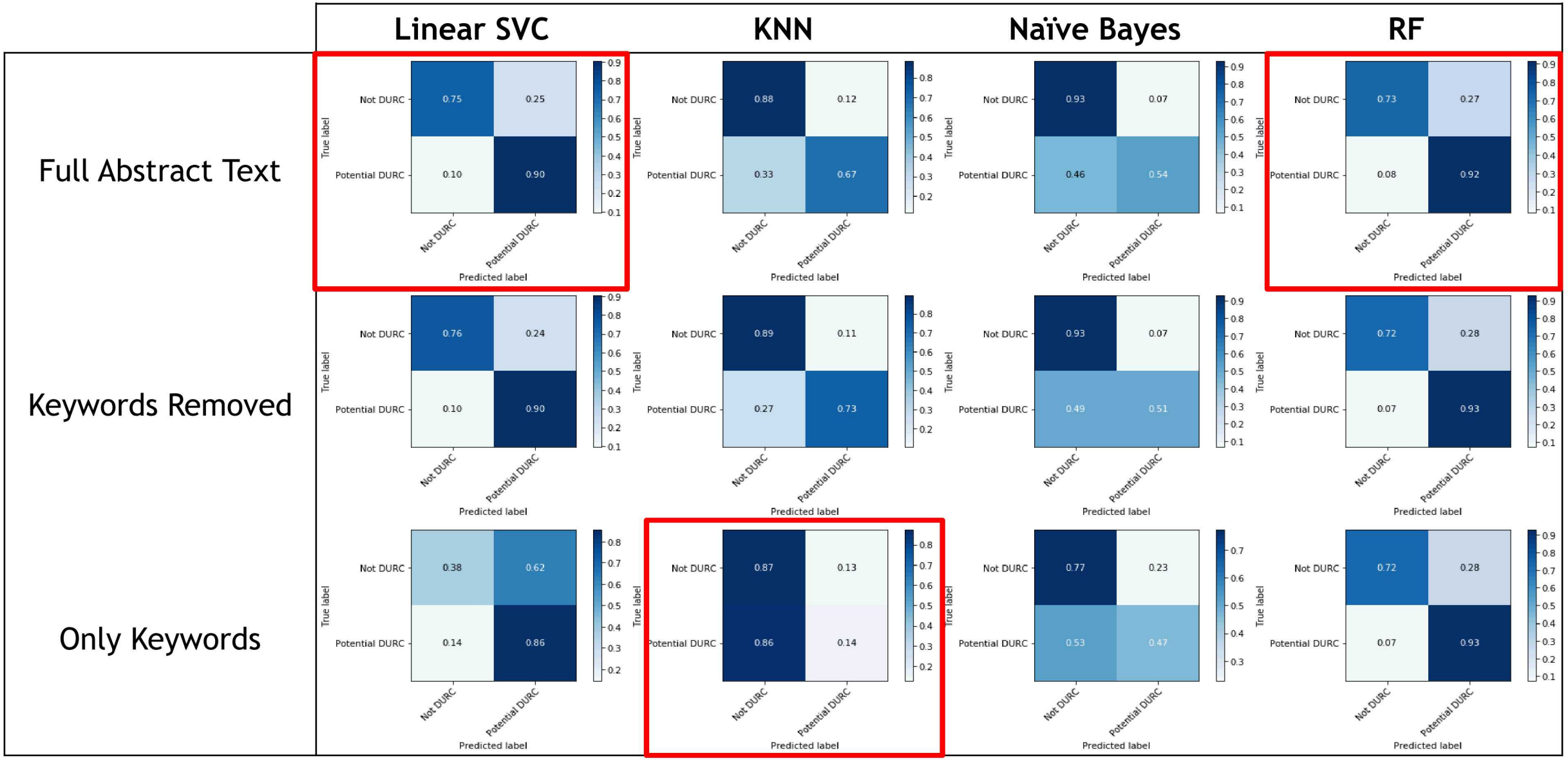
cytomodulin
modulate
virion
titer
build
lethal
recombinant
reversible
encapsulate
pathogenesis
reconstruct
pure
bind
coexpress
modularity
toxin
traceless
endocytose
cellular activity
conjugate
redesign
potent

antibiotic resistance
replace
fuse
virus
potency
expand
synthetic
mutate
vary
enzymatic activity
modify
specific
construct
heterologous
versatility
infectivity
maximal
rapid
purer
higher
accelerate
broaden

