REPORT DOCUMENTATION PAGE

PLEASE DO NOT RETURN YOUR FORM TO THE ABOVE ORGANIZATION.

1. REPORT DATE	2. REPORT TYPE	3. DATES COVERED	
		START DATE	END DATE
MARCH 2023	TECHNICAL PAPER	JANUARY 2019	MARCH 2020

4. TITLE AND SUBTITLE

A COMPARISON OF LANGUAGE REPRESENTATION MODELS ON SMALL TEXT CORPORA OF SCIENTIFIC AND TECHNICAL DOCUMENTS

5a. CONTRACT NUMBER	5b. GRANT NUMBER	5c. PROGRAM ELEMENT NUMBER
FA8750-18-C-0133	N/A	63260F/64536A

5d. PROJECT NUMBER	5e. TASK NUMBER	5f. WORK UNIT NUMBER
		R23G

6. AUTHOR(S)

Michael T. Gorczyca, Tavish M. McDonald, Thadeous A. Goodwyn, Peter F. David

7. PERFORMING ORGANIZATION NAME(S) AND ADDRESS(ES)

1. Air Force Research Laboratory/RIEA, 525 Brooks Rd, Rome NY 13441-4505
2. Decisive Analytics Corp., 1400 Crystal Dr, Ste 1400, Arlington, VA 22202-4153

8.PERFORMING ORGANIZATIONREPORT NUMBER

9. SPONSORING/MONITORING AGENCY NAME(S) AND ADDRESS(ES)	10.SPONSOR/MONITOR'S ACRONYM(S)	11.SPONSOR/MONITOR'S REPORT NUMBER(S)
Air Force Research Laboratory/RIEA 525 Brooks Road Rome NY 13441-4505	AFRL/RI	AFRL-RI-RS-TP-2023-020

12. DISTRIBUTION/AVAILABILITY STATEMENT

Approved for Public Release; Distribution Unlimited. PA# 88ABW-2020-1241; Date Cleared: 05 April 2020

13. **SUPPLEMENTARY NOTES**

 Proceedings Volume 11413, Artificial Intelligence and Machine Learning for Multi-Domain Operations Applications II; 11413IT. This work was funded in whole or in part by Department of the Air Force contract number FA8750-18-C-0133.

14. ABSTRACT

Text mining for the identification of emerging technology is becoming increasingly important as the number of scientific and technical documents grows. However, algorithms for developing text mining models require a large amount of training data, which carries heavy costs associated with data annotation and model development. The need for avoiding these associated costs has in part motivated recent work in text mining, which indicate value in leveraging language representation models (LRMs) on domain-specific text corpora for domain-specific tasks. However, these results are demonstrated predominantly on large text corpora, which do not address concerns associated with the ability of LRMs to transfer to domains where training data may be scarce. Due to this, we benchmarked the performance of LRMs on identifying quantities and units of measure from text when the number of training samples is small.

15. SUBJECT TERMS

Artificial intelligence, machine learning, multi-domain operations, text mining

16. SECURITY CLASSIFICATION OF:			17. LIMITATION OF ABSTRACT	18. NUMBER OF PAGES
a. REPORT	b. ABSTRACT	C. THIS PAGE		
U	U	U	SAR	10

19a. NAME OF RESPONSIBLE PERSON	19b. PHONE NUMBER *(Include area code)*
DANIEL DASKIEWICH	N/A

PREVIOUS EDITION IS OBSOLETE.

STANDARD FORM 298 (REV. 5/2020)

Prescribed by ANSI Std. Z39.18

A comparison of language representation models on small text corpora of scientific and technical documents

Michael T. Gorczyca[a*], Tavish M. McDonald[a], Thadeous A. Goodwyn[a], Peter F. David[a]

[a]Decisive Analytics Corporation, Arlington, VA 22202.

