NUKELM: Pre-Trained and Fine-Tuned Language Models for the Nuclear and Energy Domains

Lee Burke and Karl Pazdernik and Daniel Fortin and Benjamin Wilson and Rustam Goychayev
Pacific Northwest National Laboratory

Abstract
Natural language processing (NLP) tasks (text classification, named entity recognition, etc.) have seen revolutionary improvements over the last few years. This is due to language models such as BERT that achieve deep knowledge transfer by using a large pre-trained model, then fine-tuning the model on specific tasks. The BERT architecture has shown even better performance on domain-specific tasks when the model is pre-trained using domain-relevant texts. Inspired by these recent advancements, we have developed NUKELM, a nuclear-domain language model pre-trained on 1.5 million abstracts from the U.S. Department of Energy Office of Scientific and Technical Information (OSTI) database. This NUKELM model is then fine-tuned for the classification of research articles into either binary classes (related to the nuclear fuel cycle [NFC] or not) or multiple categories related to the subject of the article. We show that continued pre-training of a BERT-style architecture prior to fine-tuning yields greater performance on both article classification tasks. This information is critical for properly triaging manuscripts, a necessary task for better understanding citation networks that publish in the nuclear space, and for uncovering new areas of research in the nuclear (or nuclear-relevant) domains.

1 Introduction
While natural language processing (NLP) has made significant strides in recent years, its application to the nuclear domain has remained rudimentary. In any domain, the ability to classify and triage information is critical when the data volume is large and growing. To enable the discovery of new connections between existing technologies or the potential use of a new technology in the nuclear domain, simple keyword searches are insufficient. To accelerate research in the nuclear domain, a language model is needed—one that “understands” nuclear terminology, “understands” terminology in similar energy domains, and can automatically uncover latent similarities between materials, methodologies, and technologies.

In addition to accelerating nuclear science, this new methodology would be valuable to the International Atomic Energy Agency (IAEA) as part of their information collection and processing system. Quantifying the threat of a nation state’s nuclear capability presents a particularly complex problem because the use, development, and transfer of nuclear technology is not itself an indication of nefarious intent. Technology itself has the added complexity of encompassing both physical items of trade, as well as social networks in academia and industry settings, where the “technology” is not a physical, tradeable good, but the knowledge and capabilities of individuals (Molas-Gallart, 1997). Further, as international scientific collaborations become more prevalent, transfer of nuclear technology may become more prevalent, including inadvertent transfers. Readily available open-source information about such research collaborations, e.g., journal papers and technical reports, can offer indications of the use or transfer of such technology.

The amazing progress of state-of-the-art NLP methods has opened up new opportunities for nuclear domain researchers to leverage powerful language models. Models like BERT (Devlin et al., 2019) have shown significant improvement in NLP benchmarking metrics, such as the General Language Understanding Evaluation (GLUE) benchmark (Wang et al., 2019). These benchmark metrics evaluate a language model’s ability to perform a variety of tasks that resemble human ability to comprehend and be language literate. Though undoubtedly one element of BERT’s success is its large architecture of stacked Transformers (Vaswani et al.), another is the widespread use...
of transfer learning: pre-training on one task then fine-tuning on another. By pre-training on general-purpose corpora, a model has a strong foundation when approaching particular benchmark tasks.

There is also evidence that the performance of pre-trained language models on some tasks can be improved even further by domain-adaptive pre-training (Gururangan et al., 2020)—that is, starting with a model pre-trained on general-purpose corpora, then continuing the pre-training process on a corpus that is more representative of the domain of interest.

Given the recent success of large, Transformer-based neural network architectures and domain-adaptive pre-training, as well as the need for nuclear-“aware” NLP models, we have developed NUKE-LM, a language model trained on nuclear-relevant research that performs best on nuclear-relevant downstream tasks.

2 Related Work

A number of scientific and computational advances in recent years have led to significant improvements in the performance of computational models for natural language inference and understanding. Notable among these is the field of transfer learning, using pre-trained models for downstream tasks perhaps markedly different from their original tasks. Often, this takes the form of semi-supervised learning, where a model is trained on unlabeled data using a self-supervised task, then fine-tuned on a supervised task in the same domain.

Word embeddings (e.g., WORD2VEC (Mikolov et al.), GLOVE (Pennington et al.), FASTTEXT (Mikolov et al., 2018)) learn a projection from the high-dimensional vocabulary space of a corpus of texts into a much smaller vector space using self-supervised training tasks like predicting nearby words. A key drawback of this approach is that each word is associated with a single vector, regardless of context.