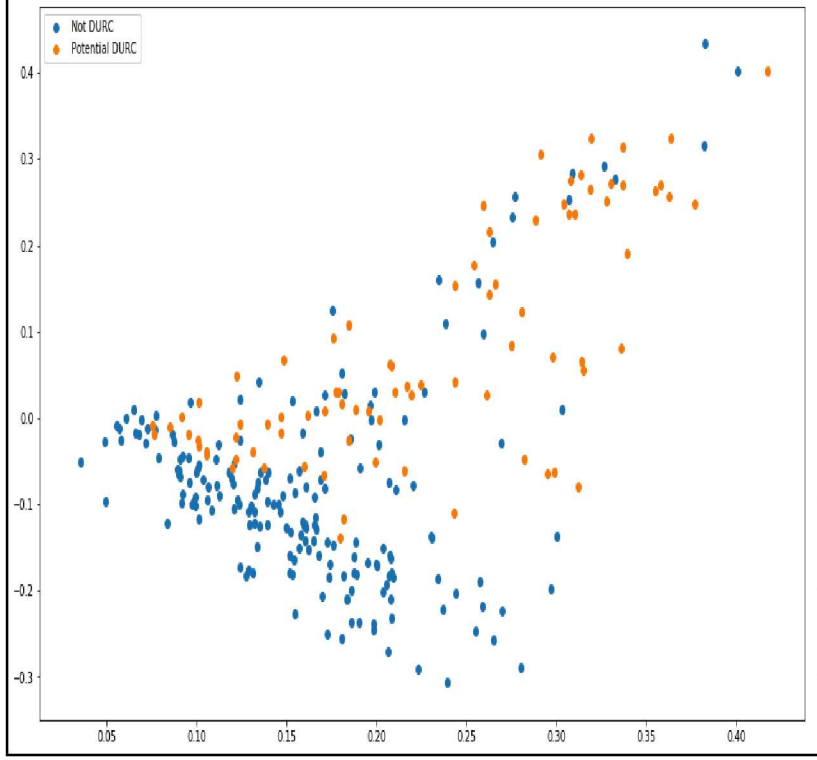
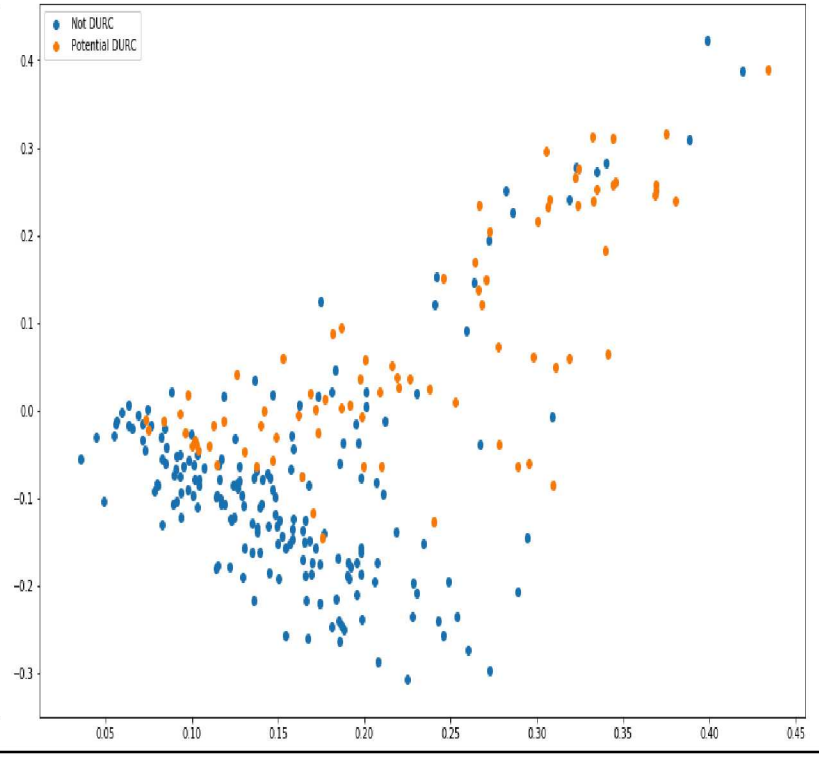
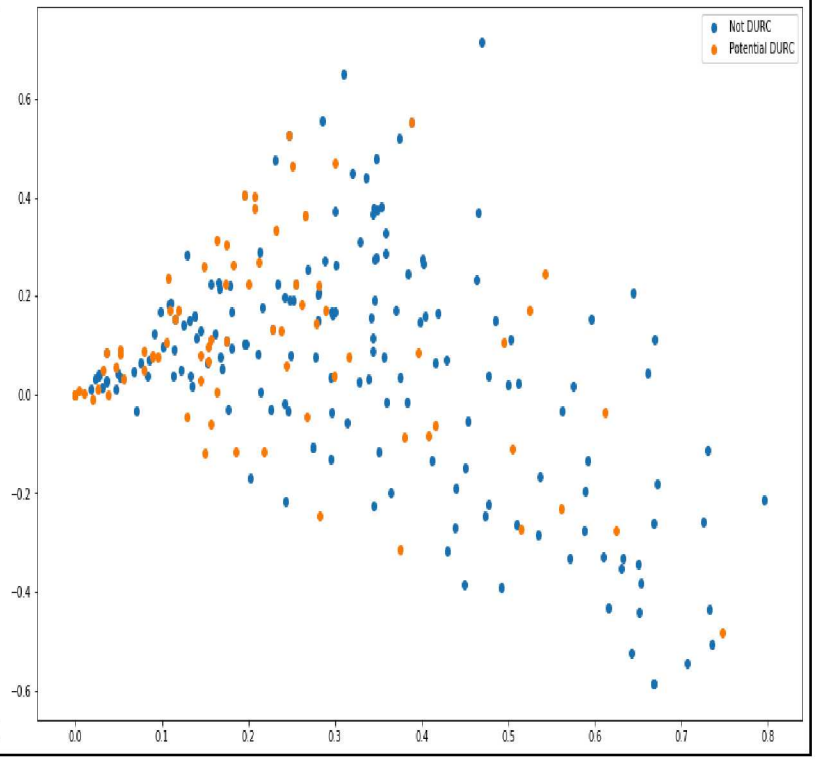
PCA Dimensionality Reduction