ABSTRACT

Text mining for the identification of emerging technology is becoming increasingly important as the number of scientific and technical documents grows. However, algorithms for developing text mining models require a large amount of training data, which carries heavy costs associated with data annotation and model development. The need for avoiding these associated costs has in part motivated recent work in text mining, which indicate value in leveraging language representation models (LRMs) on domain-specific text corpora for domain-specific tasks. However, these results are demonstrated predominantly on large text corpora, which do not address concerns associated with the ability of LRMs to transfer to domains where training data may be scarce. Due to this, we benchmarked the performance of LRMs on identifying quantities and units of measure from text when the number of training samples is small.

Keywords: language representation models, text mining, natural language processing, information extraction, small sample size

1. INTRODUCTION

Scientific and Technical (S&T) documents provide valuable information about advances in science, engineering accomplishments, and capabilities of novel systems. When S&T documents are available in large quantities, such as in online repositories that contain millions of articles, manual analysis of all content is impractical. As a consequence, scientists, analysts, and other knowledge workers who are confronted with large quantities of S&T data must rely on tools that mine these repositories for information.

Recent progress in mining text repositories has been driven by the adoption of deep learning models[1]. But, naively training such models often requires large amounts of annotated data, which is expensive to gather. This has in turn led to the development of language representation models (LRMs) such as BERT[2] and its S&T variant, SCIBERT[3]. LRMs are readily available, large scale deep learning models trained on large corpora to output contextualized embeddings for each input token; these embeddings can then be utilized as inputs for task-specific model development.

While the successes of these models have been widely reported, some researchers have demonstrated that these successes may be due to spurious statistical cues in the training dataset[4]. In this work, we provide some additional scrutiny to LRM performance on the task of identifying quantities and units of measure from text when the number of training samples is small relative to the number of samples available in standard NLP tasks. Specifically, we make the following contributions: (i) we compare the performance of models developed from token embeddings output by 50 dimensional GloVe embeddings (GloVe maps pre-defined words in a dictionary to context-agnostic token embedding)[5]; 768 dimensional BERT embeddings; and 768 dimensional SCIBERT embeddings. (ii) We feature engineer a set of auxiliary variables based on our own observations of the training data, and empirically demonstrate their value on this task. (iii) We provide basic importance measures to assess the effect of token embeddings and auxiliary variables on prediction outputs.

2. MATERIALS AND METHODS

2.1 Problem Overview

Scientific and technical documents contain a large amount of discussion concerning measurements and their numeric values. The meaning of these numeric values, or quantities, is central to understanding and building on scientific

literature, as they describe what phenomena are measured. Two difficult problems facing analysts – identifying relevant scientific information and searching for undiscovered connections and trends – may be considered problems partially solvable by extracting and evaluating quantities discussed across large collections of scientific literature.

A viable solution to these problems led to the development of a Quantity and Unit Extraction (QUE) capability that tags measurements in text. We note that a solution to this problem is nontrivial. For example, one may naively define a QUE capability that searches for instances of quantities and units of measure using a pre-defined dictionary (we note that we have constructed this capability, and will refer to this as the "QUE algorithm"). But, regular expressions are not robust to noise that may occur within text – unstructured text often contains text spans that may appear to be a quantity and unit of measure, but not actually express a quantity and unit of measure (we refer to such text spans as "false positives"). Some examples of these text spans are displayed in Figure 1. The problem of false positives motivated the development of a machine learning model that could robustly discriminate between false positives and true positives.