A number of approaches have been proposed to learn contextualized word embeddings. For instance, ELMo (Peters et al.) trains separate forward- and backward-oriented models for next-word prediction, then learns linear combinations of the deep representations for downstream tasks. In contrast, BERT (Devlin et al., 2019) learns to encode context from both left and right at once using a very large architecture of stacked Transformers (Vaswani et al.), pre-training with both a word prediction task (masked language modeling, MLM) and a task to predict whether a given sample follows another in the original text, relative to being chosen randomly from the corpus (next-sentence prediction, NSP).

ROBERTa (Liu et al., 2019) leverages the same Transformer-based architecture as BERT, but shows improvements on downstream tasks with some changes to its pre-training strategy: it removes the NSP objective, pre-training only with MLM; it allow samples to cross document boundaries in pre-training, ensuring all pre-training samples are as long as possible; it determines which tokens to predict in each batch rather than deciding offline, before training; it uses much larger batch sizes; it uses byte-level tokenization instead of character-level; and finally, it considers much more pre-training data, including those from the Common Crawl corpora.

SciBERT (Beltagy et al., 2019) clones BERT’s stacked Transformer architecture and pre-training methodology but replaces the BERT training corpus with a large, multi-domain corpus of scientific publications. This results in better performance on scientific domain tasks because of the better match between the domains of pre-training and fine-tuning tasks.

In contrast to training a domain-specific model from scratch like SciBERT, Gururangan et al. (Gururangan et al., 2020) demonstrate that continued pre-training of a general-purpose language model on in-domain text (called domain-adaptive pre-training, DAPT) can lead to improved performance on downstream tasks, but that continued pre-training on out-of-domain text can worsen performance. They explore several ways to bootstrap a targeted continued-pre-training corpus and explore the tradeoff between performance and computational expense.

Similarly, several domain-specific models have been proposed that continue pre-training from a BERT checkpoint. BIOBERT (Lee et al., 2019) continues pre-training on biomedical corpora. NUKEBERT (Jain et al., 2020) continues pre-training on a nuclear-domain corpus, with the addition of newly initialized vocabulary entries specific to the nuclear domain. However, in contrast to NUKE-LM, the pre-training corpus for the NUKEBERT model was generated from a relatively small corpus consisting of about 7000 internal reports from the Indira Gandhi Center for Atomic Research,
largely focused on fast breeder reactors; the NUKEBERT language model is somewhat narrowly focused on nuclear reactor research for power generation rather than defining topics broadly associated with the nuclear fuel cycle. Furthermore, it is not clear if the NUKEBERT language model is publicly available, and the associated dataset is not available under a standard open-source license.

3 Data

We consider scientific abstracts from the U. S. Department of Energy Office (DOE) Scientific and Technical Information (OSTI) database (OSTI) obtained in November 2018, amounting to nearly two million abstracts from over 70 years of research results from DOE and its predecessor agencies.

For fine-tuning, we consider only abstracts labeled with a subject category. The possible categories are formalized by OSTI, and all products submitted to OSTI are encouraged to provide at least one, listing the primary category first. If more than one category is specified, we consider only the first. In addition to the multi-class labels induced by the OSTI subject categories, we formulate binary labels by identifying OSTI subject categories that correspond to the top level of the IAEA Physical Model (Liu and Morsy, 2001), which describes acquisition pathways. The topics described in the IAEA Physical Model include ore mining and milling, pre-conversion, uranium enrichment, post-conversion, fuel fabrication, nuclear reactors, heavy water production, and reprocessing of irradiated fuels. Using this criterion, the following OSTI topic categories are considered related to the nuclear fuel cycle for the binary classifier: nuclear fuels, isotope and radiation sources, nuclear fuel cycle and fuel materials, management of radioactive and nonradioactive wastes from nuclear facilities, specific nuclear reactors and associated plants, general studies of nuclear reactors, radiation chemistry, instruments related to nuclear science and technology, and nuclear physics and radiation physics. The list of all OSTI categories and their binary categorization designation is provided in Appendix A.

4 Experimental Setup

We begin with pre-trained checkpoints implemented in HuggingFace's transformers framework (Wolf et al., 2019), available from the HuggingFace model database with the following slugs: roberta-base and roberta-large are base and large versions of the RoBERTa model, respectively, and allenai/scibert_scivocab_uncased is the recommended uncased version (i.e., inputs are converted to lower case) of SciBERT.