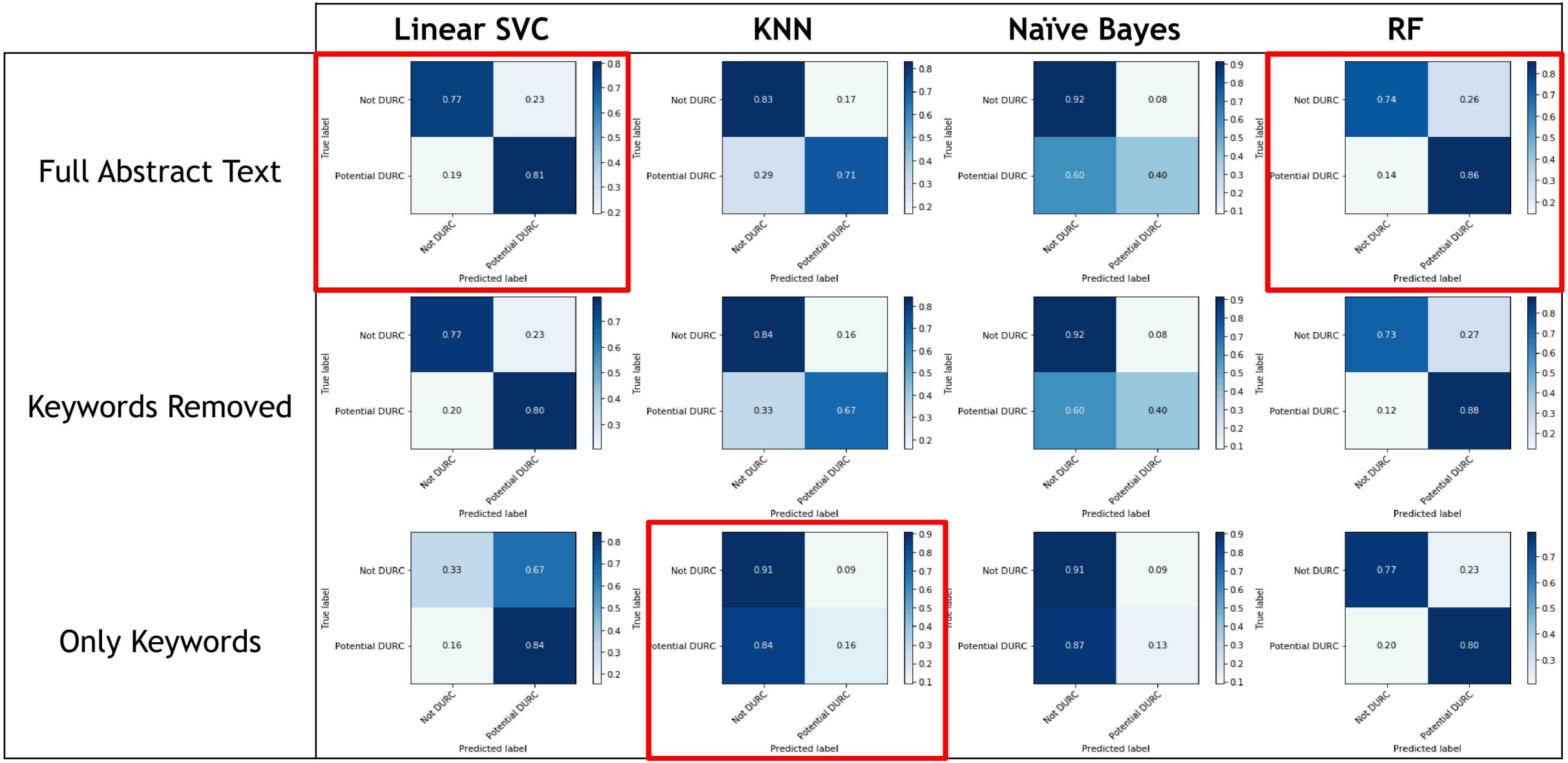
Classification Performance on PCA-Reduced Data