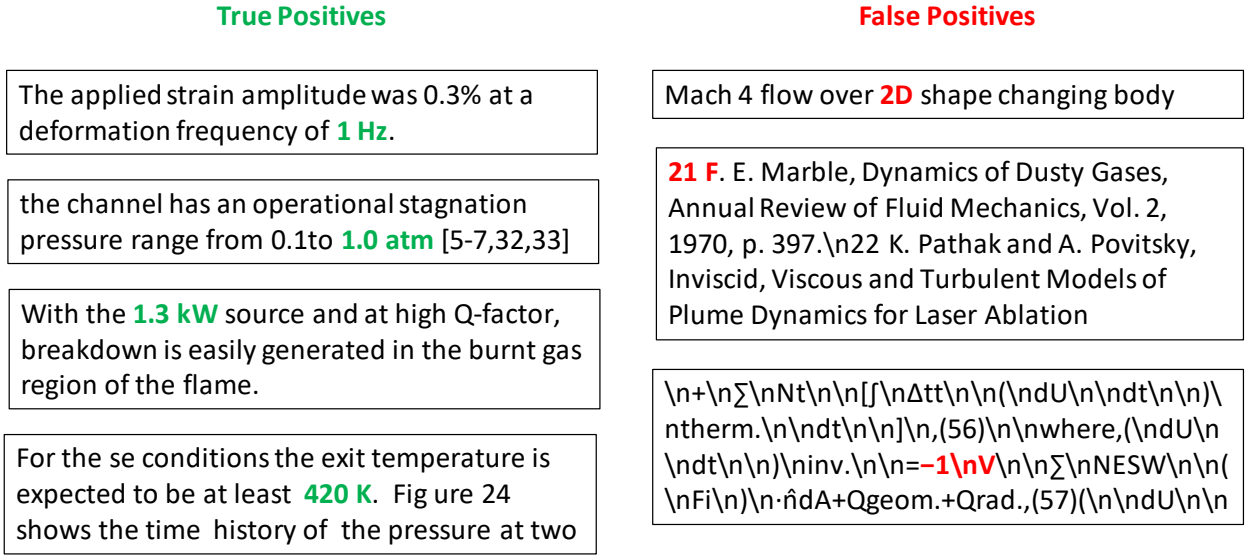

Figure 1. Numbers followed by unit abbreviation may not reflect an actual quantity and unit of measure. The highlighted numeric value denotes a quantity, which are followed by a potential unit of measure (True Positive example "Hz" denotes the unit of measure "hertz"; False Positive example "V" does not denote the unit of measure "Volts").

2.2 Dataset Overview

Three datasets were constructed for model development and assessment. Dataset construction consisted of searching for potential instances of quantities and units of measure in a repository of text documents, and extracting the corresponding text window. A text window consists of the quantity and unit measure as well as a moderate number of characters that appear before and after the identified quantity and unit of measure – in this study, a text window consisted of 60 characters. These text windows then received a score of "0" if the extracted quantity and unit of measure was a false positive, and a score of "1" otherwise; the scoring process was performed manually.

The first dataset was created from 708 Small Business Innovation Research (SBIR) solicitations, from which 33,022 instances of quantities and units of measures were "observed" by the QUE algorithm. The second dataset was created from every American Society for Materials (ASM) news article available from ASM International as of September 20th, 2019. There were 3,580 articles in total, from which 3,741 quantities and units of measure were detected. Lastly, the third dataset (the "Hypersonics" dataset) was created from 3,018 randomly selected quantities and units of measure extracted from 1,256 files retrieved from Defense Technical Information Center via a publicly accessible search for the term "hypersonics". A summary of each dataset is provided in Table 1.

	# of True Positives	# of False Positives	# of Distinct Units	# of Instances
SBIR	27,493	6,529	322	33,022
ASM	2,066	1,675	108	3,741
Hypersonics	1,726	1,292	163	3,018
Total	31,156	9,625	352	40,781

Table 1. Summary of units of measures extracted from each dataset. The Number of Instances were extracted by a regex search with a dictionary containing pre-defined units of measure. The Number of True Positives, Number of False Positives, and the Number of Distinct Units was manually determined.

2.3 Model Development

Problem formulation consisted of converting a text window into a feature vector that may be used to train a model. This feature vector consists of either an embedding output by an LRM considered in this study, a set of auxiliary variables engineered from the input text (Appendix), or both an embedding output and the set of auxiliary variables engineered from the text. For BERT and SCIBERT, we used the embedding for the token that represented the quantity given the surrounding context of the input text window. For GloVe, we used the average embedding of the token that preceded the quantity and immediately after the unit of measure. We avoid using the token embedding for the unit of measure, as this may cause a model to discriminate against that unit of measure.