Following (Gururangan et al., 2020), we perform domain-adaptive pre-training. We continue pre-training all three models, SciBERT, ROBERTA BASE, and ROBERTA LARGE, on 80% of the OSTI abstracts, and for the remainder of this manuscript, we use the naming convention NUKELM to define the latter. The remaining 20% of documents are held out from the pre-training process and split evenly into two data sets (200 K each) to be used for fine-tuning and testing the classification models. When forming each batch, 512-token segments are taken irrespective of document boundaries, and 15% of the tokens are masked for prediction. We train for 13 K steps with a batch size of 256, for a total of 3.3 M samples consisting of 1.7 B tokens (similar in size to the corpora in (Gururangan et al., 2020)). Other hyperparameters follow (Gururangan et al., 2020).

We perform some exploratory analysis of the impact of domain-adaptive pre-training on OSTI abstracts, including performance metrics and an example of masked word modeling.

For fine-tuning, we begin with the six models described above: ROBERTA BASE and LARGE and SciBERT, both with and without OSTI domain-adaptive pre-training. We then follow (Gururangan et al., 2020) by passing the final layer $[\text{CLS}]$ token representation to a task-specific fully connected layer for prediction (see the transformers documentation for details). A validation set is held out, consisting of 10% of the overall fine-tuning set.

We consider two tasks: multi-class prediction over the original OSTI subject categories, and binary prediction over the relevance of an abstract’s subject category to one of the steps of the nuclear fuel cycle. The fine-tuning data set consisted of 198,564 documents, of which 23,268 are related to the nuclear fuel cycle according to our definition.

A small hyperparameter search is performed on the binary task (details in Appendix B), selecting a learning rate of $10^{-5}$ and a batch size of 64. We train for five epochs (14.7 K steps), evaluating at 20 checkpoints (about every 750 steps) and saving the best model according to loss on the validation set. Other hyperparameters follow (Gururangan et al., 2020).
5 Results of the Language Modeling Task

5.1 Metrics

The MLM task is evaluated based on the categorical cross-entropy between the one-hot true distribution over a model’s vocabulary and a model’s predicted distribution. This MLM loss is shown before and after domain-adaptive pre-training for each of the three baseline models in Table 1.

Continued pre-training improves the performance of RoBERTA BASE more than that of SciBERT, to the point where it performs better than the much larger RoBERTA without continued pre-training. The RoBERTA pre-training strategies may have yielded an easier-to-train model than the SciBERT methodologies, but this may be due solely to the larger vocabulary size, 50 K tokens for RoBERTA vs. 30 K for SciBERT. Regardless, NUKELM shows improvement over RoBERTA LARGE, and remains the most accurate of the models.

Table 1: Masked language modeling loss, based on categorical cross-entropy between true and predicted probability distributions, on the evaluation sub-set of the OSTI pre-training data. Lower is better. The symbol “+ OSTI” denotes continued pre-training on OSTI abstracts. The best performing model is in bold.

| Model       | MLM Loss |
|-------------|----------|
| RoBERTA BASE| 1.39     |
| RoBERTA BASE + OSTI | 1.11 |
| RoBERTA LARGE | 1.13     |
| NUKELM      | 0.95     |
| SciBERT     | 1.34     |
| SciBERT + OSTI | 1.18     |

5.2 MLM Example

We present an example of masked language modeling to illustrate the task and performance improvement after domain-adaptive pre-training. The **bolded** word is masked, and the models are asked to predict what word should fill in the blank.

The use of heavy **water** as the moderator is the key to the PHWR system, enabling the use of natural uranium as the fuel (in the form of ceramic UO2), which means that it can be operated without expensive uranium enrichment facilities. (Wikipedia)

Table 2 summarizes the top five predicted tokens and their associated likelihood score from each of the six models after domain-adaptive pre-training (if any) but before fine-tuning. Before continued pre-training, all three models include the correct answer in their top five predictions, but RoBERTA BASE and SciBERT predict the more common but incorrect phrase “heavy metal,” albeit with low confidence; only RoBERTA LARGE predicts the correct answer. After continued pre-training, all three models succeed in predicting the correct answer with high confidence.