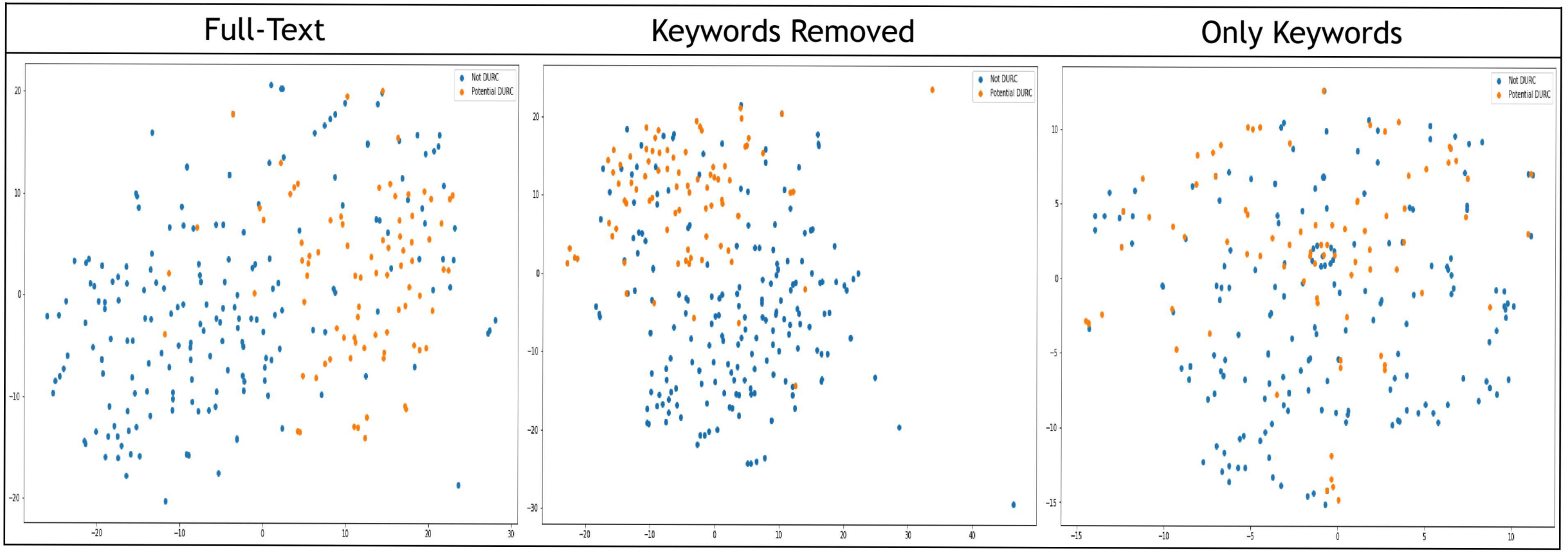
SVD Dimensionality Reduction

Full-Text	Keywords Removed	Only Keywords
		

Classification Performance on SVD-Reduced Data



TSNE Dimensionality Reduction



Classification Performance on TSNE-Reduced Data