This problem formulation is equivalent to a binary classification task, which involves developing a machine learning model that can output the probability a text window contains a quantity and unit of measure given the input feature vector for that text window. We utilized the random forest algorithm for developing these machine learning models[6] with modular 3-fold cross-validation[7]. To develop these random forest models, a random search for hyperparameter optimization[8] was performed using the hyperparameter space summarized in Table 2.

Random Forest	
# of Trees	$\boldsymbol{\mathcal{U}}(1, 250)$
# Variables in Split	$\boldsymbol{\mathcal{U}}(1, 100)$
Max. Tree Depth	$\boldsymbol{\mathcal{U}}(1, 100)$
Min. # of Obs. in Leaf Node	$\boldsymbol{\mathcal{U}}(1, 30)$

Table 2. Hyperparameters considered for model development. (a, b) denotes uniform distribution from a to b.

2.4 Experimental Setup and Model Selection

We developed three different sets of models for comparison, where each set of models had a distinct input feature vector – (i) the models were developed from either BERT, SCIBERT, or GloVe embeddings only; (ii) the models were developed from auxiliary feature engineered variables (Appendix) and either BERT, SCIBERT, or GloVe embeddings; (iii) the models were developed from auxiliary feature engineered variables only. Each set of models were developed from four different subsets of the dataset – (i) the SBIR dataset; (ii) the ASM dataset; (iii) the Hypersonics dataset; (iv) the combined SBIR, ASM, and Hypersonics datasets.

Model selection was based on which model had the lowest logistic loss from cross-validation – logistic loss is a metric of model quality, where smaller values indicate better performance (LL)[7]. LL has a minimal value of zero. We also give consideration to the F_1 Score, which is a threshold metric that assesses the accuracy of a model (F1)[9], as well as the area under the receiver operating curve, which is a ranking metric that assesses the ability of a model to discriminate between classes (AUC)[10]. Higher values indicate better performance for F1 and AUC, where both have a maximal value of one. If a selected model is developed from either the SBIR, ASM, or Hypersonics dataset (not all three), then that model is assessed on the out-of-sample datasets.

We note that the selected model with the best performance metrics overall (based on LL) is utilized to assess feature importance. We consider two types of assessments for feature importance: (i) a global assessment (general trend of how the model weighs each feature) and (ii) a local assessment (trend of how the model weighs each feature for a specific instance). The global assessment is based on the variable importance measure for each variable output by a

random forest model[6]. A variable importance measure summarizes the contribution of each input variable in improving a performance metric during training (LL) divided by the contribution of every input variable in improving this performance metric. We place a slight variation on computing variable importance measures, as our feature vector has a hierarchical representation – for instance, we considered 50 dimensional GloVe vectors, which would represent 50 input variables in our feature vector rather than 1. This variation involves taking the summation of variable importance measures to represent an aggregated importance for hierarchical features such as token embeddings.

We also utilize this model to gather Shapley additive explanations, or "SHAP values", which describe the contribution of each input variable in influencing the prediction output for an instance[11]. These SHAP values represent a new dataset, from which we perform principal component analysis (PCA) to determine whether or not our aggregated variable importance measures hold for all instances of data[12].

3. RESULTS

3.1 Performance Comparison

Table 3 compares the selected models developed from SBIR, ASM, and Hypersonics datasets by their cross-validation performance metrics; Table 4 compares the selected models developed from the SBIR dataset by their performance metrics on out-of-sample ASM and Hypersonics datasets; Table 5 compares the selected models developed from the ASM dataset by their performance metrics on out-of-sample SBIR and Hypersonics datasets; Table 6 compares the selected models developed from the Hypersonics dataset by their performance metrics on out-of-sample SBIR and ASM datasets. In these tables, LL denotes logistic loss; F1 denotes F_1 score; AUC denotes area under the receiver operating characteristic curve. GloVe denotes the selected model developed solely from GloVe embeddings; BERT the selected model from solely BERT embeddings; SCIBERT the selected model from solely SCIBERT embeddings; Auxiliary the selected model solely from the feature engineered auxiliary variables. A "+" denotes both token embeddings from an LRM and auxiliary variables were input into the model.