Table 2: An example of masked language modeling. Column 2 contains the top five tokens considered most likely (the true token, “water”, is in bold), and column 3 contains the associated likelihood scores (the highest confidence for the true token is also in bold). The character “#” indicates the token is a sub-word, i.e., a prediction of “heavywater” rather than “heavy water”. The symbol “+ OSTI” denotes continued pre-training on OSTI abstracts.

| Model                | Top-5 Preds. | Score |
|----------------------|--------------|-------|
| RoBERTA BASE         | metal        | 0.252 |
|                      | metals       | 0.149 |
|                      | uranium      | 0.145 |
|                      | water        | 0.130 |
|                      | iron         | 0.026 |
| RoBERTA BASE + OSTI  | water        | 0.955 |
|                      | metal        | 0.008 |
|                      | elements     | 0.008 |
|                      | metals       | 0.008 |
|                      | oil          | 0.003 |
| SciBERT              | water        | 0.951 |
|                      | metal        | 0.013 |
|                      | metals       | 0.011 |
|                      | fuel         | 0.004 |
|                      | carbon       | 0.002 |
| NUKELM               | water        | 0.996 |
|                      | metals       | 0.001 |
|                      | oil          | 0.001 |
|                      | #water       | <0.001|
|                      | metal        | <0.001|
| SciBERT              | metal        | 0.225 |
|                      | metals       | 0.117 |
|                      | water        | 0.068 |
|                      | iron         | 0.052 |
|                      | argon        | 0.042 |
| SciBERT + OSTI       | water        | 0.929 |
|                      | metal        | 0.024 |
|                      | metals       | 0.011 |
|                      | iron         | 0.003 |
|                      | oil          | 0.003 |
6 Results of Downstream Tasks

6.1 Multi-Class Classification

The results of fine-tuning of the multi-class classification task are presented in Table 3. SciBERT’s advantage over RoBERTA BASE persists after domain-adaptive pre-training, perhaps because its scientific-domain pre-training corpora are more closely related to the OSTI task than are RoBERTA’s. However, neither overcomes RoBERTA LARGE even without the added advantage of continued pre-training, likely because the latter contains an order of magnitude more trainable parameters.

Table 3: Results of fine-tuning on the multi-class classification task. Precision, Recall, and F1 scores are an average of all classes, weighted by class size. The best performing model by each metric is presented in bold.

| Model      | Accuracy | Precision | Recall | F1 score |
|------------|----------|-----------|--------|----------|
| RoBERTA BASE | 0.6745   | 0.6564    | 0.6745 | 0.6603   |
| RoBERTA BASE + OSTI | 0.6972   | 0.6884    | 0.6972 | 0.6863   |
| RoBERTA LARGE | 0.7056   | 0.7008    | 0.7056 | 0.7013   |
| NUKELM     | 0.7201   | 0.7164    | 0.7201 | 0.7168   |
| SciBERT    | 0.6972   | 0.6866    | 0.6972 | 0.6883   |
| SciBERT + OSTI | 0.7047   | 0.6981    | 0.7047 | 0.6973   |

6.2 Binary Classification

The results of fine-tuning on the binary classification task are presented in Table 4. Without domain-adaptive pre-training, SciBERT performs even better than RoBERTA LARGE, possibly because of its more closely related pre-training corpora. However, unlike in the multi-class task, both SciBERT and RoBERTA BASE see degraded recall, outweighed by a moderate increase in precision only due to class imbalance. Only NUKELM sees improvement across all measured metrics, likely due again to its large size. It is worth noting that the much smaller RoBERTA BASE is able to achieve performance comparable to the unwieldy RoBERTA LARGE via continued pre-training, which may be useful in resource-constrained applications.

Table 4: Results of fine-tuning on the binary classification task. Precision, Recall, and F1 scores consider NFC-related to be the positive class. The best performing model by each metric is presented in bold.

| Model      | Accuracy | Precision | Recall | F1 score |
|------------|----------|-----------|--------|----------|
| RoBERTA BASE | 0.9506   | 0.7938    | 0.7816 | 0.7876   |
| RoBERTA BASE + OSTI | 0.9544  | 0.8237    | 0.7773 | 0.7998   |
| RoBERTA LARGE | 0.9506   | 0.7995    | 0.7722 | 0.7856   |
| NUKELM     | 0.9573   | 0.8270    | 0.8038 | 0.8152   |
| SciBERT    | 0.9548   | 0.8061    | 0.7910 | 0.7984   |
| SciBERT + OSTI | 0.9532  | 0.8285    | 0.7747 | 0.8007   |

6.3 Performance under Different Training Set Sizes

One reported advantage of domain-adapted language models is the ability to fine-tune on smaller numbers of labeled examples. We test this ability with the binary classification task described above.

We randomly select increasingly large proportions of the binary classification fine-tuning set, ignoring the rest, so that each larger subset contains the earlier, smaller subsets. We train the off-the-shelf RoBERTA LARGE and NUKELM with the same experimental set-up as in Section 6.2 and track the log-loss computed on the hold-out evaluation set. Figure 1 summarizes the results.

The domain-adapted model achieves significantly better performance with smaller amounts of data, however this advantage shrinks as the training set size increases. This could be the result of continued pre-training priming the model for performance in this domain. However, this pattern is not entirely consistent which is likely an effect of the random selection of documents.

Interestingly, and contrary to prior expectation, the disparity between models actually decreases at the lowest sample size tested (0.4% of the full corpus). While NUKELM maintains its superiority, with so few examples used for fine-tuning, neither model performs well. We suspect that at a sample size this low (754 versus 1000 documents used for training and test, respectively), both models may be overfitting to the limited trained data, which is why the dominance of NUKELM is muted.
Figure 1: Binary classification performance, measured by log-loss on a hold-out evaluation set, as the training set size is increased, for RoBERTA LARGE (orange) and NUKELM (blue).

6.4 Qualitative Assessment

Beyond model performance on the MLM task and document classification, an important question regarding these trained language models is whether or not any reasonable interpretation can be gleaned from the generally unintelligible vectors representations of the input text. While there is not a clear consensus on how useful these embeddings can be in providing explanations, with arguments from both sides (Jain and Wallace, 2019; Wiegreffe and Pinter, 2019), there is undoubtedly some information contained within these transformer-based language models because their predictive ability is state-of-the-art. So, while a direct interpretation of an embedding produced by NUKELM may be questionable, the transformation of this high-dimensional space that results from pre-training should provide some explanation as to why prediction was improved.

As a first step toward interpreting the impact of domain-adaptive pre-training, we visualize output embeddings from the most accurate model, RoBERTA LARGE, both with and without continued pre-training on OSTI abstracts, and after fine-tuning on the binary classification task. We use UMAP (McInnes et al., 2018) with all default parameters to project the output corresponding to the special token [CLS] down to two dimensions, training separate UMAPs for each model. Figure 2 (top row) depicts the result of this process performed on a 1000-sample random subset of the binary classification task validation set.

In both models, the positive class is generally clustered together; indeed, both models are able to learn relatively accurate decision boundaries. However, in the version without domain-adaptive pre-training, the cluster looks like a single manifold, eventually connecting to the mass of negative samples like an isthmus. In contrast, continued pre-training appears to encourage the model to form more complicated structures, with an isolated cluster of mostly positive samples in addition to a similar but much smaller isthmus connected to a large mass of negative samples.

To explore these differences further, we apply BERTopic (Grootendorst, 2020), a clustering and topic modeling approach for understanding the output embeddings of a transformer model. BERTopic also uses a UMAP projection for dimension reduction, in this case to 100 dimensions, but then uses HDBSCAN (Campello et al., 2013) to cluster documents and a class-based TF-IDF (cb-TF-IDF) score for topic modeling. Here, all documents within the same cluster are concatenated into a single document and then the usual TF-IDF score (Teller, 2000) is computed as follows:

\[
    cb - TF - IDF_i = \frac{t_i}{w_i} \times \log \frac{m}{\sum_{j=1}^{n} t_j},
\]

where \( t_i \) is the frequency of each word in class \( i \), \( w_i \) is the the total number of words in class \( i \), \( m \) is the number of documents, and \( n \) is the number of classes.

We visualize the BERTopic clusters found in the RoBERTA LARGE binary classification models in Figure 2 (bottom row). The three words most representative of each cluster, as determined by the cb-TF-IDF model, are listed in Table 5. Without continued pre-training, we see seven clusters on a variety of topics, from cosmology to biology, with the NFC-related samples mostly relegated to a single nuclear cluster or left as outliers. In contrast, with continued pre-training, non-NFC topics are grouped into a single cluster and nuclear topics are given two clusters, one focused on nuclear facilities and another on nuclear physics. This provides evidence that continued pre-training taught the model additional knowledge of the nuclear domain, allowing it to characterize different subsets of positive examples, and recognize the irrelevance of other distinctions to the fine-tuning task.
Figure 2: Visualization of UMAP-transformed output embeddings from RoBERTa Large for 1000 randomly sampled documents from the validation set after fine-tuning on the binary classification task, both without (left) and with (right) domain-adaptive pre-training on OSTI abstracts, colored by the true binary labels (top) and BERTopic clusters (bottom). Each point in these plots is a low-dimensional representation of the embedding for a document’s abstract.
Table 5: Top three representative words for each BERTopic cluster of output embeddings from RoBERTa LARGE for 1000 randomly sampled documents from the validation set after fine-tuning on the binary classification task, both without and with domain-adaptive pre-training on OSTI abstracts. Column two, the cluster number, corresponds with the legends in Figure 2 (bottom row).