Tables 3-6 indicate that in each experiment, GloVe has consistently poor performance – this may be due to GloVe token embeddings being static, regardless of the surrounding context for a token embedding. Despite this, the use of GloVe token embeddings with the feature engineered auxiliary variables results in strong performing models, typically outperforming models developed from BERT or SCIBERT token embeddings alone. This may be indicative that this classification task has a low complexity – there may be a simple set of auxiliary variables that render the use of LRMs unnecessary[13]. We emphasize that this is not currently true, as models that utilized SCIBERT embeddings consistently produced the best performing models, which indicates value in utilizing LRMs fine-tuned for a specific domain in a downstream task.

We emphasize that in the experiments where the models were developed from one of the three datasets and assessed by their performance metrics on the remaining two (Tables 4-6), it is apparent that models utilizing SCIBERT token embeddings attain the best predictive performance. These results are particularly notable because these SCIBERT models experienced the least amount of performance degradation relative to their performance metrics in Table 3, which is indicative of the value of SCIBERT token embeddings in tasks for S&T documents. These results are also notable because the number of training instances are small. Specifically, in Table 5, each model is developed from 3,741 training instances, and assessed on 37,040 out-of-sample test instances; in Table 6, each model is developed from 3,018 training instances and assessed on 37,763 test instances.

	LL	F1	AUC
GloVe	0.311	0.921	0.901
BERT	0.218	0.956	0.967
SCIBERT	0.168	0.963	0.976
Auxiliary	0.515	0.927	0.904
GloVe+Auxiliary	**0.148**	0.969	**0.985**
BERT+Auxiliary	0.181	0.968	0.976
SCIBERT+Auxiliary	0.168	**0.970**	0.981

Table 3. Cross-validation model performance when models are developed from every dataset considered in this study. The best performance metric attained, shown in bold, is split across GloVe + Auxilary and SCIBERT + Auxiliary.

	LL	F1	AUC
GloVe	0.616	0.789	0.833
BERT	0.434	0.770	0.888
SCIBERT	0.313	0.921	0.973
Auxiliary	0.356	0.878	0.940
GloVe+Auxiliary	0.284	0.898	0.962
BERT+Auxiliary	0.354	0.868	0.935
SCIBERT+Auxiliary	**0.268**	**0.923**	**0.975**

Table 4. Out-of-sample model performance from models developed with the SBIR dataset on ASM and Hypersonics datasets. The best performance metric attained, shown in bold, unanimously falls within SCIBERT + Auxiliary for this dataset.

	LL	F1	AUC
GloVe	0.431	0.854	0.800
BERT	0.309	0.926	0.931
SCIBERT	0.284	0.941	0.943
Auxiliary	0.284	0.933	0.907
GloVe+Auxiliary	0.270	0.928	0.923
BERT+Auxiliary	0.272	0.938	0.939
SCIBERT+Auxiliary	**0.257**	**0.950**	**0.948**

Table 5. Out-of-sample model performance from models developed with the ASM dataset on SBIR and Hypersonics datasets. The best performance metric attained, shown in bold, unanimously indicates SCIBERT + Auxiliary.

	LL	F1	AUC
GloVe	0.475	0.818	0.802
BERT	0.454	0.906	0.915
SCIBERT	**0.315**	0.944	0.950
Auxiliary	0.321	0.921	0.915
GloVe+Auxiliary	0.319	0.939	0.927
BERT+Auxiliary	0.486	0.937	0.934
SCIBERT+Auxiliary	0.318	**0.947**	**0.951**

Table 6. Out-of-sample model performance from models developed with the Hypersonics dataset on SBIR and ASM datasets. The best performance metric attained, shown in bold, is split between SCIBERT and SCIBERT + Auxiliary.

3.2 Feature Importance Assessment

Figure 2 displays the global feature importance measures extracted for the best performing model overall when trained on all three datasets. This indicates that the use of measurements may discriminate against certain units of measure. For example, the unit of measure "Debye" is associated with the measurement "Electric Dipole Moment". However, the unit abbreviation for debye is "D" and, depending on the text repository used for model development, "D" may frequently be utilized as an abbreviation for dimensionality (Figure 1).

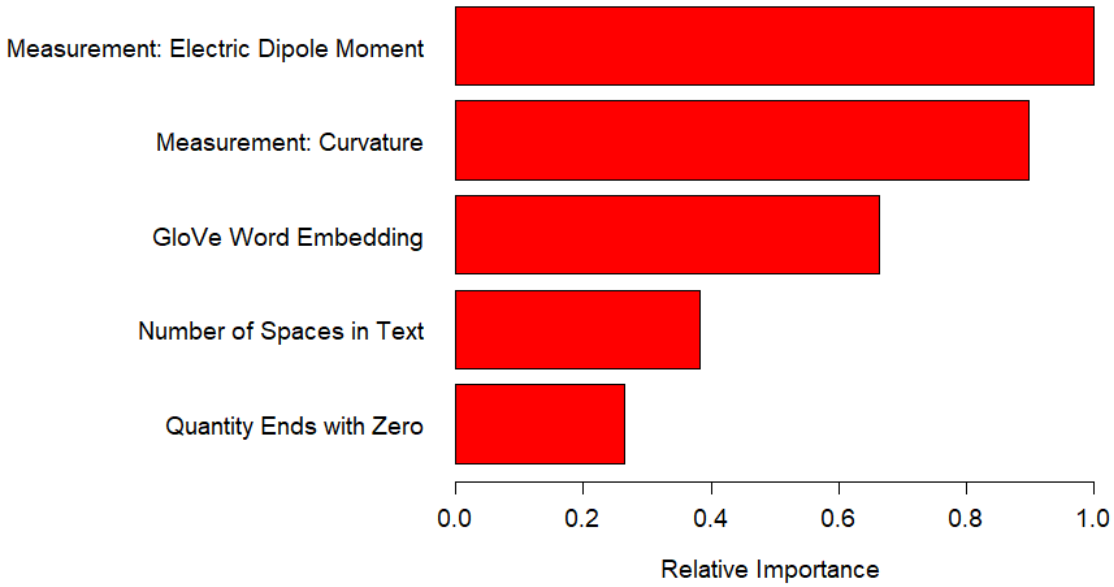

Figure 2. Five most important variable importance measures. Note that the importance measures are scaled by "relative importance" – the variable importance measure of each feature is divided by the largest variable importance measure observed, scaling the variable importance measures from zero to one.

Figure 3 shows the output of PCA on the SHAP values for local feature independence. This plot indicates that there is no noticeable clustering for misclassified instances. Due to this, it may be reasonable to assume that the global feature measures found in Figure 2 are indeed strong predictors for discriminating between true positives and false positives for the datasets considered in this study.

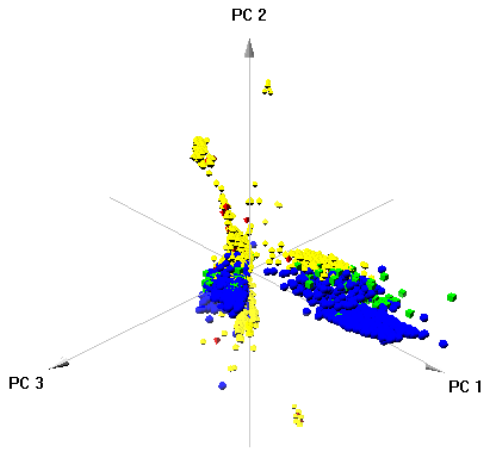

Figure 3. PCA output from SHAP values for local feature importance assessment. A yellow diamond denotes that the model labeled the unit of measure correctly when the unit of measure was a true positive; green square denotes an incorrect model label when the unit of measure was true positive; a blue circle denotes an incorrect model label when the unit of measure was false positive; a red triangle denotes an incorrect model label when the unit of measure was false positive.