| Model     | No. | Top-3 Words                              |
|-----------|-----|------------------------------------------|
|           | 1   | galaxies, galaxy, dust                   |
|           | 2   | coal, gas, energy                        |
| RoBERTa   | 3   | cells, protein, kinase                    |
| LARGE     | 4   | beam, laser, plasma                       |
|           | 5   | waste, nuclear, radiation                 |
|           | 6   | alloy, temperature, conductivity          |
|           | 7   | water, alkanolamine, lake                 |
| NUKELM    | 1   | nuclear, reactor, facility                |
|           | 2   | energy, system, model                     |
|           | 3   | nuclei, neutron, energies                 |

7 Conclusion and Future Work

In this work, we leveraged abstracts from the OSTI database to train state-of-the-art language models for nuclear-domain-specific classification tasks and as a general-purpose language model in the nuclear domain. We explored a number of base models for transfer learning and applied domain-adaptive pre-training to improve performance on the downstream tasks. To the best performing model in this process, RoBERTa LARGE + OSTI, we apply the name NUKELM.

We consider the NUKELM language model to be a general purpose resource for supporting development of NLP models in the nuclear domain. The NUKELM model can be leveraged for task training on relatively small labeled data sets, making it feasible to manually label training for targeted objectives and easily fine-tune the NUKELM model for various tasks.

Future versions of NUKELM will improve the model training pipeline, including considering full article text and data sets other than OSTI, expanding the model vocabulary to better capture a nuclear domain vocabulary without losing RoBERTa’s more robust pre-training, and exploring multilingual capabilities via models like XLM-RoBERTa (Conneau et al.).

We introduced a binary categorization of the OSTI subject categories aimed at identifying documents related to the nuclear fuel cycle and fine-tuned the NUKELM model on this task. This fine-tuned classification model can be immediately useful as triage tool or to support NLP workflows in nuclear science or nuclear nonproliferation.

The NUKELM binary classification model demonstrated superior performance for the classification task, though performance gains were minor. Because of computational constraints, multiple runs of the training process were not made to establish the statistical significance of the classification metrics, but the large set of training data and consistent trends across model types and tasks make it unlikely that the rank order of these models would change with resampling and retraining. Furthermore, we demonstrate that the performance gain may be even higher with smaller-scale fine-tuning sets.

Moreover, the whole story does not lie within the F1-score, because our qualitative visual assessment of the NUKELM binary classification embeddings reveal intriguing structural differences. The NUKELM embeddings appear to have more distinct clusters and increased separation among clusters, particularly among NFC-related documents. By applying BERTopic to these embeddings, we confirmed that these clusters correspond to identifiable topics. Potential future work would be needed to quantify these structural changes and assess differences among various models, as an in-road toward explaining how the models reach their conclusions.

Acknowledgments

The authors thank Aaron Luttman and Matthew Oster for their helpful feedback; and Gideon Juve, Dan Corbiani, and George Bache for helping to build our computing infrastructure. This work was supported by the NNSA Office of Defense Nuclear Nonproliferation Research and Development, U.S. Department of Energy, and Pacific Northwest National Laboratory, which is operated by Battelle Memorial Institute for the U.S. Department of Energy under Contract DE-AC05–76RLO1830. This article has been cleared by PNNL for public release as PNNL-SA-159410.

References

Iz Beltagy, Kyle Lo, and Arman Cohan. 2019. SciBERT: A Pretrained Language Model for Scientific Text. arXiv:1903.10676 [cs]. ArXiv: 1903.10676.

Ricardo J. G. B. Campello, Davoud Moulavi, and Joerg Sander. 2013. Density-Based Clustering Based on Hierarchical Density Estimates. In Advances
in Knowledge Discovery and Data Mining, Lecture Notes in Computer Science, pages 160–172, Berlin, Heidelberg. Springer.

Alexis Conneau, Kartikay Khandelwal, Naman Goyal, Vishrav Chaudhary, Guillaume Wenzek, Francisco Guzmán, Edouard Grave, Myle Ott, Luke Zettlemoyer, and Veselin Stoyanov. Unsupervised cross-lingual representation learning at scale. Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. arXiv:1810.04805 [cs]. ArXiv: 1810.04805.

Maarten Grootendorst. 2020. Bertopic: Leveraging bert and c-tf-idf to create easily interpretable topics.

Suchin Gururangan, Ana Marasović, Swabha Swayamdipta, Kyle Lo, Iy Beltagy, Doug Downey, and Noah A. Smith. 2020. Don’t Stop Pretraining: Adapt Language Models to Domains and Tasks. In Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics, pages 8342–8360, Online. Association for Computational Linguistics.

Ayush Jain, Dr N. M. Meenachi, and Dr B. Venka-traman. 2020. NukeBERT: A Pre-trained language model for Low Resource Nuclear Domain. arXiv:2003.13821 [cs, stat]. ArXiv: 2003.13821.

Sarthak Jain and Byron C. Wallace. 2019. Attention is not Explanation. arXiv:1902.10186 [cs]. ArXiv: 1902.10186.

Jinhyuk Lee, Wonjin Yoon, Sungdong Kim, Donghyeon Kim, Sunkyu Kim, Chan Ho So, and Jaewoo Kang. 2019. BioBERT: a pre-trained biomedical language representation model for biomedical text mining. Bioinformatics, page btz682.

Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. 2019. RoBERTa: A Robustly Optimized BERT Pretraining Approach. arXiv:1907.11692 [cs]. ArXiv: 1907.11692.

Z. Liu and S. Morsy. 2001. Development of the physical model.

Leland McInnes, John Healy, Nathaniel Saul, and Lukas Grossberger. 2018. Umap: Uniform manifold approximation and projection. The Journal of Open Source Software, 3(29):861.

Tomas Mikolov, Ilya Sutskever, Kai Chen, Greg S Corrado, and Jeff Dean. Distributed representations of words and phrases and their compositionality. In C. J. C. Burges, L. Bottou, M. Welling, Z. Ghahramani, and K. Q. Weinberger, editors, Advances in Neural Information Processing Systems 26, pages 3111–3119. Curran Associates, Inc.

Jordi Molas-Gallart. 1997. Which way to go? Defence technology and the diversity of ‘dual-use’ technology transfer. Research Policy, 26(3):367–385.

OSTI. The department of energy (doe) office of scientific and technical information (osti). https://www.osti.gov/. Accessed: 2018-11.

Jeffrey Pennington, Richard Socher, and Christopher Manning. Glove: Global vectors for word representation. In Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP), pages 1532–1543. Association for Computational Linguistics.

Matthew E. Peters, Mark Neumann,Mohit Iyyer, Matt Gardner, Christopher Clark, Kenton Lee, and Luke Zettlemoyer. Deep contextualized word representations.

Virginia Teller. 2000. Speech and language processing: an introduction to natural language processing, computational linguistics, and speech recognition. Computational Linguistics, 26(4):638–641.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Lukasz Kaiser, and Illia Polosukhin. Attention is all you need. 30:5998–6008.

Alex Wang, Amanpreet Singh, Julian Michael, Felix Hill, Omer Levy, and Samuel R. Bowman. 2019. GLUE: A Multi-Task Benchmark and Analysis Platform for Natural Language Understanding. arXiv:1804.07461 [cs]. ArXiv: 1804.07461.

Wikipedia. Pressurized heavy-water reactor. Page Version ID: 996963562.
### A  OSTI Subject Categories

Table 6: List of OSTI subject category labels, their description where available, and whether they related directly to the nuclear fuel cycle.