4. CONCLUSIONS

The method of processing text described here provides a robust way to mine quantities and units of measure out of S&T data. We empirically found that SCIBERT, a specialized language representation model for scientific and technical documents, consistently achieved the best performance metrics out of the three language representation models considered in the study, and that these results hold when model development is performed on as few as 3,000 training instances. We also found that utilizing our feature engineered variable set with the token embeddings output by a language representation model further improved the performance metrics attained. We emphasize that this capability may be utilized to find quantities and units of measure in new domains with minimal manual effort.

5. ACKNOWLEDGEMENTS

This material is based upon work supported by the Air Force Research Laboratory Information Directorate under Contract No. FA8750-18-C-0133. Any opinions, findings and conclusions or recommendations expressed in this material are those of the author(s) and do not necessarily reflect the views of the Air Force Research Laboratory.

6. REFERENCES

[1] Goodfellow, I., Bengio, Y., and Courville, A., [Deep Learning], MIT PRESS, Cambridge.

[2] Devlin, J., Chang, M.-W., Lee, K., and Toutanova, K., “Bert: Pretraining of deep bidirectional transformers for language understanding,” Proc. NAACL-HLT, 4171–4186 (2019).

[3] Beltagy, I., Cohan, A., and Lo, K., “Scibert: Pretrained contextualized embeddings for scientific text”, CoRR abs/1903.10676, (2019).

[4] Niven, T., and Hung-Yu, K., “Probing neural network comprehension of natural language arguments”, Proc. ACL, 4658–4664 (2019).

[5] Pennington, J., Socher, R., and Manning, C. D., “Glove: Global vectors for word representation”, Proc. EMNLP, 1532– 1543 (2014).

[6] Breiman, L., “Random forests”, Machine learning 45(1), 5–32 (2001).

[7] Hastie, T., Tibshirani, R., and Friedman, J. H., [The Elements of Statistical Learning], Springer, New York.

[8] Bergstra, J., and Bengio, Y., “Random search for hyper-parameter optimization”, JMLR 13:281–305 (2012).

[9] Sokolova, M., and Lapalme, G., “A systematic analysis of performance measures for classification tasks”, Inf Process Manage 45, 427–437 (2009).

[10] Provost, F., and Fawcett, T., "Robust classification for imprecise environments", Machine Learning, 42(3):203–231 (2001).

[11] Lundberg, S. M., and Lee, S. I., ‘‘A unified approach to interpreting model predictions,’’ Proc. NIPS, 4768–4777 (2017).

[12] Jolliffe, I. T., [Principal Component Analysis], Springer-Verlag, New York (1986).

[13] Mackay, D. J. C., [Information Theory, Inference, and Learning Algorithms], Cambridge University Press Cambridge, (2003).

7. APPENDIX

The following list defines the set of auxiliary feature engineered variables.

- Number of characters representing quantity.
- Presence of decimal in quantity.
- Numeric value of quantity.
- Quantity is between 1800 and 2050 (years that may appear in S&T documents).
- Quantity starts with zero.
- Quantity ends with zero.
- Number of characters representing unit.
- Number of uppercase characters in unit.
- Number of lowercase characters in unit.
- Measurements present in an input text window (a measurement encompasses a family of units of measure; e.g., “meters” and “feet” are units in the measurement family “length”). This is represented by 317 dimensional vector of binary indicator variables.
- Number of total units in an input text window.
- Text window contains a “problem unit” (a unit that has ambiguous meaning; Table A1).
- Number of spaces in an input text window.
- Number of newline characters in an input text window.
- Proportion of alphabetical characters in an input text window.
- Proportion of “ “ characters in an input text window.
- Proportion of “[“ or “]” characters in an input text window.
- Proportion of “.” characters in an input text window.
- Proportion of “-” in an input text window.
- Proportion of “+” in an input text window.
- Proportion of other characters not previously defined in an input text window.
- Tokens that had empirically appeared to have high value in discrimination. This is represented by 26 dimensional vector of binary indicator variables.

Unit
at (technical atmosphere)
atm (atmosphere)
C (Celsius, coulomb)
in (inches)
K (Kelvin)
m (meters)
rd (Rutherford)
T (Tesla)
V (volts)
W (watts)

Table A1. Units that models initially had difficulty discriminating.