| Label | Description | NFC | Label | Description | NFC |
|-------|-------------|-----|-------|-------------|-----|
| 1     | Coal, Lignite, and Peat |     | 44    | Military Technology, Weaponry, and National Defense |     |
| 2     | Petroleum |     | 45    | Instrumentation Related | Y |
| 3     | Natural Gas |     | 46    | To Nuclear Science and Technology |     |
| 4     | Oil Shales and Tar Sands |     | 47    | Other Instrumentation |     |
| 5     | Nuclear Fuels | Y | 48    | Y |     |
| 6     | Isotope and Radiation Sources | Y | 49    | Y |     |
| 7     | Hydrogen |     | 50    | Y |     |
| 8     | Biomass Fuels |     | 51    | Y |     |
| 9     | Synthetic Fuels |     | 52    | Y |     |
| 10    | Nuclear Fuel Cycle and Fuel Materials | Y | 53    | Y |     |
| 11    | Management of Radioactive and Non-Radioactive Wastes From Nuclear Facilities | Y | 54    | Environmental Sciences |     |
| 12    | Hydro Energy |     | 55    | Other Instrumentation |     |
| 13    | Solar Energy |     | 56    | Y |     |
| 14    | Geothermal Energy |     | 57    | Y |     |
| 15    | Tidal and Wave Power |     | 58    | Geosciences |     |
| 16    | Wind Energy |     | 59    | Basic Biological Sciences |     |
| 17    | Fossil-Fueled Power Plants |     | 60    | Applied Life Sciences |     |
| 18    | Specific Nuclear Reactors and Associated Plants | Y | 61    | Radiation Protection and Dosimetry |     |
| 19    | General Studies of Nuclear Reactors | Y | 62    | Radiology and Nuclear Medicine |     |
| 20    | Power Transmission and Distribution |     | 63    | Radiation, Thermal, and Other Environ. Pollutant Effects On Living Orgs. and Biol. Mat. |     |
| 21    | Energy Storage |     | 64    | Physics |     |
| 22    | Energy Planning, Policy, and Economy |     | 65    | Plasma Physics and Fusion Technology |     |
| 23    | Direct Energy Conversion |     | 66    | Physics |     |
| 24    | Energy Conservation, Consumption, and Utilization |     | 67    | Classical and Quantum Mechanics, General Physics |     |
| 25    | Advanced Propulsion Systems |     | 68    | Physics Of Elementary Particles and Fields |     |
| 26    | Arms Control |     | 69    | Nuclear Physics and Radiation Physics | Y |
| 27    | Material Science |     | 70    | Atomic and Molecular Physics |     |
| 28    | Inorganic, Organic, Physical and Analytical Chemistry |     | 71    | Condensed Matter Physics |     |
| 29    | Radiation Chemistry, Radiochemistry, and Nuclear Chemistry | Y | 72    | Superconductivity and Superfluidity |     |
| 30    | Chemistry |     | 73    | Nanoscience and Nanotechnology |     |
| 31    | Engineering |     | 74    | Astronomy and Astrophysics |     |
| 32    | Particle Accelerators |     | 75    | Knowledge Management and Preservation |     |
| 33    | Physics |     | 76    | Mathematics and Computing |     |
| 34    | Nuclear Disarmament, Safeguards, and Physical Protection |     | 77    | General and Miscellaneous |     |
| 35    | System |     | 78    | |     |
| 36    | |     | 79    | |     |
| 37    | |     | 80    | |     |
| 38    | |     | 81    | |     |
| 39    | |     | 82    | |     |
| 40    | |     | 83    | |     |
| 41    | |     | 84    | |     |
| 42    | |     | 85    | |     |
| 43    | |     | 86    | |     |
| 44    | |     | 87    | |     |
| 45    | |     | 88    | |     |
| 46    | |     | 89    | |     |
| 47    | |     | 90    | |     |
| 48    | |     | 91    | |     |
| 49    | |     | 92    | |     |
| 50    | |     | 93    | |     |
| 51    | |     | 94    | |     |
| 52    | |     | 95    | |     |
| 53    | |     | 96    | |     |
| 54    | |     | 97    | |     |
| 55    | |     | 98    | |     |
| 56    | |     | 99    | |     |
B Hyperparameter Tuning

A hyperparameter tuning experiment is performed on the binary classification task using RoBERTA LARGE, both with and without domain-adaptive pre-training. We perform a grid search over maximum learning rates of $1 \cdot 10^{-5}$, $2 \cdot 10^{-5}$, and $5 \cdot 10^{-5}$ and minibatch sizes of 16 and 64. Results on the validation set are summarized in Table 7. Both with and without continued pre-training, a small learning rate and large batch size yield the best loss, though the impact on accuracy and F1 score is both smaller and less clear.

Table 7: Results of a hyperparameter tuning experiment. F1 scores consider NFC-related to be the positive class. The best result for each model is **bolded**.

| Model   | Learning Rate | Batch Size | Accuracy | F1 score | Loss   |
|---------|---------------|------------|----------|----------|--------|
|         | $1 \cdot 10^{-5}$ | 16         | 0.9545   | 0.9537   | 0.1173 |
|         |               | 64         | 0.9506   | 0.9502   | 0.1081 |
| RoBERTA | $2 \cdot 10^{-5}$ | 16         | 0.9397   | 0.9409   | 0.1568 |
| LARGE   |               | 64         | 0.9524   | 0.9523   | 0.1118 |
|         | $5 \cdot 10^{-5}$ | 16         | 0.9206   | 0.9097   | 0.2260 |
|         |               | 64         | 0.9363   | 0.9338   | 0.1699 |
|         | $1 \cdot 10^{-5}$ | 16         | 0.9573   | 0.9568   | 0.1127 |
|         |               | 64         | 0.9573   | 0.9570   | 0.0967 |
| RoBERTA | $2 \cdot 10^{-5}$ | 16         | 0.9520   | 0.9516   | 0.1340 |
| LARGE   |               | 64         | 0.9557   | 0.9559   | 0.0977 |
| + OSTI  | $5 \cdot 10^{-5}$ | 16         | 0.9328   | 0.9279   | 0.2093 |
|         |               | 64         | 0.9525   | 0.9518   | 0.1108 |